University of Naples "Federico II"

PhD Programme in Neuroscience

Director Prof. Lucio Annunziato

PhD Thesis

"Fetal brain growth and development"

Tutor

Prof. Pasquale Martinelli

Student

Dr Raffaele Napolitano

ACADEMIC YEAR 2016-2017

to Linda

TABLE OF CONTENTS

ABBREVIATIONS

AC: abdominal circumference

AV: anterior ventricle

BPD: biparietal diameter

CM: cisterna magna

CRL: crown-rump length

CTG: cardiotocography

DV: ductus venosus

FGLS: Fetal Growth Longitudinal Study

FGR: fetal growth restriction

FL: femur length

HC: head circumference

k: kappa coefficient

OFD: occipito-frontal diameter

QC: quality control

SF: sylvian fissure

SGA: small for gestational age

TC: transcerebellar

TCD: transcerebellar diameter

TT: transthalamic

TV: transventricular

POF: parieto-occipital fissure

PV: posterior ventricle

INTRODUCTION

Fetal brain growth and development is routinely studied using prenatal ultrasound. Since the introduction of ultrasound in antenatal care several reports confirmed the safety of this health care technology and the benefit in improving maternal and perinatal outcome and long term neurodevelopmental outcome. Currently ultrasound is recommended worldwide as the screening and diagnostic technique of choice in pregnancy by international guidelines.[1-5] Ultrasound is used mainly in antenatal care to diagnose fetal abnormalities.[6-14] Fetal anomalies of the central nervous system are a major component of fetal abnormalities detected antenatally. It is estimated that the incidence from long term studies can be as high as 1 in 100 births.[2]

Ultrasound can also be used in several diseases and for numerous screening purposes[15-26] including the assessment of fetal growth as part of antenatal care.[27] Fetal growth restriction (FGR) along with preterm delivery is a major cause of stillbirth, perinatal mortality and abnormal neurodevelopment.[28] In several studies FGR babies born both preterm and at term showed having an increased risk of behavioural problems, cognitive deficiency, attentional problems and aggressive behaviour at school age.[29, 30]

The prevention and treatment of FGR has the potential of reducing the incidence of abnormal neurodevelopmental outcome. This has been particularly been demonstrated in the management of severe preterm FGR (Appendix 1).[31, 32] Controversies still exist on the management of FGR at term. This is due to the difficulty in the screening for FGR at term as most of fetuses might not be small for gestational age (SGA), and to the lack of effective treatments.[33] Challenges in clinical care include the prevention, screening, diagnosis and treatment of FGR to prevent perinatal morbidity and impaired long term neurological outcome.

The morphology of the brain in FGR fetuses has not been demonstrated to have abnormal findings compared with normally grown fetuses.[34] However, in order to study the brain growth and development normally grown fetuses without congenital abnormality have to be selected. One the aims of the main study reported in this thesis was to create standards for fetal structures brain size charts based on serial ultrasound measurements.

The INTERGROWTH-21st is large multicentre, multiethnic, population-based project, conducted between 2008 and 2013 in eight countries. The Fetal Growth Longitudinal Study (FGLS) involves women whose fetuses had both two-dimensional and three-dimensional serial scans every 5 weeks from 14+0 to 41+6 weeks.[35] Women participating in this study have low-risk pregnancies that fulfil well defined and strict inclusion criteria at recruitment.[27]

This study uses a ‘prescriptive’ other than a ‘descriptive’ design to study the fetal growth, i.e. only children from populations with minimal environmental constraints on growth were included. Previous studies on fetal growth are associated with high risk of bias in the methodology used (Appendix 2). As a results of such approach the pregnancy outcome of the FGLS had a low incidence of common obstetric complications (preterm rate: 5%, birth weight at term less than 2.5 kg: 3%, preeclampsia: 1%). As in many diseases in obstetrics, risk factors are similar and therefore a population at low risk of growth problems is also at low risk of other complications. The above findings confirmed the true low risk of the population recruited and the fact that this cohort represents the ideal candidate sample to construct fetal brain structures international standards.

Ultrasound technology requires the input of several software analysis to increase the diagnostic performance and the clinical use. A second line of research associated with this theme is reported in this manuscript (Appendix 3, 4, 5).

One of the source of high variability between different charts reporting of fetal growth is the absence of a comprehensive quality control strategy in fetal ultrasound.[36-38] The above has been a novel component of the FGLS study[39-42] and its result is reported in this manuscript along with other studies involving strategies to implement quality control in fetal ultrasound (Appendix 6, 7, 8).

A systematic review of the literature has been performed to identify all the studies aimed to create brain structures charts. Only studies reporting on six specific fetal brain structures charts were reported.

There is substantial heterogeneity in the methodology used in previous studies aimed to create brain structures charts.[43-52] There is high risk of bias in several domains including the selection of the population, the ultrasound protocol and the analysis of the data. Less than 10% of the identified studies reported on maternal and fetal inclusion criteria, pregnancy outcome, ultrasound quality control and statistical description. Most importantly, no studies reported on long term infant outcome, most probably due to the retrospective descriptive design of the study. The data collection was in fact non-specific for the purpose of the study. Not surprisingly, these are common finding in creating fetal biometry charts as found in previous systematic reviews.[36, 38]

In the last chapter of this manuscript it is reported the study focused on the main objective which is to create international standards for six fetal brain structures by antenatal ultrasound and provide further understanding into the fetal brain development process (Appendix 9). The study was conducted in women taking part in the INTERGROWTH-21st Project whose babies have a low risk of abnormal neurological outcome.

FETAL GROWTH AND FETAL BRAIN DEVELOPMENT

FGR is defined as the failure of a fetus to reach his own growth potential. Most FGR fetuses are SGA but not all of them are growth restricted.

FGR affects between 5 and 10% of fetuses. More than 20 millions newborns have a birth weight less than 2,5 kg worldwide, two third of them have evidence of FGR at birth. These figures are increased in low income countries and they contribute to 95% of low birth weights infants around the world. Most of those babies are born at term (> 37 weeks of gestation) and therefore the vast majority of FGR cannot be attributable to preterm delivery.[53]

Chronic placental insufficiency is a common cause of FGR. Placental insufficiency or utero-placental dysfunction results in insufficient blood flow to the placenta during pregnancy and inadequate supply of nutrients and oxygen to support normal growth of the fetus. Thus, the fetus develops in a chronic hypoxic environment. Placental insufficiency can result in changes in fetal metabolism, hormones, haematology, immunology and cardiovascular function.[54]

One of the major challenge of modern obstetric practice is to screen, diagnose and furnish tools to treat FGR, mainly using ultrasound. The usefulness and limitations of such screening methods have been

evaluated in randomised controlled trials over the last two decades.[55] In some pregnancies and newborns, especially those that are preterm, there is a need to monitor growth more closely to decide if clinical interventions are required. So far non optimal timing for delivery and no treatment has been reported in the management of SGA and FGR at term to improve the neurodevelopmental outcome.[56]

It is unclear why newborns who suffered from FGR have an increased risk of neurological delay, independently from the gestation at delivery.[57] Most studies report an increased incidence of hypoxia leading to hypoxemia. Other causes such as neuroinflammation[54] and abnormal metabolites production can have an impact on the developing brain.[58]

Independently from the above factors it is largely recognised that there is a 'fetal programming' of the adult life. A fetus whose mother is exposed to factors that led to adverse intrauterine milieu is more susceptible to adult diseases.[59] One of the mechanism through which the fetus compensate with the chronic hypoxia is the phenomenon of redistribution of the blood flow. This consists into the increase in the blood flow to the brain, the upper part of the body and the most vital organs (surrenal gland, brain and heart). This mechanism has been demonstrated to be associated with abnormal neonatal and long term neurological outcome. In two systematic reviews of the literature, SGA and FGR with abnormal cerebral redistribution is associated with increased risk of abnormal neurodevelopment in the cognitive, language, motor, behaviour, vision and hearing domain.[28, 57] The above evidence reinforces the findings of higher

risk of abnormal development with increasing severity of the growth restriction.

From the ultrasound point of view the above findings are associated with increased resistance in the umbilical artery blood flow, a decrease in the resistance in the middle cerebral artery, and increased resistance in the ductus venosus (DV) (the latter mainly in severe preterm FGR). All the above fetal vessels can be studies through Doppler assessment at antenatal ultrasound. Those blood flow alterations reflect a change in the cardiovascular function of the fetus. The sympathetic and parasympathetic nervous system is also affected by the hypoxic status which can be evaluated by fetal cardiac responses to stimuli. The cardiotocography (CTG) is a health technology which can register the heart rate fluctuations generated by the nervous system and record the fetal wellbeing status. It has been largely used as an intrapartum monitoring technique but the use of computerised assessment has been reported as an antenatal predictor of hypoxia in FGR fetuses, especially preterm.

Despite no interventional trials are reported on the management at term a recent study was published on the management of FGR before 32 weeks of gestation. The TRUFFLE study is a prospective, European multicentre, unblinded, randomised study, where women with singleton fetuses at 26–32 weeks of gestation who had very preterm FGR (SGA associated with increased resistances in the umbilical artery) were randomly allocated to three timing of delivery plans, which differed according to antenatal monitoring strategies. They were based on the computerised ultrasound

assessment of the fetal heart rate (CTG) and the assessment of Doppler velocimetry in the DV. Delivery plans differed according to three antenatal monitoring strategies: CTG abnormality, early DV changes or late DV changes. The primary outcome was survival without cerebral palsy or neurosensory impairment, or a Bayley III developmental score of less than 85, at 2 years of age. Outcomes assessed were surviving infants with known outcomes at 2 years.[31, 32] 542 eligible women were randomly allocated to monitoring groups. The median gestational age at delivery was 30.7 weeks and mean birthweight was 1019 grams. The proportion of infants surviving without neuroimpairment did not differ between the CTG arm (111 [77%] of 144 infants with known outcome), early DV changes (119 [84%] of 142), and late DV changes (133 [85%] of 157) groups. 12 fetuses (2%) died in utero and 27 (6%) neonatal deaths occurred. Of survivors, more infants where women were randomly assigned to delivery according to late ductus changes (133 [95%] of 144) were free of neuroimpairment when compared with those randomly assigned to CTG (111 [85%] of 131), but this was accompanied by a non-significant increase in perinatal and infant mortality. The conclusion of the study was that timing of delivery based on the study protocol using late changes in the DV waveform might produce an improvement in developmental outcomes at 2 years of age.

Being this cohort of women high risk, many infants in the TRUFFLE study were delivered because of other maternal and fetal indications. It was the objective of a secondary study to present a post-hoc sub-analysis to

investigate the indications for delivery in relation to outcome at 2 years in infants delivered before 32 weeks, to come to a further refinement of management proposals of severely FGR babies (Appendix 1).[60] The study findings were that overall only 32% of fetuses born alive were delivered according to the specified monitoring parameter for indication for delivery. There was an increase rate of intact neurological survival in fetuses randomised into the DV arms. Therefore, the optimal timing of delivery can be achieved by combined longitudinal monitoring using both computerised CTG and DV.

Appendix 1: The TRUFFLE study; fetal monitoring indications for delivery in 310 IUGR infants with 2 year's outcome delivered before 32 weeks of gestation

This is study has been accepted for publication but the final proof is under the review of the TRUFFLE scientific group and it might undergo substantial review of the data before the final publication in the journal.

Gerard H.A.Visser (1), C.M Bilardo (2), J.B.Derks (1), E Ferrazzi (3), N.Fratelli (4), T. Frusca (5), W. Ganzevoort (6), C.Lees (7), R. Napolitano (8), T.Todros (9), H.Wolf (6), K.Hecher (10) on behalf of the TRUFFLE group investigators*

1. Department of Perinatology, University Medical Center, Utrecht, Netherlands
2. Department of Obstetrics and Gynaecology, University Medical Center, University of Groningen, Netherlands
3. Department of Woman, Mother and Neonate, Buzzi Children's Hospital, University of Milan, Milan, Italy
4. Maternal-Fetal Medicine Unit, University of Brescia, Brescia, Italy
5. Department of Obstetrics and Gynecology, University Hospital, Parma, Italy

6. Department of Obstetrics and Gynecology, Academic Medical Centre, Amsterdam, Netherlands

7. Department of Surgery and Cancer, Imperial College London, London, UK; and Department of Development and Regeneration, KU Leuven, Leuven, Belgium

8. Department of Gynecology and Obstetrics, University Federico II of Naples, Naples, Italy

9. Department of Obstetrics and Gynecology, Sant' Anna Hospital, Turin, Italy

10. Department of Obstetrics and Fetal Medicine, University Medical Center, Hamburg-Eppendorf, Germany

*TRUFFLE group investigators: N Marlow (11) , B Arabin (12) , C Brezinka (13), A Diemert (10), JJ Duvekot (14), P Martinelli (7), E Ostermayer (15), AT Papageorghiou (16), D Schlembach (17), KTM Schneider (15), B Thilaganathan (16), A Valcamonico (4).

11. Department of Neonatology, UCL Institute for Women's Health, London, UK

12. Center for Mother and Child of the Phillips University, Marburg, Germany

13. Obstetrics and Gynecology, Universitatsklinik fur Gynakologische Endokrinologie und Reproduktionsmedizin, Department fur Frauenheilkunde, Innsbruck, Austria

14. Department of Obstetrics and Gynaecology, Erasmus MC, Rotterdam, Netherlands

15. Division of Perinatal Medicine, Department of Obstetrics and Gynecology, Technical University, Munich, Germany

16. Department of Obstetrics, St George's, University of London, London, UK

17. Department of Obstetrics, Friedrich Schiller University of Jena, Jena, Germany

TRUFFLE group collaborators: A Aktas (Marburg), S Borgione (Turin), R Chaoui (Berlin), JMJ Cornette (Rotterdam), T Diehl (Hamburg), J van Eyck (Zwolle), IC van Haastert (Utrecht), J Kingdom (Toronto), S Lobmaier (Munich), E Lopriore (Leiden), H Missfelder Lobos (Cambridge), G Mansi (Naples), P Martelli (Brescia), G Maso (Trieste), K Marsal (Lund), U Maurer-Fellbaum (Graz), N Mensing van Charante (Amsterdam), S Mulder-de Tollenaer (Zwolle), M Oberto (Turin), D Oepkes (Leiden), G Ogge (Turin), JAM van der Post (Amsterdam), F Prefumo (Brescia/London), L Preston (Cambridge), F Raimondi (Naples), H Rattue (London), IKM Reiss (Rotterdam), LS Scheepers (Nijmegen/Maastricht), A Skabar (Trieste), M Spaanderman (Nijmegen), J Thornton (Nottingham), H Valensise (Rome), N Weisglas–Kuperus (Rotterdam), A Zimmermann (Munich).

ABSTRACT

Objective: In the TRUFFLE study on outcome of early fetal growth restriction women were allocated to three timing of delivery plans according to antenatal monitoring strategies based on reduced computerized cardiotocographic heart rate short term variation (c-CTG STV) , early Ductus Venosus (DV p95) or late DV (DV noA) changes. However, many infants were per protocol delivered because of 'safety net' criteria, or for maternal indications, or 'other fetal indications' or after 32 weeks of gestation when the protocol was not applied anymore. It was the objective of the present post-hoc sub-analysis to investigate the indications for delivery in relation to outcome at 2 years in infants delivered before 32 weeks, to come to a further refinement of management proposals.

Methods: We included all 310 cases of the TRUFFLE study with known outcome at 2 years corrected age and 7 perinatal and infant deaths, apart from 7 cases with an inevitable death. Data were analysed according to the randomization allocation and specified for the intervention indication.

Results: Overall only 32% of fetuses born alive were delivered according to the specified monitoring parameter for indication for delivery. 38% were delivered because of safety net criteria, 15% because of other fetal

reasons and 15% because of maternal reasons. In the c-CTG arm 51% of infants were delivered because of reduced STV. In the DV p95 arm 34% were delivered because of an abnormal DV and in the DV no A wave arm only 10% of cases were delivered accordingly. The majority of fetuses in the DV arms delivered for safety net criteria were delivered because of spontaneous decelerations. Two year's intact survival was highest in the combined DV arms as compared to the c-CTG arm (p=0.05 when life born, p= 0.21 including fetal death), with no difference between the DV arms. Poorer outcome in the c-CTG arm was restricted to fetuses delivered because of decelerations in the safety net subgroup. Infants delivered because of maternal reasons had the highest birth weight and a non-significant higher intact survival.

Conclusions: In this sub-analysis of fetuses delivered before 32 weeks the majority of infants were delivered for other reasons than according to the allocated CTG or DV monitoring strategy. Since in the DV arms CTG criteria were used as safety net criteria, but in the c-CTG arms no DV safety net criteria were applied, we speculate that the slightly poorer outcome in the CTG arm might be explained by absence of DV data. Optimal timing of delivery of the early IUGR fetus may therefore best be achieved by monitoring them longitudinally with DV and CTG monitoring.

INTRODUCTION

The 2 year outcome data of the TRUFFLE study ('Trial of Umbilical and Fetal Flow in Europe') on outcome of early intrauterine growth restricted (IUGR) fetuses has shown that overall outcome of these fetuses was more favourable than published in the past (1). Timing of delivery was randomized and based on reduced computerized cardiotocograph heart rate short term variation (c-CTG STV), and early or late pulsatility changes in the Ductus Venosus (DV), with safety net criteria in all three intervention strategies.

Impaired outcome (mortality and severe morbidity) did not differ significantly between cases delivered in the three arms of the trial, but data on intact two year's neurological outcome showed that a conservative approach to the timing of delivery by waiting for late DV changes, was associated a better outcome in the survivors as compared to the c-CTG arm. Data were analysed according to intention to treat. However, a considerable proportion of infants was delivered per protocol because of co-called 'safety net' criteria (i.e. severely reduced c-CTG STV, occurrence of spontaneous unprovoked heart rate decelerations, or -after 30 weeks- because of reversed end-diastolic flow velocities (ReDV) in the umbilical artery, without abnormalities in DV flow velocity waveform patterns).

Since cardiotocography is the standard of care in monitoring of IUGR fetuses at risk of impaired intra-uterine condition, c-CTG STV safety net

criteria were established for patients randomized to the DV groups only, while DV was not evaluated in patients randomized to CTG monitoring. Moreover, in all 3 arms of the trial many infants were delivered because of maternal indications or 'other fetal indications' or after 32 weeks of gestation, when delivery occurred according to local protocols and not according to the intention to treat arms of the protocol. Therefore, there is the need for a post-hoc sub-analysis of the TRUFFLE data, especially for infants delivered before 32 weeks to investigate outcome at 2 years in relation to the indications for delivery, to come to a further refinement of management proposals.

METHODS

In the multicenter, unblinded, randomised TRUFFLE study we included women with singleton fetuses at 26–32 weeks of gestation who had very preterm fetal growth restriction (ie low abdominal circumference [<10th percentile] and a high umbilical artery Doppler pulsatility index [>95th percentile]). We randomly allocated women 1:1:1, with randomly sized blocks and stratified by participating center and gestational age (<29 weeks vs ≥29 weeks), to three timing of delivery plans, which differed according to antenatal monitoring strategies: reduced c-CTG STV (STV <3.5 ms at <29 weeks of gestation or STV <4 ms at ≥29 weeks of

gestation), early DV changes (pulsatility index >95th percentile; DV p95), or late DV changes (zero or reversed A wave; DV no A). The safety net c-CTG STV criteria as used in the two DV groups were set considerably lower than in the CTG STV arm, namely ≥26 - <29 weeks if STV < 2.6 and ≥30 - <32 weeks if STV< 3. Joint safety-net criteria for all three randomisation arms included the occurrence of spontaneous decelerations and, after 30 weeks, reversed end-diastolic flow velocities (REDV) in the umbilical artery. The primary outcome was survival without cerebral palsy or neurosensory impairment, or a Bayley III developmental score of more than 85, at 2 years of age. This study was registered with ISRCTN, number 56204499. Between January 2005 and October 2010 503 women were included. Results on direct neonatal and 2 year's outcome have been published before (1,2).

In this post-hoc sub-analysis we included all 310 live born cases of the TRUFFLE study with known outcome at 2 years corrected age, that were delivered before 32 weeks of gestation and 7 fetal deaths. Cases in which it was refrained from intervention before birth because of suspected poor prognosis of the infant (n=5) and 2 cases born with a lethal congenital malformation were not included (2). There were 25 neonatal deaths that were included in the analyses. Most of the analyses were made on the 310 life born cases. However, for comparison with the data from the original TRUFFLE study and where appropriate, data are also shown for 2 year's survivors only. Data were analysed according to the randomization allocation specified for the intervention indication. Data were analysed by

anova or chi-square test as appropriate, using IBM SPSS version 22 (New York, USA).

RESULTS

We included 310 infants born alive before 32 weeks of gestation and 7 fetal deaths. The number of infants born alive according to randomization arm and intervention indication is shown in Table 1. Overall two-third of the infants were delivered according to the specified criteria of the randomization strategies. Slightly more than half of these were delivered because of safety net criteria. The remaining one-third of the study population was delivered because of other off-protocol fetal indications or for maternal indications. In the c-CTG STV arm 54 of 104 infants (51%) were delivered because of reduced STV. In 19 of these cases also decelerations were present. In the DV arms delivery because of a DV>95th centile was the reason for delivery in 34% of cases allocated to that arm and in the DV no A wave arm only 10% of cases were delivered for absent or reversed A-wave. In the latter group over 50% of cases were delivered because of safety net criteria and almost 40% because of other fetal or maternal indications. The 7 fetal deaths occurred in the latter two groups (3 in the DV P95 and 4 in the DV no A wave group).

The Supplementary Table shows gestational age and weight at delivery according to randomisation and indication for delivery. There were no

significant differences between the subgroups, although birth weight was higher in infants delivered for maternal indications (anova, corrected for multiple testing, p=0.02), while gestational age was similar, as compared to the other indication groups.

Outcome at 2 years is shown in Table 2. Overall 83% of live born infants were alive without neurological impairment at 2 years of age. This percentage was 86 for infants delivered in both DV arms and 77 for those delivered in the c-CTG arm (p=0.049 for live born infants if comparing CTG-STV to both DV groups combined).

There were 7 fetal deaths, all in the DV arms. When these deaths were included, intact outcome in the DV arms decreased to 83% (p=0.21 when compared to the CTG arm). Overall the most favourable outcome (92%) occurred in infants delivered because of maternal reasons and this held for all 3 randomization arms (p=0.09 for maternal versus all other indications, excluding fetal death). The lowest incidence of intact outcome (15 of 26; 58%) occurred in the infants in the CTG arm delivered because of safety net criteria. Outcome in this group was significantly poorer than that in the DV arms in which delivery took place on the basis of safety net criteria (p=0.001) . In fact, the poorer outcome in the CTG arm was only due to a poorer outcome in the safety net subgroup. There was no difference in 2 year's outcome between infants that were delivered based on the c-CTG STV criteria (favourable outcome in 44 of 54; 82%), as compared to those delivered based on DV criteria (combined group n=45, favourable outcome in 36, 80%). Results were similar when the 25

neonatal deaths were excluded. In the c-CTG arm 81 of 95 survivors (85%) had a normal neurological outcome, as compared to 176 of 190 (93%) in the combined DV groups (Table 2; p=0.049). The lowest incidence of intact survival occurred in the infants in the c-CTG group delivered because of safety net (15 of 22 (68%), versus 80 of 85 in the combined DV groups (94%)), with no differences in intact survival in case delivery was based on the specified CTG abnormality in the c-CTG arm (44 of 50; 88%) or DV abnormality in the combined DV arms (36 of 41; 88%).

Table 3 shows a sub-division of the safety net criteria according to the randomization arms. Low STV was only a safety net criterion in the DV groups. The other criteria held for all 3 groups (joined criteria). 67% of infants in the safety net group were delivered because of decelerations, 12% because of a low STV, another 15 % because of a combination of both and only 6% because of ReD velocities in the umbilical artery at a gestational age >30 weeks. In the combined DV arms very low STV alone was an indication for delivery in only 14 out of 92 cases (15%) and a very low STV combined with decelerations in another 18 cases (20%); decelerations, with or without low STV were by far the most important determinant for delivery in the DV arms (79%). When delivery was indicated by decelerations then adverse 2-year infant outcome was significantly more frequent in the CTG-STV arm than in the DV-groups (p=0.003). For the other safety net criteria outcome was not significantly different from the overall 2-year infant outcome. (Table 3), although all 7

cases delivered because of ReDV in the umbilical artery after 30 weeks did well.

In 19 of the 54 cases in the CTG arm delivered because of STV criteria (Table 1) also decelerations were present. In a further 24 only decelerations were present (Table 4). In other words, in the CTG arm slightly more fetuses were delivered because of reduced STV than because of decelerations. When leaving out infants delivered because of maternal reasons, ReD flow umbilical artery or off-protocol (i.e.: infants in which there were no recorded CTG or DV abnormalities), altogether 210 infants were delivered because of CTG (STV with decelerations) or DV abnormalities. 165 of these infants were delivered because of CTG and 45 because of DV. Of the infants delivered because of an abnormal DV, 80 % were normal at follow-up (36 of 45) and that held for 83% delivered because of CTG abnormalities. (132 of 165). The only fetuses monitored with both CTG and DV, were those in the two DV arms. Even in these arms twice as many infants (n=87) were delivered because of CTG safety net STV and/or decelerations than because of DV changes (n=45).

Slightly more infants delivered because of CTG were normal at follow-up (75 of 87, 86%; see Table 3), as compared to 80% delivered because of DV (36 of 45; see Table 3). So these data indicate that overall outcome of infants delivered because of CTG changes was at least similar to those delivered because of DV abnormalities. Only in the subgroup, monitored with only c-CTG without DV, outcome was poorer.

DISCUSSION

We have performed a post-hoc sub-analysis of outcome of infants from the TRUFFLE trial who were delivered before 32 weeks of gestation. By doing so we excluded infants born ≥ 32 weeks, who were likely to be at lower risk for impaired outcome and were delivered according to local management criteria and not according to the initial randomization arms (1). This analysis was done to obtain more insight in 2 year's outcome in relation to the actual indications for delivery. A disadvantage of the smaller size of this study was the fact that it was not powered for the questions raised. Conclusions have, therefore, to be drawn with caution.

We found that 2 year's outcome was better in the DV arms as compared to the CTG arm and this is in line with that of the total study population (1). In the original TRUFFLE study primary outcome, i.e.: survival without CP or neurosensory impairment, was not significantly different between the randomization arms, but neurological outcome in survivors was significantly better in the DV no A wave arm as compared to that in the CTG arm, with a trend towards better outcome in the DV>95th centile arm. When specified for the actual indication for delivery (specified CTG or DV abnormality, safety net, other fetal indications, maternal indications) we found no differences between groups in two year's outcome, although those delivered for maternal indication had a non-significantly better outcome. The latter may be related to a significantly higher birth weight at the same age at delivery. In the DV no A group more fetuses were

delivered because of other fetal indications or maternal indications, than in the other arms of the trial. The reason is unclear also since ‘’other fetal indications’’ was not specified enough by the participating centres, apart from cases with partial placental abruption. Waiting for late DV changes may have increased the chance for CTG and other fetal indications to arise.

The better outcome in the DV groups appears initially somewhat difficult to explain given the fact that only 35 and 10% of infants in the DV p95 and DVnoA arm, respectively, were actually delivered because of the allocated DV abnormalities, whereas 52 and 73%, respectively, were delivered because of safety net or other fetal indications. The safety net criteria largely relate to the occurrence of fetal heart rate decelerations or a very reduced STV, i.e. CTG criteria. Altogether more infants in the DV arms were delivered on the basis of CTG safety net criteria than on the basis of an abnormal ductus flow velocity pattern. This implies that in the majority of cases CTG abnormalities (STV and/or decelerations) preceded DV changes. From longitudinal studies it is known that c- CTG STV and DV changes occur more or less at the same time in early IUGR fetuses (3,4). In other words in half of the cases changes in c-CTG STV precede DV changes, but also the opposite holds true. The differences in outcome may, therefore, be related to the study design in which in the DV groups CTG safety net criteria were included, whereas in the CTG arm no DV measurements were obtained. From earlier studies we know that survival in early IUGR is higher if either CTG or DV anomalies had been present as

compared to cases in which both had been abnormal (3,4,5). The poorer outcome in the c-CTG group may therefore, be due to the fact that in this arm in a substantial number of cases both CTG and DV abnormalities had been present. Outcome of fetuses in the CTG arm delivered on the basis of c-CTG STV was identical to that of those delivered in the combined DV arms on the basis of DV abnormalities. It therefore seems essential to include c-CTG when determining the timing of delivery. The significantly poorer outcome in the CTG safety net group delivered because of decelerations, as compared to the DV arms delivered because of this criterion, may well indicate, that absence of knowledge on DV in this subgroup has delayed delivery and has been causal to the poorer outcome. In this context it has to be realised the TRUFFLE study was a comparison of CTG monitoring only, with combined DV and CTG monitoring. Our data stress the importance of monitoring early IUGR fetuses with both CTG and DV. In clinical practice this implies that when monitoring early IUGR fetuses with both techniques, the majority will be delivered because of CTG abnormalities before DV changes occur. DV may therefore be considered the ‘’safety net’’ for CTG monitoring. Such a safety net seems useful, also since the data from the original TRUFFLE trial and the ones from the present sub-analysis have shown that monitoring with CTG alone (without a DV safety net), resulted in a poorer outcome, than when combining both assessment techniques.

STV threshold values for normality may not be clear at this moment. We have defined normal STV as a STV > 3.5 ms in between 26-28 weeks of

gestation and > 4 in between 29 till 32 weeks (1). These threshold values were set taken into account the increase in STV with increasing gestational age (6,7), the absence of fetal acidaemia in case of a STV>4 ms (8) and presence of acidaemia or hypoxaemia in the majority of cases when STV was in between 3.5-4 ms (9). The 2.5th centile of STV in normal populations has been found to be around 4-5 ms in the early third trimester in recordings of variable length (10) or around 4.4-5.4 in CTG recordings of one hour duration (6,7). Therefore, we have used a lower STV threshold values in the present study. However, it is known that fetal heart rate decelerations occur on average at the same time as heart rate variation falls below the normal range (11). Since in the present study slightly more fetuses in the c-CTG arm were delivered on the basis of reduced STV than because of decelerations, it seems unlikely that the STV threshold values in the CTG arm were set too low.

The fact that most fetuses in the DV arms that were delivered on safety net indications were delivered on the basis of decelerations and not on the applied very low STV cut-off values, suggests that the latter values might have been set too low. Therefore, it may be that the same criteria used in the c-CTG arm should be used. The more so since outcome in the c-CTG arm of fetuses delivered according to the specified monitoring parameter, was identical to that of cases delivered in the DV arms because on an abnormal DV. However, the optimal STV cut-off values might be subject to further analysis. The more so, since we had no information on DV in the c-CTG arm and it may therefore be that cases with a reduced STV

according to the c-CTG arm might have been identified by DV abnormalities. It should also be noted that the TRUFFLE STV threshold values were based on one hour CTG recordings. Shorter recordings may give less accurate results (1,2,6). Moreover, possible effects of medication like betamethasone and MgSO4 should be taken into account, since both drugs may reduce STV without affecting the occurrence of decelerations (12-16).

Taken into account the restriction that the present post-hoc sub-analysis was not powered for the questions raised in this paper, the present data suggest some refinement in the management protocol of early IUGR fetuses delivered before 32 weeks of gestation:

1- the optimal timing of delivery may best be achieved by combined longitudinal monitoring using both c-CTG and DV. Given that low STV (<2.6 before 29 weeks and <3 between 30 and 32 weeks) do not appear to be associated with an increase in adverse outcome and it may be safe to wait for such abnormalities to occur as long as DV remains normal.

2- the favourable outcome in the small group of fetuses delivered because of reversed end-diastolic velocities in the umbilical artery after 30 weeks of gestation, supports the use of this criterion after this gestational age.

The data from this sub-analysis based on the actual indications for delivery in infants delivered before 32 weeks of gestation, support those of the whole TRUFFLE study, whereby it has to be realised that almost 2/3rd of cases will be delivered per protocol because of other indications than

CTG in the c-CTG arm, or abnormal DV in the DV arms. This held especially for fetuses allocated to the DV arms. Overall, outcome of IUGR fetuses delivered before 32 weeks, appears to be better than historical data have shown and this is likely to be due to the close multi-modality (Doppler and c-CTG) monitoring.

Table 1: Number of infants born alive (n=310) before 32 weeks of gestation according to randomisation arm (intention to treat) and intervention indication. ReDV: Reversed end-diatolic velocities umbilical artery.

	c-CTG STV	**DV p95**	**DV no A**	**All**
Indication for delivery:				
According to randomization arm:				
- Specified CTG or DV abnormality	54	34	11	99 (32%)
- Safety net, total	26	37	55	118 (38%)
- DV STV safety net - criteria*		11	21	
- Joint safety net criteria:				
Spontaneous decel	24	22	33	
ReDV >30 weeks	2	4	1	
Other fetal indications	9	15	22	46 (15%)
Maternal	16	13	18	47 (15%)
Total	105	99	106	310

*STV<2.6 before 29 weeks and <3 after 29 weeks

Table 2: Number of infants with normal neurological follow-up and total number with known outcome, specified for the indication of delivery and randomization allocation. Selected were only infants delivered before 32 weeks, fetal death due to inevitable poor prognosis and neonatal death due to a lethal anomaly were excluded. In 'total including fetal death' the 7 antepartum deaths were included and in 'total, survivors only' outcome in the 285 survivors is shown.

Indication for delivery	**c-CTG STV**	**DV p95**	**DV no A**	**All**
According to randomization arm:				
- Specified CTG or DV abnormality	44/54 (82%)	26/34 (77%)	10/11 (91%)	80/99 (81%)
- Safety-net	15/26 (58%)	34/37 (92%)	46/55 (84%)	95/118 (81%)
Other fetal indications*	7/9 (78%)	14/15 (93%)	18/22 (82%)	39/46 (85%)
Maternal	15/16 (94%)	11/13 (85%)	17/18 (94%)	43/47 (92%)
Total, liveborn infants with known outcome	81/105 (77%)	85/99 (86%)	91/106 (86%)	257/310 (83%)
Total included fetal death	81/105 (77%)	85/102 (83%)	91/110 (83%)	257/317 (81%)
Total, survivors only	81/95 (85%)	85/93 (91%)	91/97 (94%)	257/285 (90%)

*including 8 cases of partial abruption (2, 2 and 4, respectively; all these infants did well)

Table 3: Sub-division of safety net criteria for randomisation allocation for infants with normal or abnormal neurological follow-up at 2 year's of age and total number. ReDV= reversed end-diastolic velocities umbilical artery

Safety-net indications for delivery	**c-CTG STV**	**DV p95**	**DV no A**	**Total**
Low STV* only	-	2	12	14
Normal outcome	-	1	9	71%
Abnormal outcome	-	1	3	
Decelerations only	24	22	33	79
Normal outcome	13	20	27	76%
Abnormal outcome	11	2	6	
Low STV* with decelerations		9	9	18
Normal outcome		9	9	100%
Abnormal outcome		0	0	
ReDV > 30 weeks	2/2	4/4	1/1	7
Normal outcome	2	4	1	100%
Abnormal outcome	0	0	0	
Total	26	37	55	118
Normal outcome	15 (58%)	34 (92%)	46 (84%)	(81%)
Abnormal outcome	11 (42%)	3 (8%)	9 (16%)	

* DV group only

Supplementary Table: Median gestational age and weight of infants born alive (n=310) before 32 weeks of gestation according to randomisation arm and intervention indication (fetal death excluded).

Indication for delivery	N	c-CTG		DV p95		DVnoA		All	
N		105		99		106		310	
		GA	BW	GA	BW	GA	BW	GA	BW
Specified CTG or DV abnormality	99	29.5 (28.6 to 30.9)	901 (198)	29.4 (28.1 to 30.6)	832 (208)	29.9 (28.6 to 30.9)	851 (275)	29.6 (28.6 to 30.9)	872 (211)
Safety net	118	29.9 (28.4 to 30.6)	832 (175)	30.0 (28.6 to 31.2)	881 (221)	29.9 (28.7 to 30.7)	885 (221)	29.9 (28.6 to 30.9)	872 (211)
Other fetal indications	46	30.0 (28.9 to 30.7)	851 (180)	30.3 (29.0 to 31.1)	932 (183)	30.4 (29.2 to 31.0)	875 (139)	30.3 (29.0 to 31.0)	889 (162)
Maternal	47	29.8 (27.9 to 31.4)	956 (251)	30.0 (29.3 to 30.9)	1019 (258)	29.8 (28.0 to 30.7)	901 (198)	29.9 (28.4 to 31.0)	952 (234)
All	310	29.7 (28.5 to 30.9)	888 (202)	29.9 (28.7 to 31.0)	890 (222)	29.9 (28.7 to 30.8)	882 (207)	29.9 (28.7 to 30.9)	887 (209)

References:

1. Lees CC, Marlow N, van Wassenaer-Leemhuis A, Arabin B, Bilardo CM, Brezinka C, Calvert S, Derks JB, Diemert A, Duvekot JJ, Ferrazzi E, Frusca T, Ganzevoort W, Hecher K, Martinelli P, Ostermayer E, Papageorghiou AT, Schlembach D, Schneider KT, Thilaganathan B, Todros T, Valcamonico A, Visser GHA, Wolf H; TRUFFLE study group. 2 year neurodevelopmental and intermediate perinatal outcomes in infants with very preterm fetal growth restriction (TRUFFLE): a randomised trial. The Lancet 2015; 385:2162–2172.

2. Lees CC, Marlow N, Arabin B, Bilardo CM, Brezinka C, Derks JB, Duvekot J, Frusca T, Diemert A, Ferrazzi E, Ganzevoort W, Hecher K, Martinelli P, Ostermayer E, Papageorghiou AT, Schlembach D, Schneider KT, Thilaganathan B, Todros T, van Wassenaer-Leemhuis A, Valcamonico A, Visser GHA, Wolf H; TRUFFLE Group. Perinatal morbidity and mortality in early-onset fetal growth restriction: cohort outcomes of the trial of randomized umbilical and fetal flow in Europe (TRUFFLE). Ultrasound Obstet Gynaecol, 2013;42:400–408.

3. Hecher K, Bilardo CM, Stigter RH, Ville Y, Hackelöer BJ, Kok HJ, Senat MV, Visser GHA. Monitoring of fetuses with intrauterine growth restriction: a longitudinal study. Ultrasound Obstet Gynecol, 2001;18:564–570.

4. Ferrazzi E, Bozzo M, Rigano S, Bellotti M, Morabito A, Pardi G, Battaglia FC, Galan HL Temporal sequence of abnormal Doppler changes in the peripheral and central circulatory systems of the severely growth-restricted fetus. Ultrasound Obstet Gynecol, 2002; 19:140–146.

5. Baschat AA, Cosmi E, Bilardo CM, Wolf H, Berg C, Rigano S, Germer U, Moyano D, Turan S, Hartung J, Bhide A, Müller T, Bower S, Nicolaides KH, Thilaganathan B,

Gembruch U, Ferrazzi E, Hecher K, Galan HL, Harman CR. Predictors of neonatal outcome in early-onset placental dysfunction Obstet Gynecol. 2007;109:253-261.

6. Nijhuis IJM , ten Hof J, Mulder EJH, Nijhuis JG, Narayan H, Taylor DJ, Visser GHA. Numerical fetal heart rate analysis: nomograms, minimal duration of recording and intrafetal consistency Pren Neon Med 1998;3:314-322.

7. Nijhuis IJM, ten Hof J, Mulder EJH, Nijhuis JG, Narayan H, Taylor DJ, Visser GHA. Fetal heart rate in relation to its variation in normal and growth retarded fetuses. Eur J Obstet Gynecol Reprod Biol. 2000;89:27-33.

8. Dawes GS, Moulden M, Redman CWG. Short-term fetal heart rate variation, decelerations, and umbilical flow velocity waveforms before labor Obstet Gynecol 1992;80:673-678.

9. Ribbert LSM, Snijders RJM, Nicolaides KH, Visser GHA. Relation of fetal blood gases and data from computer-assisted analysis of fetal heart rate patterns in small for gestation fetuses. Brit.J Obstet Gynaecol 1991;98:820-823.

10. Serra V Bellver J, Moulden M, Redman CWG. Computerized analysis of normal fetal heart rate pattern throughout gestation. Ultrasound Obstet Gynecol 2009; 34: 74–79.

11. Snijders RJM, Ribbert LSM**,** Visser GHA, Mulder EJH. Numeric analysis of heart rate variation in intrauterine growth-retarded fetuses: a longitudinal studyAm J Obstet Gynecol 1992;166:22-27.

12. Mulder EJH, Derks JB, Visser GHA. Antenatal corticosteroid therapy and fetal behaviour: a randomised study of the effects of betamethasone and dexamethasone Br J Obstet Gynaecol. 1997;104:1239-1247.

13. Multon O1, Senat MV, Minoui S, Hue MV, Frydman R, Ville Y. Effect of antenatal betamethasone and dexamethasone administration on fetal heart rate variability in growth-retarded fetuses. Fetal Diagn Ther 1997;12:170-177.

14. Frusca T, Soregaroli M, Valcamonico A, Scalvi L, Bonera R, Bianchi U. Effect of betamethasone on computerized cardiotocographic parameters in preterm growth-restricted fetuses with and without cerebral vasodilation. Gynecol Obstet Invest. 2001;52:194–197.

15. Nensi A, De Silva DA, von Dadelszen P, Sawchuck D, Synnes AR, Crane J, Magee LA. Effect of magnesium sulphate on fetal heart rate parameters: a systematic review. J Obstet Gynaecol Can. 2014;36:1055-64.

16. Hiett AK, Devoe LD, Brown HL, Watson J. Effect of magnesium on fetal heart rate variability using computer analysis. Am J Perinatol. 1995;12:259-61.

PRESCRIPTIVE GROWTH CHARTS METHODOLOGY

Growth monitoring is essential in antenatal and newborn care worldwide, as it is for infants and children and it requires comprehensive anthropometric standards.[61]

These tools have been available to evaluate term infants' postnatal growth, but not fetal growth, newborn size, or the postnatal growth of preterm newborn infants. 'Descriptive' *reference* charts, rather than 'prescriptive' *standards*, are used in obstetric and neonatal practice. Standards are preferable because they describe aspirational, biologic norms that are achieved by healthy populations ('how a population should grow'). References, on the other hand, describe the distribution of variables that are observed in unselected samples at a given time and place ('how a population has grown').[62] Furthermore, the higher is the risk of developing perinatal complications the higher is the risk that a reference derived would be influenced by clinical management causing the impossibility of establishing how a baby should have grown.

Reference charts to assess fetal growth, as for example, the popular Hadlock charts of estimated fetal weight, are based on selected populations not reflecting current standards of growth (109 fetuses from Texas in the 1980).[63] One of the issues with the use of reference charts is the large number and limited methodologic quality of the charts that are available to obstetricians and neonatologists. In a series of systematic

reviews, several domains at high risk of bias where found in describing the fetal and neonatal growth.[36, 38, 64, 65] In a review aimed to evaluate the methodology used in studies to create fetal growth measurements substantial heterogeneity was found. 83 published fetal size charts for monitoring growth by ultrasound scanning were selected and there were high risk of bias in pregnancy dating, ultrasound methodology, sample selection, statistical analysis. Even selecting the best quality studies there was such a high variability in centiles reported that the 10th centile for abdominal circumference in one study at a specific gestation was similar to the 50th centile at the same gestation in another study.

Similar findings were found in studies aimed to create pregnancy dating charts by crown-rump length (CRL). In the study presented in this manuscript a systematic review was performed, out of 29 studies selected, 4 studies were reported having the lowest percentage of methodological bias (Appendix 2). Despite the high quality of the four studies selected, using one dating equation rather than another would lead to variability on average between 0 and 4 days in estimating the date of delivery.

Because of the above reason the INTERGROWTH-21st Project was conducted in order to complement the World Health Organization (WHO) Child Growth Standards study,[66] that was derived from healthy newborn infants from populations with few growth-restricting factors whose mothers followed breastfeeding recommendations. This study revealed no significant differences in growth patterns according with the country of origin.

The INTERGROWTH-21st is a multicentre, multiethnic, population-based project done between 2009 and 2014, in eight sites in eight countries: Brazil, Italy, Oman, UK, USA, China, India, Kenya. The INTERGROWTH-21st Project's main aim was to study growth, health, nutrition and neurodevelopment from less than 14 weeks and 0 days of gestation to 2 years of age, so as to produce prescriptive growth standards to complement the existing WHO Child Growth Standards, and to develop a new phenotypic classification of the fetal growth restriction and preterm birth syndromes. The populations that contributed participants to the project were first selected at the geographical level and then at the individual level within each study site. At the population level, urban areas were identified where most deliveries occurred in health facilities. The areas had to be located at an altitude of 1600 m, the area had to have low levels of non-microbiological contamination such as pollution, domestic smoke, radiation, or any other toxic substances. For the fetal component of the study (FGLS) women were recruited with characteristics at low risk of abnormal growth (optimal health, nutrition, education, and socioeconomic status). For example, maternal height (≥153 cm), body-mass index (BMI; ≥18.5 and <30 kg/m2), haemoglobin concentration (≥110 g/L), absence of medical conditions. The study methodology (ultrasound protocol, statistical analysis etc…) was set to have low risks of bias. For the neonatal study 'FGLS-like' newborns were selected to create neonatal charts. Results of the studies showed striking similarity in linear growth in children from the eight sites, thereby justifying pooling data to

construct one international growth standard from the antenatal period to 2 years of age.

The INTERGROWTH-21st Project has produced an integrated set of standards and tools for antenatal and postnatal care, early[67] and late gestational age estimation,[68] first-trimester fetal size,[67] fetal growth[35] and estimated fetal weight standards,[69] symphysis-fundal height standards,[70] pregnancy weight gain standards,[71] newborn size for gestational age,[72] postnatal growth of preterm infants.[73]

Growth velocity charts and infant development at 2 years old standards are also under preparation.[74]

Appendix 2: Pregnancy dating by fetal crown–rump length: a systematic review of charts

DOI: 10.1111/1471-0528.12478
www.bjog.org

Systematic review

Pregnancy dating by fetal crown–rump length: a systematic review of charts

R Napolitano,[a] J Dhami,[a] EO Ohuma,[a] C Ioannou,[a] A Conde-Agudelo,[b] SH Kennedy,[a,c] J Villar,[a,c] AT Papageorghiou[a,c]

[a] Nuffield Department of Obstetrics & Gynaecology, University of Oxford, Oxford, UK [b] Perinatology Research Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland and Detroit, Michigan, USA [c] Oxford Maternal & Perinatal Health Institute, Green Templeton College, University of Oxford, Oxford, UK

Correspondence: Dr AT Papageorghiou, Nuffield Department of Obstetrics & Gynaecology, John Raddiffe Hospital, Oxford, OX3 9DU, UK. Email aris.papageorghiou@obs-gyn.ox.ac.uk

Accepted 2 September 2013. Published Online 6 January 2014.

Background Fetal crown–rump length (CRL) measurement by ultrasound in the first trimester is the standard method for pregnancy dating; however, a multitude of CRL equations to estimate gestational age (GA) are reported in the literature.

Objective To evaluate the methodological quality used in studies reporting CRL equations to estimate GA using a set of predefined criteria.

Search strategy Searches of MEDLINE, EMBASE, and CINAHL databases, from 1948 to 31 January 2011, and secondary reference sources, were performed.

Selection criteria Observational ultrasound studies, where the primary aim was to create equations for GA estimation using a CRL measurement.

Data collection and analysis Included studies were scored against predefined independently agreed methodological criteria: an overall quality score was calculated for each study.

Main results The searches yielded 1142 citations. Two reviewers screened the papers and independently assessed the full-text versions of 29 eligible studies. The highest potential for bias was noted in inclusion and exclusion criteria, and in maternal demographic characteristics. No studies had systematic ultrasound quality-control measures. The four studies with the highest scores (lowest risk of bias) satisfied 18 or more of the 29 criteria; these showed lower variation in GA estimation than the remaining, lower-scoring studies. This was particularly evident at the extremes of GA.

Author's conclusions Considerable methodological heterogeneity and limitations exist in studies reporting CRL equations for estimating GA, and these result in a wide range of estimated GAs for any given CRL; however, when studies with the highest methodological quality are used, this range is reduced.

Keywords Crown–rump length, dating chart, gestational age, pregnancy dating, ultrasound.

Please cite this paper as: Napolitano R, Dhami J, Ohuma EO, Ioannou C, Conde-Agudelo A, Kennedy SH, Villar J, Papageorghiou AT. Pregnancy dating by fetal crown–rump length: a systematic review of charts. BJOG 2014;121:556–565.

Introduction

A dating of pregnancy is important, as up to 30% of women attending an antenatal clinic have uncertain or unreliable menstrual dates.[1] Antenatal care and interventions aimed at improving pregnancy outcome rely on our knowledge of the gestational age (GA).[2] The potential benefits of correct ultrasound dating in the first trimester include: the improved performance of first-trimester screening for chromosomal abnormalities[3]; reducing the number of pregnancies classified as preterm[4]; and the reduced incidence of post-term delivery.[5] It has also been shown that dating the pregnancy in the first rather than the second trimester can lead to a reduction in the number of unnecessary inductions of labour.[6–8]

Crown–rump length (CRL) is the most commonly used fetal measurement for pregnancy dating in the first trimester. The first equation that correlated CRL with GA was reported by Robinson and Fleming in 1975.[9] Several studies proposing and validating different CRL equations have been reported since then.[9–37] Although the original Robinson equation remains widely used, there is variation in practice and no consensus exists on which formula is the most appropriate for pregnancy dating. The prevailing practice for GA assessment is often dictated by operator preference, the default equation setting in the ultrasound

equipment, local hospital policy, or national guidance.[38] The use of different formulae can lead to a discrepancy in GA estimation for the same CRL measurement of several days.[39]

Assessing the accuracy of CRL formulae is difficult, as it requires an independent gold standard for GA estimation: for instance, some studies have compared CRL dates with GAs based on the date of embryo transfer in pregnancies following *in vitro* fertilisation (IVF).[39–41] The problem with this approach is that IVF pregnancies may not be biologically equivalent to spontaneous conceptions. They are associated with higher perinatal risks and congenital malformation rates.[42,43] Therefore, it is possible that early fetal growth in IVF pregnancies is also different to that in spontaneously conceived pregnancies.

Another way to evaluate existing CRL equations is to assess the methodological quality of the studies from which they were derived, in a manner similar to assessing the quality of randomised controlled trials, in order to evaluate the potential sources of bias and to identify the best equations to be used. The objective of this systematic review was therefore to perform such an evaluation.

Methods

This systematic review of observational studies was conducted and reported using the checklist proposed by the Meta-analyses of Observational Studies in Epidemiology (MOOSE) group.[44] Three major electronic databases (MEDLINE, EMBASE, and CINAHL) were systematically searched from 1948 to 31 January 2011. Studies were included if they reported GA estimation from first-trimester CRL measurements using ultrasound. Only articles written in English were considered. Articles were excluded if they did not report a new equation for CRL dating. For instance, reviews and studies performing validation of previously published dating equations were excluded from the review.

A search strategy was formulated in collaboration with a professional information specialist: we searched MEDLINE (OvidSP; 1948–31 January 2011), EMBASE (OvidSP; 1974–31 January 2011), and CINAHL (EbscoHOST; 1980–31 January 2011). A cited reference search was conducted on the Science Citation Index (Web of Knowledge; 1945–31 January 2011) for two seminal papers.[9,45] The following keywords were entered: crown–rump length OR CRL OR fetal OR foetal OR fetus OR foetus AND length OR embryo* AND (pole OR length) AND ultrasound OR ultrasonogra* OR ultra-sonogra* OR sonic* OR scan*AND gestational age OR gestation* OR expected gestation OR expected date* OR date delivery* OR dating delivery OR dating AND (formula or model or chart) OR dating.

Two reviewers (R.N. and J.D.) screened the titles and abstracts of all identified citations, and selected potentially eligible studies. The same reviewers independently assessed the full-text versions of eligible studies, and any disagreements were resolved by consensus or consultation with a third reviewer (A.T.P.). Reference lists of retrieved full-text articles were examined for additional, relevant citations. The flow chart of the literature search, plus the inclusions and exclusions, is presented in Figure 1.

The quality of the studies included was assessed using a modified version of the methods used in our previous evaluation of fetal growth charts.[46] A list of methodological quality criteria (listed in Table S1) was devised *a priori* and divided into two domains: study design (12 criteria); and statistical and reporting methods (17 criteria). Studies were assessed against each criterion within the checklist and were scored as either 0 or 1 if there was a 'high' or 'low' risk of bias, respectively. The overall quality score was defined as the sum of 'low risk of bias' marks (with the range of possible scores being 0–29).

The studies included were reviewed and study details entered into an EXCEL spreadsheet (Microsoft 2007). The methodological quality of each study was then assessed by two obstetricians (R.N. and J.D.) and a medical statistician (E.O.O.). Disagreements were resolved by consensus or consultation with a fourth reviewer (A.T.P.).

Results

The searches yielded 1142 citations, of which 62 were considered for potential inclusion (Figure 1). Thirty-three studies were excluded because they described growth with GA ($n = 16$), assessed or compared existing chart(s) ($n = 3$), were reviews or practice guidelines ($n = 3$), or had other aims ($n = 11$; Table S2).[38–41,45,47–74] Finally, 29 studies providing data on over 11 000 pregnancies met the inclusion criteria and were included in the final analysis (Table 1).[9–37]

The main characteristics and overall quality score for each study included are presented in Table 1. The earliest study was published in 1975 and the latest in 2011.[9,12] Data collection was prospective in 17 studies, retrospective in seven studies, and not reported or uncertain in five studies (Figure 2A; Table S3). Eighteen studies had a cross-sectional design, five were longitudinal, and five were mixed cross-sectional and longitudinal; the design of the remaining study was not reported.

Unselected, low-risk pregnancies were included in only eight (28%) studies. Overall, the demographic characteristics of the populations and any inclusion or exclusion criteria were not well described. Although almost all of the studies reported some of the inclusion/exclusion criteria used in the scoring, in no study were all of them used (Table S1).

The independent method used to assess GA was the first day of the last menstrual period (LMP) in 16 studies. In

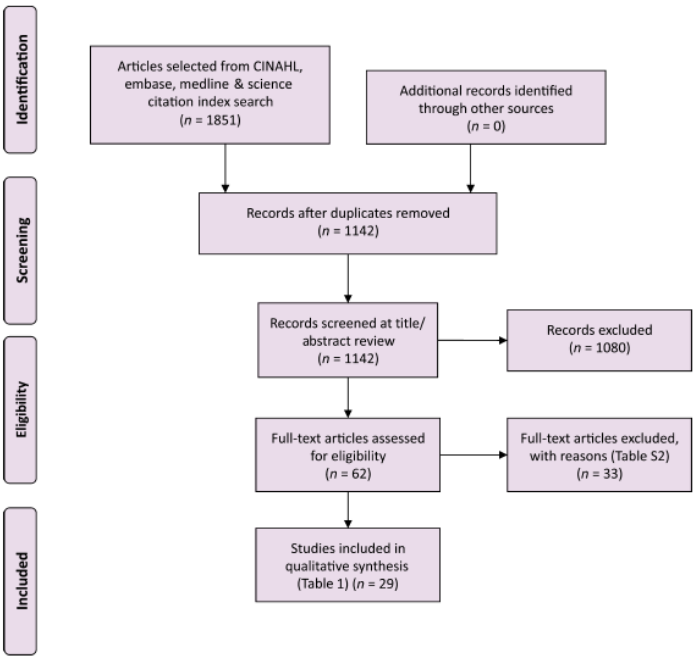


Figure 1. Study selection process.

the remainder, GA was assessed using dates relevant to assisted reproduction, e.g. the date of oocyte retrieval, luteinising hormone surge, embryo transfer, basal body temperature rise, or intrauterine insemination.

Overall, the ultrasound aspects of the studies were well described (Figure 2B; Table S4). Transabdominal ultrasound was used in 12 studies, transvaginal ultrasound was used in five studies, and both were used in six studies. In 14 studies, more than one sonographer obtained scans. The method of image acquisition was well described (26 studies); however, none of the studies employed a comprehensive strategy for ultrasound quality control.

Although all studies had pregnancy dating as their main purpose, the regression equation of GA versus CRL was not reported in four studies. Assessment of the goodness of fit of the proposed equation was performed in 18 studies.

This review has identified four studies that satisfied more than 18 of the 29 quality criteria (Table 2).[9,23,29,34] Figure 3 shows the variation of GA estimation using these four 'best-scoring' charts, compared with the remaining 22 lower-scoring studies (in 3 further studies it was not possible to caclulate the regression equation). It is notable that the best charts are very similar, and that the remaining 22 studies give a wide range of estimated GA, particularly at the extremes of CRL.

Discussion

Main findings

The aim of this review was to investigate the methodology used in studies reporting GA estimation based on CRL measurement. Using a set of 29 criteria the studies were scored as having a low or high risk of bias based on study design, and on the statistical and reporting methods used. This produced a wide range of scores, showing that the quality of the studies was variable (median 15, range 5–21): nine studies scored >15/29 and six studies scored <12/29. We previously used this approach for the case of ultrasound chart creation in fetal biometry.[46] In our view, this is the most scientific way to compare the methodological rigour of studies, improve consistency in fetal growth research, and highlight limitations that should be avoided in future research.

We found that there is considerable heterogeneity and that limitations exist in studies reporting CRL equations for estimating GA. The four studies with the highest scores (lowest risk of bias) satisfied 18 or more of the 29 criteria; these showed lower variation in GA estimation than the remaining, lower scoring studies, and this was particularly evident at the extremes of GA.

Table 1. Main characteristics of studies included

Study	Country	Study period (months or year range)	Data collection	Study design	Conception	GA estimation method	GA range (weeks)	Women (*n*)	Measurements (*n*)	Scanning method	Quality score
Bovicelli et al.[10]	Italy	NR	P	CS	Spont	LMP	7–13	237	NR	NR	11
Campbell et al.[11]	UK	39	R	CS	Spont	LMP	7–14	316	NR	NR	13
Chalouhi et al.[12]	France	2001–2006	P	CS	AC/Spont*	Oocyte retrieval	11–14	331	NR	TV/TA	15
Chevernak et al.[13]	USA	47	R	NR	Mixed	hCG, body temperature, cervical mucous, consistency, endometrial biopsy, IVF	–12	77	NR	NR	5
Daya[14]	Canada	NR	R	CS	AC	Embryo transfer	NR	94	94	TV/TA	15
Drumm et al.[15]	Ireland	15	P	CS	Spont	LMP	6–14	253	253	TA	17
Goldstein and Wolfson[16]	USA	NR	P	CS	Spont	LMP	NR	143	143	TV	12
Grisolia et al.[17]	Italy	NR	P	CS	Spont	LMP	5–12	236	NR	TV	17
Hadlock et al.[18]	USA	9	P	CS	Spont	LMP	5–18	416	NR	TV/TA	16
Izquierdo et al.[19]	USA	12	NR	CS	Spont	LMP	8–12	92	NR	TV	13
Joshi[20]	Nepal	12	P	CS	Spont	LMP	7–14	123	NR	NR	15
Kurjak et al.[21]	Yugoslavia	NR	NR	Mixed	Spont	LMP	6–14	220	390	TA	10
MacGregor et al.[22]	USA	NR	NR	CS	AC	LH, body temperature, follicular collapse on ultrasound	7–13	72	72	TA	14
McLennan and Shulter[23]	Australia	22	P	CS	Mixed	LMP, embryo transfer	5–14	396	NR	TV/TA	18
Nelson[24]	USA	NR	P	CS	Spont	LMP	7–14	83	NR	TA	15
Papaioannou et al.[25]	UK	77	R	Longit	Mixed	Previous CRL dating equation	6–13	4698	NR	TV/TA	16
Pedersen[26]	Denmark	NR	P	Longit	Spont	LMP, body temperature	7–14	101	289	TA	15
Piantelli et al.[27]	Italy	NR	P	Longit	Spont	LMP	7–13	72	NR	TA	14
Robinson et al.[9]	UK	NR	P	CS	Spont	LMP	6–14	334	334	TA	18
Rossavik et al.[28]	USA	NR	P	Longit	Mixed	Embryo transfer, follicular collapse on ultrasound	7–15	35	106	TA	10
Sahota et al.[29]	China	24	P	CS	Spont	LMP	6–15	393	393	NR	21
Selbing[30]	Sweden	NR	P	Longit	Spont	LMP	–9	13	NR	TA	15
Selbing and Fjällbrant[31]	Sweden	NR	P	CS	AC	Insemination, body temperature	NR	24	24	TA	13
Silva et al.[32]	USA	24	P	CS	AC	LH, ovulation induction	5–9	36	36	TV	12
Van de Velde et al.[33]	the Netherlands	NR	P	Mixed	Spont	Body temperature	7–14	60	118	TA	13
Verburg et al.[34]	the Netherlands	46	P	CS	Spont	LMP	6–14	2079	2079	TV/TA	20
Vollebergh et al.[35]	the Netherlands	1981–1986	R	CS	Spont	Body temperature	6–13	47	47	TA	9
Westerway et al.[36]	Australia	22	P	Mixed	NR	LMP	5–14	NR**	NR	NR	10
Wisser et al.[37]	Germany	56	P	Mixed	AC	Oocyte retrieval, insemination	5–14	139	274	TV	15

AC, assisted conception; CS, cross-sectional; hCG, human chorionic gonadotrophin; IVF, *in vitro* fertilisation; L, longitudinal; LH, luteinising hormone; LMP, last menstrual period; Mixed, mixed cross-sectional and longitudinal; NR, not reported or unable to ascertain; P, prospective; R, retrospective; Spont, spontaneous conception; TA, transabdominal; TV, transvaginal.

*Formula derived on assisted-conception pregnancies and validated on spontaneous-conception pregnancies.

**This paper reported a number of charts, and it was not possible to establish the number of women included only for CRL.

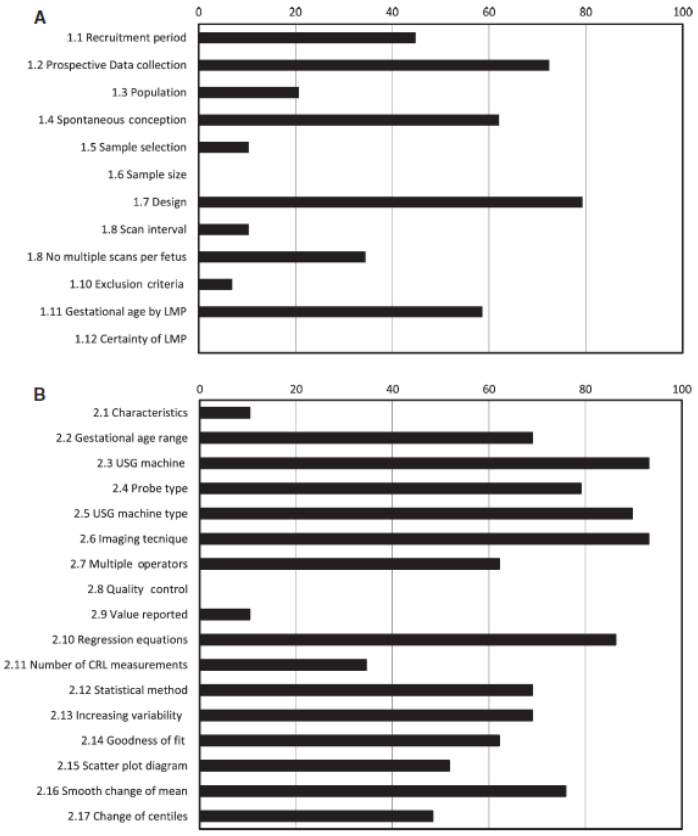


Figure 2. Overall methodological quality of studies included in the review. (A) Study design (percentage of low risk of bias). (B) Reporting and statistical methods (percentage of low risk of bias).

Strengths and limitations of the review

This review has several strengths. The use of a quality score allowed for an objective and quantitative assessment of study methodology: the quality criteria were formulated *a priori*, and were based on a previously published quality checklist used in studies of fetal biometry.[46] One limitation is that an English language restriction was used; however, unlike systematic reviews of randomised controlled trials, where it is imperative that all available evidence is included to estimate the effect of treatment, this is less likely to be a significant limitation in reviews of methodological quality.

Interpretation

There is a debate regarding how best to select samples in research studies that aim to create reference equations of fetal size.

Some authors propose using markers of ovulation or oocyte retrieval/embryo transfer dates in IVF pregnancies as the gold standard[12]; however, uncertainties remain in modelling GA estimation charts in such pregnancies, including the potential time lag between ovulation and conception, differences in early embryonic growth *in vitro*, and, more importantly, differences arising from the selected nature of the population undergoing assisted conception. There are conflicting results about first-trimester fetal growth in IVF pregnancies.[75] Both underestimation and overestimation have been reported between assisted and spontaneous conception populations.[14,22,31,32] Moreover, pregnancies achieved by assisted reproduction may be at higher risk of perinatal complications than normally conceived pregnancies.[76] Finally, we consider that using a sample of women undergoing assisted repro-

Table 2. Gestational age estimation by crown–rump length according to the four studies with the highest quality scores

Fetal crown-rump length (mm)	Gestational age (weeks + days)			
	McLennan and Schluter[23]	Robinson and Fleming[9]*	Sahota et al.[29]*	Verburg et al.[34]
5	6 + 0	6 + 0	6 + 2	6 + 2
10	7 + 0	7 + 1	7 + 2	7 + 4
15	7 + 6	7 + 6	8 + 1	8 + 2
20	8 + 4	8 + 4	8 + 6	9 + 0
25	9 + 1	9 + 2	9 + 3	9 + 4
30	9 + 5	9 + 6	9 + 6	10 + 0
35	10 + 2	10 + 2	10 + 3	10 + 3
40	10 + 5	10 + 6	10 + 6	10 + 6
45	11 + 1	11 + 2	11 + 2	11 + 2
50	11 + 4	11 + 5	11 + 5	11 + 5
55	12 + 0	12 + 1	12 + 1	12 + 0
60	12 + 2	12 + 3	12 + 3	12 + 3
65	12 + 5	12 + 6	12 + 6	12 + 5
70	13 + 0	13 + 1	13 + 1	13 + 0
75	13 + 2	13 + 4	13 + 3	13 + 3
80	13 + 3	13 + 6	13 + 6	13 + 5
85	–	14 + 1	14 + 1	14 + 0
Formula	GA (days) = 32.61967 + (2.62975 × CRL) − [0.42399 × log(CRL) × CRL]	GA (days) = 8.052 × (CRL × 1.037)1/2 + 23.73**	GA (days) = 26.643 + 7.822 × CRL1/2	GA (weeks) = exp[1.4653 + 0.001737 × CRL + 0.2313 × log(CRL)]

*Derived from formula reported.
**Includes correction of 3.7%.

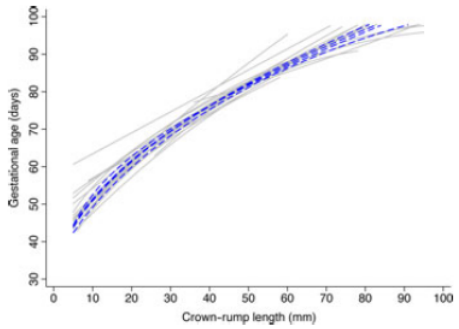


Figure 3. Gestational age charts from the different studies included. The four studies with the highest score for methodological quality (see text) are shown in blue.[9,23,29,34] In three studies the chart is based on data extracted from the tables.[11,26,27] In five cases data were not plotted because two studies did not provide a table or an equation,[10,21] whereas in three studies we were unable to reproduce the figures from the equations given and no tables were provided.[28,30,33]

duction to create dating charts that are then applied to a population of women with spontaneous conception is questionable.

Some authors have proposed using a sample that is as unselected as possible to best represent the underlying population.[77] The problem with this strategy is that a number of pathological conditions may be prevalent, which are likely to affect the reference equations derived. We believe that the purpose of a reference equation is to demonstrate how fetuses should grow (prescriptive), rather than how they do grow (descriptive).[46] Pathological processes, such as smoking,[78] hypertension and pre-eclampsia,[79] maternal disease, abnormal fetal karyotype and congenital anomalies,[80] preterm delivery,[81] and stillbirth,[82] are known to affect fetal size later in pregnancy. There is now evidence to suggest that early fetal growth restriction can be evident as early as the first trimester.[83] Therefore, when producing reference CRL equations, efforts should be made to ensure the sample consists of women at low risk of developing such complications.

A number of studies reporting CRL measurements in the first trimester have been excluded from this review because they attempted to answer a different question: to describe fetal growth in the first trimester.[49,51–55,57,59,61,65,66,68,70,72–74] In some of the studies the authors considered both of these concepts, and such reports were included if GA estimation was one of the stated aims of the study and if a GA

estimation formula was provided, regardless of how the data were analysed.[10,11,15,21] The study by McLennan and Schluter illustrates the differences between the two concepts.[23] The scatter plot of CRL (the independent variable) against GA is reported first, deriving the equation for GA estimation. In the second figure the scatter plot of GA (the independent variable) against CRL is reported. Both charts can be derived from the same population, but differences are seen relating to the analysis performed (i.e. modelling GA estimation rather than fetal size).[84] Sahota et al.[29] elegantly demonstrate how the assessment of size and maturity should not be considered interchangeable, as just 'flipping' a regression can lead to an over- or underestimation of GA, especially at the extremes of the CRL range.

We believe that the recommended study design should be a prospective study of normally conceived, singleton pregnancies, with a pre-defined analysis plan and a prior sample size estimation. Reporting of the demographic characteristics, recruitment period, and estimated GA is essential information for such observational studies; in the present review, <50% of the studies identified satisfied these criteria (Figure 2A, B); in addition, <60% had a prospective design. Most hospitals now routinely collect information using ultrasound software databases, and retrospective analysis of such databases can very easily generate a large sample size. However, retrospective studies are fraught with potential bias as data quality may be variable and the ability to perform continuing ultrasound quality assurance is curtailed.

It has previously been argued that reference studies should be performed by a single operator in order to reduce inter-observer error; however, ultrasound scans in most clinical services are performed by multiple operators, and so variability is inevitable and it would be illogical to ignore it. Reference studies should account for this when using multiple operators, and quality assurance steps should be taken to improve the quality and consistency of measurements, including the standardisation of contributing ultrasonographers.[85]

In the analysis of studies a table of included observations should show how many women were recruited in each GA window. Both the median and variance should be modelled as a function of GA in a way that accounts for the increasing variability with gestation, and should provide smooth percentile curves. A goodness-of-fit assessment, with graphical evaluation of the superimposed centiles, is essential to compare the predictive model. To assess the model, a smooth change of the mean should be represented, superimposed onto the raw data.[12,14,16,17,19,23–25,27,29,31,34,37] While many studies described the statistical method used, more than half did not satisfy the above criteria (Figure 2B).

When adopting reference equations for use in clinical service it is reasonable to choose the publications with the lowest risk of methodological bias (Table 2). This review has identified four studies that satisfied more than 18 of the 29 quality criteria. In Figure 3 it is evident that using any of these four charts leads to very small differences in GA estimation, when compared with the remaining charts.

Conclusion

This systematic review has demonstrated considerable heterogeneity of design in the studies of pregnancy dating by CRL: this results in a wide range of estimated GA for any given CRL. The use of any one of the four studies identified that satisfy most quality criteria lead to very small differences in GA estimations. Consensus in methodology is essential in order to appraise population differences in CRL measurement. A checklist of recommended design is proposed to aid such consensus and potentially reduce the variability in application for clinical practice.

Disclosure of interests

The authors declare they have no conflict of interests.

Contribution to authorship

RN, JD, SHK, JV, and ATP designed the study. JV, ACA, and CI defined the quality criteria *a priori*. RN, JD, and EOO extracted the data. RN, JD, EOO, and ATP scored the studies, analysed the data, interpreted the results, drafted the article, and made the decision to publish. All authors assisted in drafting the article, submitting it, and revising it for important intellectual content, and all authors edited and approved the final version to be published.

Details of ethics approval

No ethical approval was required.

Funding

All authors are part of the INTERGROWTH–21st Project, an international study of fetal growth (www.intergrowth21.org.uk) funded by the Bill & Melinda Gates Foundation to the University of Oxford, for which we are very grateful. C. Ioannou and A. T. Papageorghiou are supported by the Oxford Partnership Comprehensive Biomedical Research Centre, with funding from the Department of Health's National Institute for Health Research (NIHR) Biomedical Research Centres funding scheme.

Acknowledgements

We would like to thank Ms Nia Wyn Roberts, Outreach Librarian, Bodleian Health Care Libraries, for her assistance with the literature search, and Prof. Doug Altman for useful discussions on the assessment of fetal dating charts.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Methodological criteria used to score the studies.

Table S2. Studies excluded after full paper review, because they did not report or develop a new method to estimate GA by CRL.

Table S3. Included studies: study design risk of bias.

Table S4. Included studies: quality scores for reporting and statistical methods. ■

References

1 Pandya PP, Snijders RJ, Psara N, Hilbert L, Nicolaides KH. The prevalence of non-viable pregnancy at 10–13 weeks of gestation. *Ultrasound Obstet Gynecol* 1996;7:170–3.

2 Nguyen TH, Larsen T, Engholm G, Moller H. Increased adverse pregnancy outcomes with unreliable last menstruation. *Obstet Gynecol* 2000;95:867–73.

3 van Heesch PN, Struijk PC, Laudy JA, Steegers EA, Wildschut HI. Estimating the effect of gestational age on test performance of combined first-trimester screening for Down syndrome: a preliminary study. *J Perinat Med* 2010;38:305–9.

4 Blondel B, Morin I, Platt RW, Kramer MS, Usher R, Breart G. Algorithms for combining menstrual and ultrasound estimates of gestational age: consequences for rates of preterm and postterm birth. *BJOG* 2002;109:718–20.

5 Taipale P, Hiilesmaa V. Predicting delivery date by ultrasound and last menstrual period in early gestation. *Obstet Gynecol* 2001;97:189–94.

6 Kalish RB, Thaler HT, Chasen ST, Gupta M, Berman SJ, Rosenwaks Z, et al. First- and second-trimester ultrasound assessment of gestational age. *Am J Obstet Gynecol* 2004;191:975–8.

7 Caughey AB, Nicholson JM, Washington AE. First- vs second-trimester ultrasound: the effect on pregnancy dating and perinatal outcomes. *Am J Obstet Gynecol* 2008;198:703.e1–5; discussion e5-6.

8 Bennett KA, Crane JM, O'Shea P, Lacelle J, Hutchens D, Copel JA. First trimester ultrasound screening is effective in reducing postterm labor induction rates: a randomized controlled trial. *Am J Obstet Gynecol* 2004;190:1077–81.

9 Robinson HP, Fleming JE. A critical evaluation of sonar "crown-rump length" measurements. *Br J Obstet Gynaecol* 1975; 82:702–10.

10 Bovicelli L, Orsini LF, Rizzo N, Calderoni P, Pazzaglia FL, Michelacci L. Estimation of gestational age during the first trimester by real-time measurement of fetal crown-rump length and biparietal diameter. *J Clin Ultrasound* 1981;9:71–5.

11 Campbell S, Warsof SL, Little D, Cooper DJ. Routine ultrasound screening for the prediction of gestational age. *Obstet Gynecol* 1985;65:613–20.

12 Chalouhi GE, Bernard JP, Benoist G, Nasr B, Ville Y, Salomon LJ. A comparison of first trimester measurements for prediction of delivery date. *J Matern Fetal Neonatal Med* 2011;24:51–7.

13 Chervenak FA, Brightman RC, Thornton J, Berkowitz GS, David S. Crown-rump length and serum human chorionic gonadotropin as predictors of gestational age. *Obstet Gynecol* 1986;67:210–3.

14 Daya S. Accuracy of gestational age estimation by means of fetal crown-rump length measurement. *Am J Obstet Gynecol* 1993;168:903–8.

15 Drumm JE, Clinch J, MacKenzie G. The ultrasonic measurement of fetal crown-rump length as a method of assessing gestational age. *Br J Obstet Gynaecol* 1976;83:417–21.

16 Goldstein SR, Wolfson R. Endovaginal ultrasonographic measurement of early embryonic size as a means of assessing gestational age. *J Ultrasound Med* 1994;13:27–31.

17 Grisolia G, Milano K, Pilu G, Banzi C, David C, Gabrielli S, et al. Biometry of early pregnancy with transvaginal sonography. *Ultrasound Obstet Gynecol* 1993;3:403–11.

18 Hadlock FP, Shah YP, Kanon DJ, Lindsey JV. Fetal crown-rump length: reevaluation of relation to menstrual age (5–18 weeks) with high-resolution real-time US. *Radiology* 1992;182:501–5.

19 Izquierdo LA, Kushnir O, Smith JF, Gilson GJ, Chatterjee MS, Qualls C, et al. Evaluation of fetal sonographic measurements in the first trimester by transvaginal sonography. *Gynecol Obstet Invest* 1991;32:206–9.

20 Joshi BR. Estimation of gestational age according to crown-rump length in Nepalese population: a comparison with previously published nomograms. *Iran J Radiol* 2009;6:167–70.

21 Kurjak A, Cecuk S, Breyer B. Prediction of maturity in first trimester of pregnancy by ultrasonic measurement of fetal crown-rump length. *J Clin Ultrasound* 1976;4:83–4.

22 MacGregor SN, Tamura RK, Sabbagha RE, Minogue JP, Gibson ME, Hoffman DI. Underestimation of gestational age by conventional crown-rump length dating curves. *Obstet Gynecol* 1987;70:344–8.

23 McLennan AC, Schluter PJ. Construction of modern Australian first trimester ultrasound dating and growth charts. *J Med Imaging Radiat Oncol* 2008;52:471–9.

24 Nelson LH. Comparison of methods for determining crown-rump measurement by real-time ultrasound. *J Clin Ultrasound* 1981;9:67–70.

25 Papaioannou GI, Syngelaki A, Poon LC, Ross JA, Nicolaides KH. Normal ranges of embryonic length, embryonic heart rate, gestational sac diameter and yolk sac diameter at 6–10 weeks. *Fetal Diagn Ther* 2010;28:207–19.

26 Pedersen JF. Fetal crown-rump length measurement by ultrasound in normal pregnancy. *Br J Obstet Gynaecol* 1982;89:926–30.

27 Piantelli G, Sacchini C, Coltri A, Ludovici G, Paita Y, Gramellini D. Ultrasound dating-curve analysis in the assessment of gestational age. *Clin Exp Obstet Gynecol* 1994;21:108–18.

28 Rossavik IK, Torjusen GO, Gibbons WE. Conceptual age and ultrasound measurements of gestational sac and crown-rump length in *in vitro* fertilization pregnancies. *Fertil Steril* 1988;49: 1012–7.

29 Sahota DS, Leung TY, Leung TN, Chan OK, Lau TK. Fetal crown-rump length and estimation of gestational age in an ethnic Chinese population. *Ultrasound Obstet Gynecol* 2009;33:157–60.

30 Selbing A. Gestational age and ultrasonic measurement of gestational sac, crown-rump length and biparietal diameter during first 15 weeks of pregnancy. *Acta Obstet Gynecol Scand* 1982;61:233–5.

31 Selbing A, Fjallbrant B. Accuracy of conceptual age estimation from fetal crown-rump length. *J Clin Ultrasound* 1984;12:343–6.

32 Silva PD, Mahairas G, Schaper AM, Schauberger CW. Early crown-rump length. A good predictor of gestational age. *J Reprod Med* 1990;35:641–4.

33 van de Velde EH, Broeders GH, Horbach JG, Esser-Rath VW. Estimation of pregnancy duration by means of ultrasonic measurements of the fetal crown-rump length. *Eur J Obstet Gynecol Reprod Biol* 1980;10:225–30.

34 Verburg BO, Steegers EA, De Ridder M, Snijders RJ, Smith E, Hofman A, et al. New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study. *Ultrasound Obstet Gynecol* 2008;31:388–96.

35 Vollebergh JH, Jongsma HW, van Dongen PW. The accuracy of ultrasonic measurement of fetal crown-rump length. *Eur J Obstet Gynecol Reprod Biol* 1989;30:253–6.

36 Westerway SC, Davison A, Cowell S. Ultrasonic fetal measurements: new Australian standards for the new millennium. *Aust N Z J Obstet Gynaecol* 2000;40:297–302.

37 Wisser J, Dirschedl P, Krone S. Estimation of gestational age by transvaginal sonographic measurement of greatest embryonic length in dated human embryos. *Ultrasound Obstet Gynecol* 1994;4:457–62.

38 Loughna P, Chitty L, Evans T, Chudleigh T. Fetal size and dating: charts recommended for clinical obstetric practice. *Ultrasound* 2009;17:160–6.

39 Sladkevicius P, Saltvedt S, Almstrom H, Kublickas M, Grunewald C, Valentin L. Ultrasound dating at 12–14 weeks of gestation. A prospective cross-validation of established dating formulae in in-vitro fertilized pregnancies. *Ultrasound Obstet Gynecol* 2005;26:504–11.

40 Porreco RP, Kaske TI, Henry GP, Shapiro H. Crown-rump length dating of pregnancy: error estimates based on measurement of embryos conceived by *in vitro* fertilization. *J Matern Fetal Med* 1992;1:289–92.

41 Grange G, Pannier E, Goffinet F, Cabrol D, Zorn JR. Dating biometry during the first trimester: accuracy of an every-day practice. *Eur J Obstet Gynecol Reprod Biol* 2000;88:61–4.

42 Helmerhorst FM, Perquin DA, Donker D, Keirse MJ. Perinatal outcome of singletons and twins after assisted conception: a systematic review of controlled studies. *BMJ* 2004;328:261.

43 Hansen M, Bower C, Milne E, de Klerk N, Kurinczuk JJ. Assisted reproductive technologies and the risk of birth defects–a systematic review. *Hum Reprod* 2005;20:328–38.

44 Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;283:2008–12.

45 Robinson HP. Sonar measurement of fetal crown-rump length as means of assessing maturity in first trimester of pregnancy. *Br Med J* 1973;6:28–31.

46 Ioannou C, Talbot K, Ohuma E, Sarris I, Villar J, Conde-Agudelo A, et al. Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size. *BJOG* 2012;119:1425–39.

47 Adam AH, Robinson HP, Dunlop C. A comparison of crown-rump length measurements using a real-time scanner in an antenatal clinic and a conventional B-scanner. *Br J Obstet Gynaecol* 1979;86:521–4.

48 Ahmed AG, Klopper A. Estimation of gestational age by last menstrual period, by ultrasound scan and by SP1 concentration: comparisons with date of delivery. *Br J Obstet Gynaecol* 1986;93:122–7.

49 Blaas HG, Eik-Nes SH, Bremnes JB. The growth of the human embryo. A longitudinal biometric assessment from 7 to 12 weeks of gestation. Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in. *Obstet Gynecol* 1998;12:346–54.

50 Bottomley C, Daemen A, Mukri F, Papageorghiou AT, Kirk E, Pexsters A, et al. Assessing first trimester growth: the influence of ethnic background and maternal age. *Hum Reprod* 2009;24:284–90.

51 Coulam CB, Britten S, Soenksen DM. Early (34–56 days from last menstrual period) ultrasonographic measurements in normal pregnancies. *Hum Reprod* 1996;11:1771–4.

52 Deter RL, Buster JE, Casson PR, Carson SA. Individual growth patterns in the first trimester: evidence for difference in embryonic and fetal growth rates. *Ultrasound Obstet Gynecol* 1999;13:90–8.

53 Dickey RP, Gasser RF, Olar TT, Curole DN, Taylor SN, Matulich EM, et al. The relationship of initial embryo crown–rump length to pregnancy outcome and abortus karyotype based on new growth curves for the 2–31 mm embryo. *Hum Reprod* 1994;9:366–73.

54 Evans J. Fetal crown-rump length values in the first trimester based upon ovulation timing using the luteinizing hormone surge. *Br J Obstet Gynaecol* 1991;98:48–51.

55 Goldstein I, Zimmer EA, Tamir A, Peretz BA, Paldi E. Evaluation of normal gestational sac growth: appearance of embryonic heartbeat and embryo body movements using the transvaginal technique. *Obstet Gynecol* 1991;77:885–8.

56 Goldstein SR. Embryonic ultrasonographic measurements: crown-rump length revisited. *Am J Obstet Gynecol* 1991;165:497–501.

57 Guirgis RR, Alshawaf T, Dave R, Craft IL. Transvaginal crown-rump length measurements of 224 successful pregnancies which resulted from gamete intra-Fallopian transfer or in-vitro fertilization. *Hum Reprod* 1993;8:1933–7.

58 Koornstra G, Wattel E, Exalto N. Crown-rump length measurements revisited. *Eur J Obstet Gynecol Reprod Biol* 1990;35:131–8.

59 Kustermann A, Zorzoli A, Spagnolo D, Nicolini U. Transvaginal sonography for fetal measurement in early pregnancy. *Br J Obstet Gynaecol* 1992;99:38–42.

60 Lagrew DC, Wilson EA, Fried AM. Accuracy of serum human chorionic gonadotropin concentrations and ultrasonic fetal measurements in determining gestational age. *Am J Obstet Gynecol* 1984;149:165–8.

61 Lasser DM, Peisner DB, Vollebergh J, Timor-Tritsch I. First-trimester fetal biometry using transvaginal sonography. *Ultrasound Obstet Gynecol* 1993;3:104–8.

62 Lindgren R, Selbing A, Leander E. Which fetal growth charts should be used? *Acta Obstet Gynecol Scand* 1988;67:683–7.

63 Mills MS. The crown-rump length in early human pregnancy: a reappraisal. *Br J Obstet Gynaecol* 1991;98:946.

64 O'Rahilly R, Muller F. Embryonic length and cerebral landmarks in staged human embryos. *Anat Rec* 1984;209:265–71.

65 Parker AJ, Davies P, Newton JR. Assessment of gestational age of the Asian fetus by the sonar measurement of crown-rump length and biparietal diameter. *Br J Obstet Gynaecol* 1982;89:836–8.

66 Pexsters A, Daemen A, Bottomley C, Van Schoubroeck D, De Catte L, De Moor B, et al. New crown-rump length curve based on over 3500 pregnancies. *Ultrasound Obstet Gynecol* 2010;35:650–5.

67 Reece EA, Gabrielli S, Degennaro N, Hobbins JC. Dating through pregnancy: a measure of growing up. *Obstet Gynecol Surv* 1989;44:544–55.

68 Rosati P, Guariglia L. Transvaginal fetal biometry in early pregnancy. *Early Human Dev* 1997;49:91–6.

69 Sande HA, Reiertsen O. Crown rump length: a comparison of real time scanning and conventional compound scanning. The interindividual measuring variation between two operators. *Ultrasound Med Biol* 1979;5:279–81.

70 Schats R, Van Os HC, Jansen CA, Wladimiroff JW. The crown-rump length in early human pregnancy: a reappraisal. *Br J Obstet Gynaecol* 1991;98:460–2.

71 Smazal SF Jr, Weisman LE, Hopper KD, Ghaed N, Shirts S. Comparative analysis of ultrasonographic methods of gestational age assessment. *J Ultrasound Med* 1983;2:147–50.

72 Tannirandorn Y, Manotaya S, Uerpairojkit B, Tanawattanacharoen S, Wacharaprechanont T, Charoenvidhya D. Transvaginal sonography for fetal crown-rump length measurement in a Thai population. *J Med Assoc Thai* 2001;84:364–70.

73 Verwoerd-Dikkeboom CM, Koning AH, Hop WC, van der Spek PJ, Exalto N, Steegers EA. Innovative virtual reality measurements for embryonic growth and development. *Hum Reprod* 2010;25:1404–10.

74 von Kaisenberg CS, Fritzer E, Kuhling H, Jonat W. Fetal transabdominal biometry at 11–14 weeks of gestation. *Ultrasound Obstet Gynecol* 2002;20:564–74.

75 Dias T, Thilaganathan B. Is first-trimester crown-rump length associated with birthweight? *BJOG* 2012;119:380; author reply 1.

76 Reddy UM, Wapner RJ, Rebar RW, Tasca RJ. Infertility, assisted reproductive technology, and adverse pregnancy outcomes: executive summary of a National Institute of Child Health and Human Development workshop. *Obstet Gynecol* 2007;109:967–77.

77 Altman DG, Chitty LS. Design and analysis of studies to derive charts of fetal size. *Ultrasound Obstet Gynecol* 1993;3:378–84.

78 DiFranza JR, Lew RA. Effect of maternal cigarette smoking on pregnancy complications and sudden infant death syndrome. *J Fam Pract* 1995;40:385–94.

79 Rasmussen S, Irgens LM. The effects of smoking and hypertensive disorders on fetal growth. *BMC Pregnancy Childbirth* 2006;6:16.

80 Hendrix N, Berghella V. Non-placental causes of intrauterine growth restriction. *Semin Perinatol* 2008;32:161–5.

81 Zeitlin J, Ancel PY, Saurel-Cubizolles MJ, Papiernik E. The relationship between intrauterine growth restriction and preterm delivery: an empirical approach using data from a European case-control study. *BJOG* 2000;107:750–8.

82 Stacey T, Thompson JM, Mitchell EA, Ekeroma AJ, Zuccollo JM, McCowan LM. The Auckland Stillbirth study, a case-control study exploring modifiable risk factors for third trimester stillbirth: methods and rationale. *Aust N Z J Obstet Gynaecol* 2011;51:3–8.

83 Salomon LJ. Early fetal growth: concepts and pitfalls. *Ultrasound Obstet Gynecol* 2010;35:385–9.

84 Altman DG, Chitty LS. New charts for ultrasound dating of pregnancy. *Ultrasound Obstet Gynecol* 1997;10:174–91.

85 Sarris I, Ioannou C, Dighe M, Mitidieri A, Oberto M, Qingqing W, et al. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. *Ultrasound Obstet Gynecol* 2011;38:681–7.

Table S1: Methodological criteria used to score the studies. For each criterion one point was given for "low risk of bias" and zero for "high risk of bias".

Domain	Low risk of bias	High risk of bias
1. STUDY DESIGN		
1.1 Recruitment Period	Reported in months	Not reported
1.2 Prospective Data collection	Prospective study and ultrasound data were collected specifically for the purpose of constructing charts for dating of pregnancy	Retrospective study, or data not collected specifically for the purpose of constructing charts (e.g. use of routinely collected data)
1.3 Population	Women were reported as coming from an unselected population, or from a population at low risk of pregnancy complications;	Women did not come from an unselected population; or were selected; or at high risk of pregnancy complications; or not reported.
1.4 Spontaneous Conception	Pregnancies following spontaneous conception	Pregnancies conceived by assisted reproductive technology
1.5 Sample selection	Women are selected either consecutively or at random;	Convenience sampling; arbitrary recruitment; or not reported;
1.6 Sample size	A priori determination / calculation of sample size and justification	Lack of *a priori* sample size determination / calculation and justification
1.7 Design	Clearly either cross-sectional or longitudinal	Not reported Mixture of cross-sectional and longitudinal data
1.8 Method of selecting the gestational ages at which the fetuses were measured **(only for longitudinal studies)**	Interval of measures prospectively pre-specified and justified	Interval of measures not prospectively pre-specified and justified or not reported
1.9 Number of occasions each fetus was measured **(only for cross-sectional studies)**	Each fetus was measured and included only once	Some fetuses were measured and included more than once
1.10 Exclusion criteria	*The study made it clear that women at high risk of pregnancy complications were not included; and that women with abnormal outcome were excluded, i.e. an effort was made to include "normal" outcome as best possible* As a minimum the study population should exclude: - multiple pregnancy - fetuses with congenital structural or chromosomal anomalies - fetal death or miscarriages	*The study population included both low-risk and high-risk pregnancies or women with abnormal outcome were not excluded* Study population that did not exclude fetuses or women with the characteristics previously described

Domain	Low risk of bias	High risk of bias
	- deliveries prior 37 weeks - women with disorders that may affect fetal growth (as a minimum this should exclude women with pre-existing hypertension, diabetes mellitus and smoking)	Exclusions which could have a direct effect on the estimated percentiles, such as fetuses found at birth to be large or small for dates.
1.11 Method of dating pregnancy	Clearly described, by LMP	Not by LMP, or not described clearly or not reported.
1.12 Certainty of LMP assessed	All the following criteria have to be reported: LMP certain Regular menstrual cycles prior to pregnancy No recent use of OCP (1 month or more) No recent breastfeeding (1 month or more) No recent pregnancy (1 month or more)	Any of the criteria were not assessed
2. REPORTING AND STATISTICAL METHODS		
2.1. Characteristics of study population	Presented in a table or clearly described and includes minimum dataset of age, weight and height (or BMI), and parity	Not presented in a table or not clearly described, or does not contain minimum data set
2.2 Gestational age range	Reported	Not reported
2.3 Ultrasound machine(s) used	Clearly specified	Not clearly specified
2.4 Probe Type (Transvaginal or Transabdominal)	Reported	Not reported
2.5 Ultrasound Machine Type (Static or real-time)	Reported	Not reported
2.6 Description of measurement techniques	The study described sufficient and unambiguous details of the measurement techniques used for fetal CRL	The study did not describe sufficient and unambiguous details of the measurement techniques used
2.7 Number of sonographers that took the measurements	Reported	Not reported
2.8 Contains quality control measures	Should include the following - assessment of intra-observer variability - assessment of inter-observer variability - image review - image storage	Does not contain quality control measures
2.9 Report of mean and SD of each measurement and the sample size for each week of gestation.	Presented in a table or clearly described	Not presented in a table or not clearly described
2.10 Report of regression equations for	Reported	Not reported

Domain	Low risk of bias	High risk of bias
the mean (and SD if relevant) for each measurement		
2.11 Number of CRL measurements taken at each scan	More than one measure per fetus per scan	Single measure or not specified
2.12 Statistical methods	Clearly described and applied	Not clearly described and applied
2.13 Assessment of increasing variability of the data with gestation	Performed	Not performed
2.14 Assessment of goodness of fit of the models	A test of goodness-of-fit of the models was reported	Goodness-of-fit of models was not reported
2.15 Scatter diagram of the data with the fitted median/mean superimposed	Study included scatter diagrams of the data with the median/mean superimposed	Study did not include scatter diagrams of the data with the median/mean superimposed
2.16 Change of mean or median across gestational age	Smooth change	Not smooth change
2.17 Change of SD or centile across gestational age	Smooth change	Not smooth change

CRL = crown-rump length; SD = standard deviation; LMP = last menstrual period

Table S2: The following studies were excluded after full paper review, because they did not report or develop a new method to estimate GA by CRL.

Author	Reasons for exclusion
Adam 1979	CRL reproducibility study
Ahmed 1986	Comparison of GA estimation by LMP, ultrasound and Schwangerschaftsprotein1
Blass 1998	Description of growth with gestational age
Bottomley 2009	Effect of maternal characteristics on 1st trimester growth
Coulam 1996	Description of growth with gestational age
Deter 1999	Description of growth with gestational age
Dickey 1994	Description of growth with gestational age
Evans 1991	Description of growth with gestational age
Goldstein S.R. 1991	Review on early embryonic size
Goldstein I. 1991	Description of growth with gestational age
Grange 2000	Assessment / comparison of existing chart(s)
Guirgis 1993	Description of growth with gestational age
Koornstra 1990	Comparison of 2 groups (optimal menstrual history v. change in basal body temperature) using existing charts.
Kustermann 1992	Description of growth with gestational age
Lagrew 1984	To compare GA estimation based on hCG, and CRL/BPD
Lasser 1993	Description of growth with gestational age
Lindgren 1988	Comparison between LMP and CRL GA estimation and birthweight
Loughna 2009	Practice guideline
Mills 1991	CRL measurements comparison between different populations
O'Rahilly 1984	Comparison between greatest embryonic length and CRL
Parker 1982	Description of growth with gestational age and comparison between ethnic groups

Author	Reasons for exclusion
Pexsters 2010	Description of growth with gestational age
Porreco 1992	Assessment / comparison of existing chart(s)
Reece 1989	Review on GA estimation
Robinson 1973	Description of methodological aspects of measuring CRL using ultrasound
Rosati 1997	Description of growth with gestational age
Sande 1979	Inter-observer assessment using two techniques
Schats 1991	Description of growth with gestational age
Sladkevicius 2005	Assessment / comparison of existing chart(s)
Smazal 1983	Comparison between CRL GA confidence interval estimation and BPD
Tannirandorn 2001	Description of growth with gestational age
Verwoerd-Dikkeboom 2010	Description of technique and growth with gestational age
Von Kaisenberg 2002	Description of growth with gestational age

CRL = crown rump length; GA=gestational age; LMP = last menstrual period; hCG = human chorionic gonadotrophin; BPD = biparietal diameter

Table S3 - Included Studies - Study Design Risk of Bias

Study	1.1 Recruitment Period	1.2 Prospective Data collection	1.3 Population	1.4 Spontaneous Conception	1.5 Sample selection	1.6 Sample size	1.7 Design	1.8 Scan Interval	1.9 No multiple scans per fetus	1.10 Exclusion Criteria	1.11 Gestational age by LMP	1.12 Certainty of LMP
Bovicelli et al. 1981[10]	0	1	0	1	0	0	1	0	0	0	1	0
Campbell et al. 1985[11]	1	0	1	1	0	0	0	0	0	0	1	0
Chalouhi et al. 2011[12]	0	1	1	0	1	0	1	0	0	0	0	0
Chevernak et al. 1986[13]	1	0	0	0	0	0	0	0	0	0	0	0
Daya 1993[14]	0	0	0	0	0	0	1	0	1	0	0	0
Drumm et al.1976[15]	1	1	0	1	0	0	1	0	1	0	1	0
Goldstein and Wolfson 1994[16]	0	1	0	1	0	0	1	0	1	0	1	0
Grisolia et al.1993[17]	0	1	0	1	1	0	1	0	0	0	1	0
Hadlock et al.1992[18]	1	1	0	1	0	0	1	0	0	0	1	0
Izquierdo et al.1991[19]	1	0	0	1	0	0	1	0	0	0	1	0
Joshi 2009[20]	1	1	0	1	0	0	1	0	0	0	1	0
Kurjak et al. 1976[21]	0	0	0	1	0	0	0	0	0	0	1	0
MacGregor et al.1987[22]	0	0	0	0	0	0	1	0	1	0	0	0
McLennan and Shulter 2008[23]	1	1	0	0	0	0	1	0	0	1	0	0
Nelson 1981[24]	0	1	0	1	0	0	1	0	0	0	1	0
Papaioannou et al. 2010[25]	1	0	0	0	1	0	1	0	0	0	0	0
Pedersen 1982[26]	0	1	1	1	0	0	1	1	0	0	1	0
Piantelli et al. 1994[27]	0	1	0	1	0	0	1	0	0	0	1	0
Robinson et al. 1975[9]	0	1	0	1	0	0	1	0	1	0	1	0
Rossavik et al.1988[28]	0	1	0	0	0	0	1	0	0	0	0	0
Sahota et al. 2009[29]	0	1	1	1	0	0	1	0	1	1	1	0
Selbing 1982[30]	0	1	0	1	0	0	1	1	0	0	1	0
Selbing and Fjällbrant 1984[31]	0	1	0	0	0	0	1	0	1	0	0	0
Silva et al.1990[32]	1	1	0	0	0	0	1	0	1	0	0	0
Van de Velde et al.1980[33]	0	1	0	1	0	0	0	1	0	0	0	0
Verburg et al. 2008[34]	1	1	1	1	0	0	1	0	1	0	1	0
Vollebergh et al. 1989[35]	0	0	0	1	0	0	1	0	1	0	0	0
Westerway et al. 2000[36]	1	1	1	0	0	0	0	0	0	0	1	0
Wisser et al. 1994[37]	1	1	0	0	0	0	0	0	0	0	0	0

Legend: Score 1 if low risk of bias, 0 if high risk of bias.

Table S4 - Included Studies – Quality scores for reporting and statistical methods

Study	2.1 Characteristics	2.2 Gestational age range	2.3 US machine	2.4 Probe type	2.5 US machine type	2.6 Imaging tecnique	2.7 Number of Operators reported	2.8 Quality Control	2.9 Value reported	2.10 Regression equations	2.11 Number of CRL measurements	2.12 Statistical method	2.13 Increasing variability	2.14 Goodness of fit	2.15 Scatter plot diagram	2.16 Smooth change of mean	2.17 Change of centiles
Bovicelli et al. 1981[10]	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	1	1
Campbell et al. 1985[11]	0	1	1	0	1	1	1	0	1	0	0	0	1	0	0	1	1
Chalouhi et al. 2011[12]	0	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	0
Chevernak et al. 1986[13]	0	0	0	0	0	1	0	0	0	1	0	1	1	0	0	0	0
Daya 1993[14]	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
Drumm et al.1976[15]	0	1	1	1	1	1	1	0	1	1	0	1	1	1	0	0	0
Goldstein and Wolfson 1994[16]	0	0	1	1	1	1	0	0	0	1	0	0	0	0	1	1	0
Grisolia et al.1993[17]	0	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1
Hadlock et al.1992[18]	0	1	1	1	1	1	1	0	0	1	1	1	1	1	0	0	0
Izquierdo et al.1991[19]	0	1	1	1	1	0	0	0	0	1	0	1	0	1	1	1	0
Joshi 2009[20]	0	1	1	0	1	1	1	0	0	1	1	1	1	1	0	0	0
Kurjak et al. 1976[21]	0	1	1	1	1	1	0	0	0	0	0	0	1	0	0	1	1
MacGregor et al.1987[22]	0	1	1	1	1	1	1	0	0	1	1	1	1	1	0	1	0
McLennan and Shulter 2008[23]	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1
Nelson 1981[24]	0	1	1	1	1	1	0	0	0	1	1	0	1	0	1	1	1
Papaioannou et al. 2010[25]	1	1	1	1	0	1	1	0	0	1	0	1	1	1	1	1	1
Pedersen 1982[26]	0	1	1	1	1	1	0	0	0	0	1	0	1	0	0	1	1
Piantelli et al. 1994[27]	0	0	1	1	1	1	1	0	0	1	0	1	0	1	1	1	0
Robinson et al. 1975[9]	0	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1	1
Rossavik et al.1988[28]	0	1	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0
Sahota et al. 2009[29]	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1
Selbing 1982[30]	0	0	1	1	1	1	0	0	0	1	0	1	1	0	1	1	1
Selbing and Fjällbrant 1984[31]	0	0	1	1	1	1	1	0	0	1	0	1	0	1	1	1	0
Silva et al.1990[32]	0	1	1	1	1	1	1	0	0	1	0	0	0	0	0	1	0
Van de Velde et al.1980[33]	0	1	1	1	1	1	1	0	0	1	0	0	1	0	1	1	0
Verburg et al. 2008[34]	1	0	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1
Vollebergh et al. 1989[35]	0	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0
Westerway et al. 2000[36]	0	0	0	0	0	1	1	0	0	1	0	1	0	1	0	0	0
Wisser et al. 1994[37]	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1

Legend: Score 1 if low risk of bias, 0 if high risk of bias.

ULTRASOUND AS A HEALTH CARE TECHNOLOGY

Ultrasound technology is one of the most used heath care technology especially in pregnancy in view of the safety, low costs, diagnostic capability and operators and women acceptability. However, this technique is associated with variability and poor reproducibility especially if it refers to studies where there is lacking of appropriate methodology and quality control.[36-38, 75]

Software for ultrasound images, volumes and video analysis have been developed to assist the operator in the use of ultrasound.[76, 77] Currently, ultrasound assessment still requires a skilled and trained operator but the progress in this area can improve the training and ultrasound performance especially in remote areas and low income settings, reduce the human workload associated with quality control and assist in the diagnostic performance of skilled operators.

Given the benefits of ultrasound imaging such as portability, real time acquisition and low costs compared to other imaging modalities, there is a great potential for this technology to be widely used in resource poor settings. A software for ultrasound video sequences analysis has been developed in collaboration with biomedical engineers (Appendix 3).[78]
Ultrasound video clips were acquired placing the probe at the symphysis and running a sweep to the fundus in 86 pregnant women recruited in the FGLS of the INETRGROWTH-21st Project. The software was developed in

order to identify video sequences containing one of the following four structures: fetal abdomen, heart, head skull, and ‘other fetal structures’ (anatomical structure which did not fall into the other three classes). The applicability of this approach would be an assistance in diagnosing the fetal lie in utero. Fisher vector methodology to develop the software for ultrasound object representation was used in this study. Normally, dense feature extraction is used as many state of art image classification methods, where features of interest are computed on a dense grid rather than sparsely using an interest point detector on an image. Given the characteristics of ultrasound images where the level of shadows and speckles are variable Fisher vector analysis was instead evaluated in this study and compared with a traditional dense feature extraction method (Bag-of-Visual-Words). Fisher vector analysis proved to be more effective in identifying the correct video sequence than traditional methods (98.9% versus 87.1%).

An automatic video acquisition analysis could potentially help in training, standardisation and quality control in basic obstetric ultrasound for evaluating for example in low income countries the fetal presentation and viability. Confirmation of pregnancy viability (presence of fetal cardiac activity) and diagnosis of fetal presentation (head or buttock in the maternal pelvis) are the first essential components of ultrasound assessment in obstetrics. The former is useful in assessing the presence of a viable pregnancy and the latter is essential for labour management. An automated framework for detection of fetal presentation and heartbeat

presence from a predefined “free hand” ultrasound sweep of the maternal abdomen is reported (Appendix 4). The framework consists of a classification regime for a frame by frame categorization of each bidimensional slice of the video. 323 videos of women taking part in the FGLS of the INTERGROWTH-21st Project were acquired in pregnancies beyond 28 weeks of gestation using the previous described approach (Appendix 3). Automatic software analysis is performed using multiple approaches in order to detect correctly one of the 4 video frame sequences of interest (head, heart, abdomen, other ultrasound structure background). The fetal skull, abdomen, and heart were detected with a mean classification accuracy of 83.4%. Furthermore, for the detection of the heartbeat presence an overall classification accuracy of 93.1% was achieved.

Another area where software analysis can assist in fetal ultrasound is the automatic extraction of plane of interest from three-dimensional volumes. An automatic measurements tool with caliper placement on structures of interest can facilitate human workload, can be used for training purposes and quality control. In another study a learning-based solution to automatically determine anatomical views and head and brain structures measurements is reported (Appendix 5).[79]

For the purpose of this study the three recommended planes for routine head biometry and fetal brain structures assessment were analysed: transthalamic (TT), transventricular (TV), transcerebellar (TC) plane.[5]

Recommended measurements for clinical use were selected: biparietal diameter (BPD), occipito-frontal diameter (OFD), and transcerebellar diameter (TCD). The model established anatomical correspondence between the detection of the plane and the placement of calipers for the measurements by the software compared with sonographers (manual annotations). 27 fetal head volumes from the FGLS of the INTERGROWTH-21st Project were analysed by 10 operators: three technical biomedical engineers experts in brain and volume analysis (expert level 1), four sonographers qualified in routine prenatal screening (expert level 2), and three clinicians specialised in fetal medicine trained and standardised in neurosonography (expert level 3). Each operator was asked to extract the appropriate planes and place the calipers for the relevant measurements. The reproducibility between different set of sonographers were reported analysing the angle of rotation and the offset of the plane extracted. Measurements reproducibility was assessed in mm. The automatic software plane detection and measurements reproducibility was better compared with sonographers manual annotation for each one of the expert level group. For example the average angle and offset of interobserver variability for TC plane were 4.71° and 1.2 mm for software analysis and 9.80° and 2.34 mm for manual annotation by the expert level 3 group respectively. Similarly, the average TC and OFD measurements were 0.72 mm and 1.02 mm for software analysis and 1.16 mm and 0.83 mm for manual annotation respectively.

Appendix 3: Fisher vector encoding for detecting objects of interest in ultrasound videos

IEEE 12th International Symposium on Biomedical Imaging (ISBI) 2015. pp. 651-654

FISHER VECTOR ENCODING FOR DETECTING OBJECTS OF INTEREST IN ULTRASOUND VIDEOS

M. A. Maraci⋆ *R. Napolitano*† *A. Papageorghiou*† *J. A. Noble*⋆

⋆ Institute of Biomedical Engineering, Dep. of Engineering Science, University of Oxford, UK
† Nuffield Department of Obstetrics and Gynaecology, University of Oxford, UK.

ABSTRACT

One of the main factors limiting the wider adoption of ultrasound imaging for diagnosis and therapy is requiring highly skilled sonographers. In this paper we consider the challenge of making this technology easier to use for non-experts. Our approach follows some of the recently proposed frameworks that break the process into firstly data acquisition through a simple and task-specific scan protocol followed by using machine learning methodologies to assist non-experts in performing diagnostic tasks. We present an object classification pipeline to identify the fetal skull, heart and abdomen from all the other frames in an ultrasound video, using Fisher vector features. We describe the full proposed method and provide a comparison with a recently proposed approach based on Bag of Visual Words (BoVW) to demonstrate that the new approach is superior in terms of accuracy (98.9% versus 87.1%).

Index Terms— Ultrasound video sweeps, Fisher vector encoding, Bag of Visual Words.

1. INTRODUCTION

Given the benefits of ultrasound imaging such as portability, real-time acquisition and lower-costs compared to other imaging modalities, there is a great potential for this technology to be widely used even in resource-poor settings. However considering the current scanning protocols, guiding the transducer to the correct diagnostic plane as well as interpreting often complex sonography patterns can be difficult for non-experts. In order to address this problem, 3D ultrasound helps to some extent by simplifying acquisition but transforms the problem of finding the diagnostic plane to finding a plane in an ultrasound volume.

The objective of this paper is to automatically identify the frames of interest in an ultrasound video sweep and investigate the merits of pre-processing images on classification accuracy for a small dataset and a relatively large one.

Here we deploy a similar approach to [1, 2, 3] during the acquisition step where an ultrasound sweep, defined by a simple standardized clinical protocol, is used to acquire the data. Machine learning is then used to automatically identify the objects of interest in the video.

Existing approaches to object detection and localization from ultrasound videos have taken various routes. For example in [1, 2], the original video is broken into smaller sequences of shorter length (subsequences). A set of kernel dynamic texture model parameters are then estimated from each sub-sequence and a metric defined between them. The distances were then used to measure the similarity of each subsequences to a set of desired sub-sequences which contain the structures of interest, thus learning a model to identify existence of structures of interest. Recently, [3] proposed an alternative strategy where local phase sift (LP-SIFT) features were employed in a Bag of Visual Words (BoVW) pipeline. In this framework the images are initially pre-processed to create intensity invariant structures using local phase based feature symmetry maps, followed by learning a BoVW encoding of densely computed SIFT features and an SVM classification.

Compared with [3] here we propose a different pre-processing step in addition to using a different feature encoding. Specifically instead of using SIFT features with a BoVW encoding; we use the Fisher Vector (FV) encoding during the learning process with a linear SVM as the classifier. We also demonstrate that pre-processing is only required on smaller data-sets where variability might be more difficult to capture due to a smaller training size. For such cases we propose to use the structured random forest edge detection [4] instead of feature symmetry used in [3]. We illustrate that this new pipeline outperforms the previously published results by a margin of 10%.

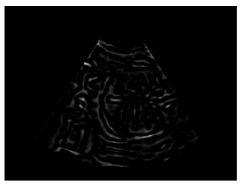
2. MATERIALS AND METHODS

Data 86 clinical 2D fetal ultrasound videos were acquired using a Philips HD9 ultrasound machine with a V7-3 transducer, by a number of experienced obstetricians who were asked to follow a simple standardized scanning protocol. All the participants included in this study were healthy pregnant volunteers at 26 weeks of gestation and over. Data acquisition was covered by appropriate ethics approval. The defined protocol consisted of moving the ultrasound probe from bottom

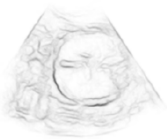
Institute of Biomedical Engineering, Dep. of Engineering Science, Center for Doctoral Training, University of Oxford, UK

978-1-4799-2374-8/15/$31.00 ©2015 IEEE

(a) Gray scale abdomen image



(b) Feature Symmetry Map of (a)



(c) SRF Detected Edges of (a)

Fig. 1: A fetal abdomen (a), and resulting feature symmetry map (b) & edge detection using structured random forest (c).

to top of the mothers abdomen at a uniform speed. The frames of the videos were extracted into the following four classes; fetal abdomen (4,087 frames), fetal heart (1,237 frames), fetal skull (3,806 frames) and the other fetal structures (13,232 frames). The other fetal structures class include any video frame that contained a fetal anatomical structure which did not fall into the other three classes

Structured Random Forests (SRF) Edge Detection The local phase feature symmetry maps are fast to compute and provide an appropriate geometric structure representation of objects but may fail to capture finer details. An alternative method is to use the SRF for edge detection [4] for a more accurate structure representation as illustrated in Fig. 1.

In random decision forests [5] a decision tree $f_t(x)$ classifies a sample $x \in X$ through a recursive process of branching left or right until a leaf node is reached where the output of the tree at the leaf node $y \in Y$ can be a target label or a distribution over all the labels Y. Structured Random Forests is an extension of a random decision forests where x represents an image patch and y encodes the corresponding annotation such as the segmentation mask. We use the implementation of [4] where all the structured labels $y \in Y$ at a given node are mapped to a discrete set of labels $c \in C$, $C = \{1, ..., L\}$, where similar structured labels are assigned to the same discrete label c. Furthermore to simplify the calculation of information gain, [4] proposed mapping Y to an intermediate space Z in order to use the Euclidean distance in Z.

Therefore a set of structured labels $y \in Y$ are mapped to a set of discrete labels $c \in C$, such that labels with similar z are assigned to the same label c. For obtaining the discrete label set C given Z, PCA quantization with $L = 2$ was used where the quantization has been based on the top $log_2(L)$ PCA dimensions. Finally to obtain a single prediction from a set of n labels, the label whose z_L minimizes the sum of distances to all other z_i is selected.

Fisher Vector & Ultrasound Object Representation Dense feature extraction has become an essential part of many state-of-art image classification methods, where features of interest (e.g. SIFT) are computed on a dense grid rather than sparsely using an interest point detector on an image. Given the characteristics of medical ultrasound images where the level of shadows, speckles and attenuation vary between subjects and also depend on the anatomical object being scanned (e.g. fetal heart, abdomen), this approach is also utilized here.

In such a pipeline the image content is described through an aggregation of the dense features encoded into a single feature vector. Here we compare the Bag-of-Visual-Words (BoVW) encoding, as used in [3], which uses a histogram to represent occurrences of vector-quantized descriptors and the Fisher Vector (FV) [6, 7] encoding which aims to reduce the loss of information caused by the vector quantization step in BoVW.

The FV encoding approach works by aggregating a large set of feature vectors, such as the dense SIFT features here, into a high-dimensional space. A common approach, also utilized in this paper, is to fit a parametric generative model such as the Gaussian Mixture Model (GMM) to the features and then encoding the derivatives of the log-likelihood of the model with respect to its parameters. First and second order differences between the dense features and each of the GMM centres can then be captured.

Therefore given $I = (x_1, ..., x_N)$ a set of D dimensional SIFT feature vectors extracted from an image, and $\Theta = (\mu_k, \Sigma_k, \pi_k : k = 1, ..., K)$ the parameters of a Gaussian Mixture Model fitting the distribution of the descriptors, the GMM associates each vector x_i to a mode k in the mixture with a strength given by the posterior probability such that

$$q_{ik} = \frac{exp\Big[-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\Big]}{\Sigma_{t=1}^{K} exp\Big[-\frac{1}{2}(x_i - \mu_t)^T \Sigma_k^{-1}(x_i - \mu_t)\Big]} \quad (1)$$

For each mode k the mean and covariance deviations vectors are defined such that

$$u_{jk} = \frac{1}{N\sqrt{\pi k}} \sum_{i=1}^{N} q_{ik} \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \quad (2)$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi k}} \sum_{i=1}^{N} q_{ik} \left[\left(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \quad (3)$$

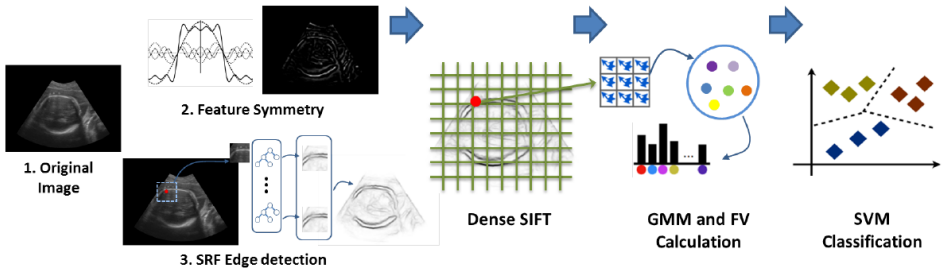


Fig. 2: Method overview: Fisher vector encoding is computed on dense SIFT features followed by the classification using an SVM classifier. Input images can either be the original images(1), feature symmetry maps(2) or the SRF edge maps(3).

where $j = 1, ..., D$ and represents the vector dimensions. The Fisher vector Φ of image I is then constructed by stacking the vectors u_k and v_k for each of the K modes in the Gaussian mixtures,

$$\Phi(I) = [u_1^\intercal, v_1^\intercal, \ldots, u_K^\intercal, v_K^\intercal]^\intercal \quad (4)$$

We used a GMM with $K = 256$ components after reducing the dimensionality of the SIFT descriptors to 80 by using PCA as it has been found to improve the accuracy and decrease the memory footprint of this representation [8]. Finally the VLFeat toolbox [9] is used for computing the features and evaluating accuracy.

3. EVALUATION & RESULTS

Experiments were designed to evaluate the performance gain between the two feature encodings discussed in Section 2. Furthermore a systematic comparison is carried out between the pre-processing step used in [3] and SRF edge detection proposed here on a small clinical dataset as used in [3]. Figure 2 shows an overview of the work.

Experiment One: Pre-processing on a small dataset Here a systematic comparison is carried out between the feature symmetry used in [3] and the structured random forests proposed here as a pre-processing step before object classification. The data and settings reported in [3] were used here. The results illustrate that using the SRF instead of the FS improves classification accuracy as shown in Table 1. The highest accuracy is achieved when the classification has been carried out on the detected edges using the SRF with a mean accuracy of 80.79% compared to 73.95% when FS is used. Eliminating the pre-processing step reduces the accuracy to 67.37%.

Experiment Two: FV & BoVW encoding The BoVW and FV encoding have been tested and evaluated on the data described in section 2. Three experiments were designed to evaluate the classification accuracy of the proposed pipeline with and without any pre-processing. In each case, the model was trained on 850 randomly selected images from each class (3400 frames over all classes) and tested on 350 randomly selected unseen images from each class (1400 frames over all classes). This experiment was repeated five times, each time with a different set of images selected for training and testing. For each experiment the accuracy and mean average precision (mAP) have been calculated and their average over the five repetitions has been reported. The results are summarised in Table 2 (bold indicates best results) and the mAP for one of the repetitions is illustrated in Figure 3.

Initially we have followed [3] to perform a four class classification on FS maps. The mean classification accuracy and the mean mAP achieved are 80.03% and 85.16% respectively using a BoVW approach and 98.10% and 99.49% using the FV. As an alternative to FS, we used structured random forests as outlined in Section 2. For this experiment, the mean classification accuracy and the mAP achieved are 84.07% and 89.12% respectively using a BoVW approach and 97.88% and 99.52% using the FV. Finally classification was performed and evaluated on the original grey scale frames. Again the mean classification accuracy and the mAP achieved are 87.12% and 91.50% respectively using a Bag-of-Visual-Words approach and 98.90% and 99.85% using the Fisher vector.

Conclusion The results obtained in this paper show that a very high accuracy can be achieved at multi-label classification in ultrasound video sequences using FV based classification. We have illustrated that the well-known bag of visual words approach can lose a high level of information during the quantization step. We further demonstrate that learning the edges of the objects of interest using the proposed method in section 2 instead of the feature symmetry that has been previously used in the literature, improves the results for smaller data-sets. The results suggest that for smaller datasets edge structures create a better representation of similar ob-

Table 1: BoVW Classification Results

(a) Results Following [3]

	min/max Accuracy (%)	Mean Accuracy (%)
BoVW_{FS}	71.1/76.3	73.95
BoVW_{SRF}	75.0/86.8	**80.79**
BoVW_{O}	54.0/76.3	67.37

Table Keys

FS	Feature Symmetry
SRF	Structured Random Forests Edges
O	Original Image

Table 2: Classification using BoVW & FV encoding.

	min/max Accuracy (%)	Mean Accuracy (%)	min/max mAP	Mean mAP (%)
BoVW_{FS}	78.4/80.9	80.03	83.9/86.2	85.16
FV_{FS}	97.7/98.7	98.10	99.3/99.8	99.49
BoVW_{SRF}	83.6/84.7	84.07	88.1/89.8	89.12
FV_{SRF}	97.7/98.1	97.88	99.5/99.6	99.52
BoVW_{O}	86.0/88.8	87.12	90.8/92.1	91.50
FV_{O}	98.5/99.4	**98.90**	99.8/99.9	**99.85**

jects than raw intensities. For example in ultrasound images shadows seem to be eliminated on the edge representation of objects and although this results in some information loss it may in fact be a desired outcome on a smaller dataset as it creates a more generalized representation of similar objects. Furthermore we have illustrated that the Fisher vector encoding can efficiently capture the variability between different classes achieve very high accuracy.

4. ACKNOWLEDGMENTS

The authors acknowledge RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation).

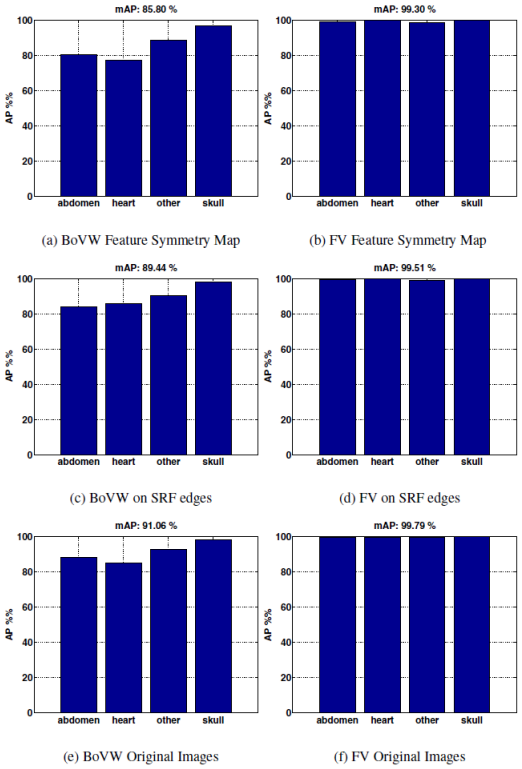
5. REFERENCES

[1] R. Kwitt, N. Vasconcelos, S. Razzaque, and S. Aylward, "Localizing target structures in ultrasound video - a phantom study," *Medical Image Analysis*, vol. 17, no. 7, 2013.

[2] M.A. Maraci, R. Napolitano, A. Papageorghiou, and J.A. Noble, "Object classification in an ultrasound video using lp-sift features," in *MICCAI Workshop on Machine Learning in Medical Imaging*. Springer, 2014.

[3] M.A. Maraci, R. Napolitano, A. Papageorghiou, and J.A. Noble, "Searching for structures of interest in an ultrasound video sequence," in *MICCAI Workshop on Medical Computer Vision*. Springer, 2014.

[4] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1841–1848.



(a) BoVW Feature Symmetry Map (b) FV Feature Symmetry Map

(c) BoVW on SRF edges (d) FV on SRF edges

(e) BoVW Original Images (f) FV Original Images

Fig. 3: mAP comparing BoVW and FV.

[5] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, no. 2–3, pp. 81–227, 2012.

[6] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*. IEEE, 2007, pp. 1–8.

[7] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, pp. 143–156. Springer Berlin Heidelberg, 2010.

[8] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," 2011.

[9] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008.

Appendix 4: A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat

Medical Image Analysis 37 (2017) 22–36

Contents lists available at ScienceDirect

Medical Image Analysis

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

A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat

M.A. Maraci[a,*], C.P. Bridge[a], R. Napolitano[b], A. Papageorghiou[b], J.A. Noble[a]

[a] Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[b] Nuffield Department of Obstetrics and Gynaecology, John Radcliffe Hospital, University of Oxford, Oxford, UK

ARTICLE INFO

Article history:
Received 3 December 2015
Revised 22 December 2016
Accepted 5 January 2017
Available online 10 January 2017

Keywords:
Ultrasound video
Fetal presentation and heartbeat
Machine learning

ABSTRACT

Confirmation of pregnancy viability (presence of fetal cardiac activity) and diagnosis of fetal presentation (head or buttock in the maternal pelvis) are the first essential components of ultrasound assessment in obstetrics. The former is useful in assessing the presence of an on-going pregnancy and the latter is essential for labour management. We propose an automated framework for detection of fetal presentation and heartbeat from a predefined free-hand ultrasound sweep of the maternal abdomen. Our method exploits the presence of key anatomical sonographic image patterns in carefully designed scanning protocols to develop, for the first time, an automated framework allowing novice sonographers to detect fetal breech presentation and heartbeat from an ultrasound sweep. The framework consists of a classification regime for a frame by frame categorization of each 2D slice of the video. The classification scores are then regularized through a conditional random field model, taking into account the temporal relationship between the video frames. Subsequently, if consecutive frames of the fetal heart are detected, a kernelized linear dynamical model is used to identify whether a heartbeat can be detected in the sequence. In a dataset of 323 predefined free-hand videos, covering the mother's abdomen in a straight sweep, the fetal skull, abdomen, and heart were detected with a mean classification accuracy of 83.4%. Furthermore, for the detection of the heartbeat an overall classification accuracy of 93.1% was achieved.

1. Introduction

There have been significant advances in the analysis of ultrasound images in the last decade due in part to increased image quality but also the introduction of modern machine learning into the medical image analysis field (Noble, 2016). Machine learning is arguably very well-suited to recognize sonographic patterns in ultrasound images, which can form the basis of image-based decision-making. By contrast, traditional biomedical image analysis methods can find the dropouts, shadows, and sonographic signatures characteristic of ultrasound images difficult to accommodate, as they are the mapping of anatomy through the ultrasound image formation process. The most successful traditional methods in the literature are model-based methods that use strong geometric models as priors to cope with missing boundaries and artefacts.

Our particular interest is in obstetric ultrasound. The majority of the image analysis literature in this area has focused on automation of fetal biometry measurement for the anomaly scan (taken at 18–22 weeks gestational age). See Challenge US (Rueda et al., 2014) for a recent challenge that looked at a variety of methods and their performances. The anomaly scan is an essential ultrasound screening examination recommended worldwide for the detection of fetal abnormalities and early fetal growth restriction (Tiran, 2005). During a scan, a skilled sonographer acquires and records a number of two dimensional (2D) images of key fetal structures in diagnostic planes, following a standardized clinical protocol (typically a minimum of 6 but often more than 20 images) (Salomon et al., 2011). The goal is to diagnose structural abnormalities and to acquire biometry measurements that are verified against fetal growth charts. Research has looked into automating biometry measurement. For instance, Carneiro et al. (2008) used a discriminative constrained probabilistic boosting tree classifier for the detection and measurement of head, femur and abdominal structures. In their framework the probabilistic boosting tree classifier was trained on a database of key structures, where the nodes of the binary tree are strong classifiers trained using AdaBoost. Rahmatullah et al. (2011b); 2011a) used Adaboost for anatomical object detection in 2D fetal abdominal ultrasound images, where their framework was designed to identify whether the correct abdominal landmarks required for a standard plane

* Corresponding author.
E-mail address: mohammad.maraci@eng.ox.ac.uk (M.A. Maraci).

http://dx.doi.org/10.1016/j.media.2017.01.003

were present. Sun (2012) applied a graph-based approach for automatic detection of the fetal skull, where initially the shortest circular path was detected. An ellipse was then fitted to the shape for finding the skull boundary. Ponomarev et al. (2012) applied a multilevel thresholding approach combined with edge detection and shape-based recognition for segmentation of the fetal skull. Imaduddin et al. (2015) used Haar-like feature with AdaBoost to detect fetal skull and femur. They further applied a Randomized Hough Transform for making biometry measurements. Anto et al. (2015) used a Random Forest to segment a head contour in fetal ultrasound scans that were acquired with a low-cost probe. Perhaps the most similar work to our own is the work of Lei et al. (2015), where densely sampled RootSIFT features were extracted and encoded using Fisher vectors for automatic recognition of fetal facial standard planes.

Three dimensional (3D) ultrasound was introduced in the 1990s as a technology designed to improve clinical workflow. It aimed to replace multiple 2D acquisitions by a single 3D acquisition, followed by standard plane finding in the volume. However, manual standard plane finding is quite time-consuming. This has led to a number of methods being proposed for automated plane finding (Chykeyuk et al., 2014; Yaqub et al., 2015) and some commercial systems now have automated plane finding as an option. However, the images from a 3D acquisition have a different appearance to those of a 2D acquisition and hence can contain different diagnostic value. It remains to be seen whether this type of solution will become accepted clinically. Quantification of 3D fetal ultrasound has, however, shown some promising results. For instance, Yaqub et al. (2011) successfully used Random Forests to perform fetal femur segmentation from 3D ultrasound volumes. This framework was later extended to automatically detect local brain structures in 3D fetal ultrasound images (Yaqub et al., 2012). Namburete et al. (2015) used Regression Forests to estimate the gestational age of a foetus from sonographic signatures in the brain. In the latter case, the accuracy of the method in the third trimester was shown to be higher than the current clinical standard.

It is important to note, though not often discussed, that in both standard 2D and 3D fetal sonography screening a sonographer follows a standardized clinical protocol, which defines criteria for the plane definition - see for instance the ISUOG guidelines for standard plane criteria (Salomon et al., 2011). Standardized 2D planes of acquisition undergo specific quality control to ensure they meet a set of predefined criteria. Moreover, sonographers need to be specifically trained to be able to meet these standards, as training programmes have previously shown to improve measurement variability (Sarris et al., 2011) and image quality (Wanyonyi et al., 2014). We refer to this standardized protocol as a **constrained scan**[1] since all images should have a similar appearance and contain certain anatomical structures, i.e. their appearance is deliberately constrained. These constraints can sometimes assist automated image analysis - for instance in abdominal circumference (AC) measurement, clear visualization of the stomach bubble, umbilical vein and often the spine is expected - but importantly reduce the degrees of variability with respect to the appearance of a general ultrasound scan of the foetus. Constrained scans are widely used in clinical practice, and simplify the image analysis challenge. However they have a key limitation. Acquisition of constrained scans requires a skilled sonographer. For wider adoption of clinical ultrasound in medicine and for uptake of ultrasound in the developing world, the need to acquire constrained scans has to be relaxed in favour of much simpler scanning protocols that a non-expert can readily learn.

[1] In the clinical setting this is referred to as a *standardized scan*.

Encouraging results from observational studies demonstrated that trained and standardized healthcare workers in developing countries can perform as well as qualified sonographers in terms of measurements reproducibility (Rijken et al., 2009). An automatic video acquisition analysis could potentially help in training, standardization and quality control in basic obstetric ultrasound for evaluating the fetal presentation and viability. The simplest scanning protocol to learn would be a linear ultrasound video sweep as illustrated in Fig. 1a. In our work, we propose the use of this type of scan and name it a **predefined free-hand** acquisition protocol. A novice sonographer can readily be trained to acquire data of this type. It is the analysis of data of this kind that we consider in this article. The question is then what useful diagnostic information can be automatically analysed from such videos?

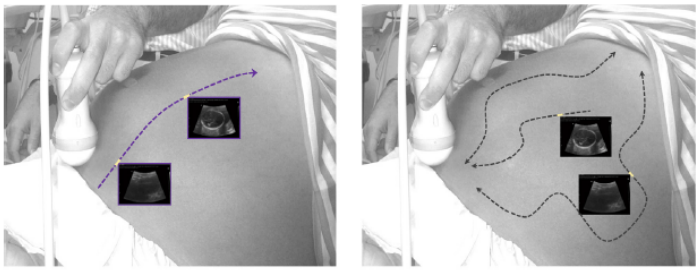
To place our work in perspective, Fig. 2 schematically summarizes how some of the current state-of-the-art literature in fetal ultrasound image analysis maps between the skill needed for acquisition and type of image interpretation and analysis (none, detection & localization, quantification). As can be seen, most image analysis literature is in the lower third of this graph (data acquired by a skilled sonographer). We have included the assisted free-hand works of Kadour and Noble (2009); Kadour et al. (2010); Brown et al. (2013), which use controlled mechanical movement of the probe or subject for elastography on the middle row. These methods generate visualization of ultrasound information and require a small amount of user input to guide probe placement.

In recent years, several methods have been proposed for automatic detection and localization of anatomical fetal structures from ultrasound videos. Linear Dynamical Systems (LDS) were used to localize structures of interest in an ultrasound video obtained from a phantom by Kwitt et al. (2013). In our own work Maraci et al. (2014b), developed independently at around the same time, a method that performed well on clinical ultrasound video sequences was proposed. In that work, the original video is broken into smaller sequences of shorter length, where all sub-sequences have the same length. The dynamics of the sequences are then learned using a linear dynamical system. Identification and classification of the sequences of interest are then based on the similarities between the estimated LDS model parameters.

In an attempt to automatically find the image best representing the fetal abdominal standard plane in a video sequence, Kumar and Shriram (2015) used a method based on the spatial configuration of key anatomical landmarks. In previous works on which the current paper builds, we have investigated the bag of visual words approach with feature symmetry filters (Maraci et al., 2014a) as well as improved Fisher vector (IFV) encoding (Maraci et al., 2015) with a support vector machine (SVM) to identify frames of interest in an ultrasound video.

Finally, CNNs are gaining popularity in medical image analysis including analysis of ultrasound images although they are best suited to very large datasets and balanced data (which we do not have in our application). Chen et al. (2015) used a convolutional neural network (CNN) for standard plane localization of the skull and abdomen from an ultrasound video although the details of acquisition were not stipulated. Gao et al. (2016) have recently used a CNN for partitioning ultrasound video and (Baumgartner et al., 2016) for standard plane detection. We discuss CNNs further in the Discussion section.

To the best of our knowledge, the automation of the task of detecting the fetal presentation and heartbeat from a "predefined free-hand" ultrasound video has not been attempted before. We propose a three-step detection framework for characterizing an ultrasound video obtained from a predefined free-hand constrained scan protocol for pregnancies beyond 28 weeks of gestation. The first step in our method automatically identifies the frames corresponding to the fetal skull, abdomen and the heart. This is used



(a) predefined free-hand scan (b) Typical obstetric scan

Fig. 1. A predefined free-hand scan vs. a typical standardized obstetric scan: (a) Sonographer follows a simple scanning protocol for automated analysis to capture some structure of interest. (b) The sonographer scans over multiple paths to locate the best visual representation of the key structures, where they are saved for further analysis.

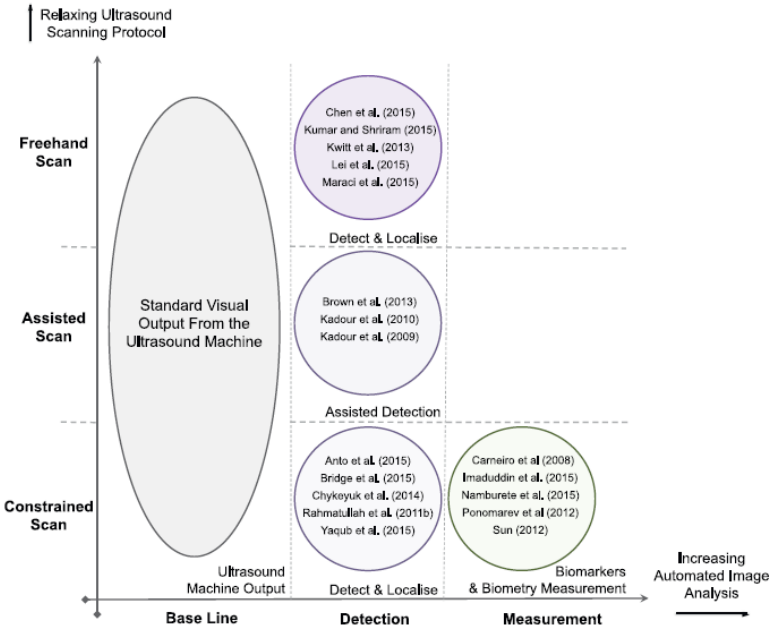


Fig. 2. Ultrasound scan spectrum: Controlled sonographer guidance and automated image analysis increases from left to right, to obtain clinically valid measurements. On the *Y* axis, data acquisition protocol changes from being constrained at the bottom to free-hand on the top.

to infer the fetal presentation as explained in Section 2.2. The second step takes candidate heart frames from the first step to localize the position of the fetal heart as explained in Section 2.3. Finally the dynamics of the fetal heart are modelled from fetal heart frames to identify whether a fetal heart is beating or not. Experiments and results are presented in Section 3, followed by a discussion and conclusion. Earlier versions of some of the component algorithms have been presented in short conference and workshop papers (Maraci et al., 2014b; 2015; Bridge and Noble, 2015). The current paper describes the complete algorithm in detail for the first time and presents substantial experimental evaluation of the complete framework to justify its design.

2. Methods

2.1. Experimental setup

323 videos were acquired from subjects participating in the INTERGROWTH-21ST project (Sarris et al., 2013; Papageorghiou et al., 2014) at the University of Oxford. Data acquisition was carried out using a mid-range ultrasound machine (Philips HD9 with a V7-3 transducer) by a number of experienced obstetricians who were trained for about 10 min to follow the simple scanning protocol. The predefined free-hand ultrasound videos were acquired while moving the transducer from the maternal cervix to the fundus following the longitudinal axis of the uterus as in Fig. 1a. All

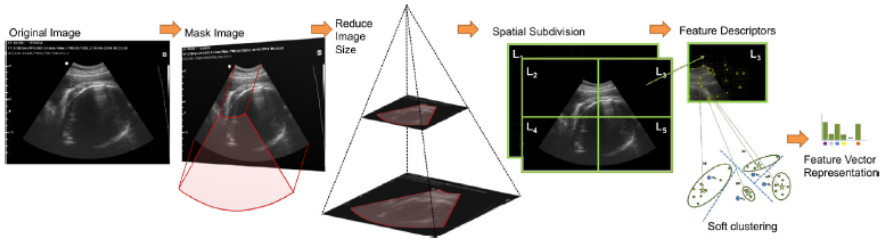


Fig. 3. Steps for feature vector extraction. Preprocessing involves masking each frame and reducing the image size to improve computational cost. Feature extraction (SIFT, rootSIFT, SURF) is then carried out on each image. The extracted features are clustered by a Gaussian mixture model (GMM) and encoded using BoVW, VLAD, or FV encoding.

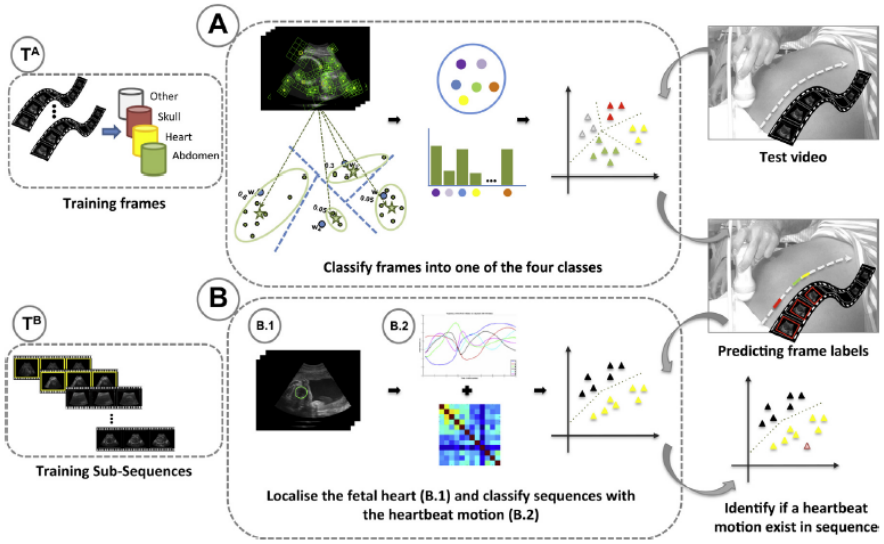


Fig. 4. The main steps of the framework. Given a new training video, all the frames are first classified into "skull", "abdomen", "heart" or "other". If a set of consecutive fetal heart frames are detected in step A, they are further analysed in step B to identify whether a heartbeat can be found. In step T^A, the green colour represents the training dataset of frames corresponding to the fetal abdominal class, yellow indicates the training dataset of fetal hearts, red indicates the dataset of fetal skulls and white indicates the dataset of frames which belong to the "other" class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

foetuses had a normal growth according to international standards (Villar et al., 2014).

The feature extraction process is illustrated in Fig. 3 and the full algorithm we have developed is shown schematically in Fig. 4. In step A, a multi-class discriminative classifier - trained using the data in T^A - is deployed to categorize ultrasound data into four classes of fetal structure: skull, abdomen, heart and "other". At test time, given a set of unseen video frames, a pre-trained classifier is used to categorize the data into the four classes.

Considering the typical heartbeat frequency of a foetus and scan speed (30 fps) employed in this work, it is assumed that heart motion can be captured in at least 30 frames, if it indeed exists. Therefore, if 30 or more consecutive frames are classified as fetal heart frames in step A, they are passed on to step B to identify whether the fetal heart beats or not. In this step, a kernel dynamic texture classifier is trained based on training sequences in T^B, where positive samples in the training set are short videos of a beating fetal heart, and negative samples are sequences that do not contain a fetal heartbeat. Moreover, it is important to note that as the ultrasound videos are intentionally kept simple and general, the likelihood of having a long sequence of a fetal heartbeat is low. In what follows, each of the steps are explained in more detail.

2.2. Step A - Video frame classification

In this subsection we describe the 4-class video frame classification step in more detail. We chose what is sometimes called a hand-crafted feature classification approach rather than deep learning because this class of method is often well-suited to problems defined by relatively small amounts of data (here we had 323 videos), there is significant class imbalance, and the relative richness of features that can represent the problem.

2.2.1. Features

Dense feature extraction, as used in this paper, has become an essential part of many state-of-art image classification methods. In this paper, the speeded up robust feature (SURF) descriptors as described by Bay et al. (2006) and the scale-invariant feature

transform (SIFT) descriptors (Lowe, 2004) were utilized and compared. The SIFT algorithm computes a histogram of local oriented gradients around an interest point and stores the bins in a 128-dimensional vector (8 orientation bins for each of the 4 × 4 location bins). The SURF descriptor describes a distribution of Haar wavelet responses at each interest point neighbourhood and exploits the integral images to estimate Haar features for speed. It results in a 64-dimensional vector and its lower feature dimensions enables a faster detection, at a cost of potentially sacrificing detection accuracy.

In this paper, both features are densely computed over each image with a stride of 4 pixels. Dimensionality reduction of SIFT features using PCA followed by square rooting the feature vectors has been shown to improve classification results (Arandjelović and Zisserman, 2012) in computer vision applications, so we also study its effect on ultrasound images. Additionally, feature vectors are encoded using the traditional bag-of-visual-words (BoVW), vector of locally aggregated descriptors (VLAD) (Jegou et al., 2010), and the improved Fisher vector (FV) (Perronnin et al., 2010) and a comparison between the results of each approach is provided.

The FV encoding approach works by aggregating a large set of feature vectors into a high-dimensional space. A common approach, which we utilize here, is to fit a parametric generative model such as a Gaussian Mixture Model (GMM) to the features and then to encode the derivatives of the log-likelihood of the model with respect to its parameters. First and second order differences between the dense features and each of the GMM centres can then be estimated.

Specifically, given $\mathbf{I} = (\mathbf{x_1}, \ldots, \mathbf{x_N})$ a set of D dimensional SIFT feature vectors extracted from an image, and $\Theta = (\mu_k, \Sigma_k, \pi_k : k = 1, \ldots, K)$ the parameters of a Gaussian Mixture Model fitting the distribution of the descriptors (where K is the number of multivariate Gaussian distributions, μ_k, Σ_k and π_k are the mean, variance and the prior probability of each Gaussian distribution k), the GMM associates each vector $\mathbf{x_i}$ to a mode k in the mixture with a strength given by the posterior probability such that,

$$q_{ik} = \frac{exp\left[-\frac{1}{2}(\mathbf{x_i} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x_i} - \mu_k) \right]}{\Sigma_{t=1}^{K} exp\left[-\frac{1}{2}(\mathbf{x_i} - \mu_t)^T \Sigma_k^{-1} (\mathbf{x_i} - \mu_t) \right]}. \quad (1)$$

Given N SIFT feature vectors, the mean and covariance deviations vectors for each mode k are defined such that,

$$u_{jk} = \frac{1}{N\sqrt{\pi k}} \sum_{i=1}^{N} q_{ik} \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}}, \quad (2)$$

$$v_{jk} = \frac{1}{N\sqrt{2\pi k}} \sum_{i=1}^{N} q_{ik} \left[\left(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right], \quad (3)$$

where $j = 1, \ldots, D$ and represents the vector dimensions. The Fisher vector Φ of image I is then constructed by stacking the vectors u_k and v_k for each of the K modes in the Gaussian mixtures,

$$\Phi(\mathbf{I}) = [\mathbf{u_1}^\intercal, \ldots, \mathbf{u_k}^\intercal, \mathbf{v_1}^\intercal, \ldots, \mathbf{v_k}^\intercal]^\intercal. \quad (4)$$

VLAD encoding utilizes a similar approach to Fisher vectors and encodes a set of local feature descriptors, $\mathbf{I} = (\mathbf{x_1}, \ldots, \mathbf{x_N})$, extracted from an image using a dictionary built using a clustering method such as Gaussian Mixture Models (GMM) or K-means clustering. More formally, let q_{ik} be the strength of the association of data vector $\mathbf{x_i}$ to cluster μ_k, such that $q_{ik} \geq 0$ and $\sum_{k=1}^{K} q_{ik} = 1$, where the association may be either soft (e.g. obtained as the posterior probabilities of the GMM clusters) or hard (e.g. obtained by vector quantization with K-means). VLAD encodes feature $\mathbf{x}$ by considering the residuals $\mathbf{v_k} = \sum_{i=1}^{N} q_{ik}(\mathbf{x_i} - \mu_k)$. The residuals are stacked together to obtain the vector $\hat{\Phi}(\mathbf{I}) = [\ldots, \mathbf{v_k}^\intercal, \ldots]^\intercal$.

2.2.2. Classification

We use support vector machines (SVM) for classification. One of the advantages of using SVM-based classification is that it allows for an efficient use of kernels. The SVM hyperparameters were tuned based on a small sub-set of the data that was randomly selected. Once the optimal parameters were estimated they were used for training the classifier. For non-linear problems, kernel functions allow the data to be projected to a higher-dimensional *feature space*, where a linear model can then be used to classify the data. Moreover, while linear kernels can be highly efficient (Joachims, 2006), non-linear kernels have shown to produce higher classification accuracy (Zhang et al., 2007). It was shown that square rooting SIFT (*sqrt(SIFT/sum(SIFT))*) is similar to using the non-linear Hellinger's kernel in the original input space, without its computational costs (Arandjelović and Zisserman, 2012).

The classifier is trained to categorize the frames into the four classes of fetal skull, fetal abdomen, fetal heart and "other" structures. As the data used in this study consists of an ordered sequence of frames, temporal information is used to regularize the classification results. In order to utilize this temporal information a conditional random field (CRF) graphical model (Lafferty et al., 2001) is constructed, where each frame of the video is represented as a node in the graph. CRFs have previously been successfully used to regularize machine learning for medical image analysis solutions for example in Bauer et al. (2011); McIntosh et al. (2013); Nowozin et al. (2011). Here the classification scores for each frame are converted into probabilities and used as the node potential in the graph. This setting smooths out the classifier scores by taking into account the neighbouring frames, where the joint probability of an assignment to all the nodes f_i (variables) is defined as the normalized product of a set of non-negative potential functions,

$$p(f_1, f_2, \ldots, f_N) = 1/Z \prod_{i=1}^{N} \phi_i(f_i) \prod_{e=1}^{E} \phi_e(f_{e_j}, f_{e_k}). \quad (5)$$

Here we have a potential function for each node i, $\phi_i()$, and edge e, $\phi_e()$, in the graph where (f_{e_j}, f_{e_k}) represents an edge between nodes j and k. As each frame of the video is treated as a node in our graphical model, the node potential $\phi_i()$ for that frame is set to the probability scores obtained from the first step. The edge potential function $\phi_e()$ between any two nodes is the probability of a node transitioning from one state to another and has been empirically set based on the videos in the training dataset. Moreover, Z is the normalization constant to ensure the distribution sums to one over all possible joint configurations of the variables. Finally, the Viterbi (Forney Jr, 1973) algorithm is used to find the most probable classification result for each frame.

2.3. Step B.1 - Locating the fetal heart

The frame classification procedure described in Section 2.2 is able to identify the frames containing the fetal abdomen. In order to assess fetal viability, it is necessary to detect the location of the heart within these frames. This task is complicated by the fact that, when simple sweeps are used, the orientation of the heart relative to the probe is variable and unknown. We therefore chose to make use of rotation invariant detection methods, first introduced for computer vision applications by Liu et al. (2014) and adapted for fetal echocardiography in Bridge and Noble (2015). An extended version of this work can be found in Bridge et al. (2017).

2.3.1. Rotation invariant features

The method for calculating rotation invariant features is based on the use of a set of complex-valued rotation invariant basis functions, b, that have a particular form that is described in polar coordinates (r, θ) by the product of a radial profile $p(r)$ and a Fourier

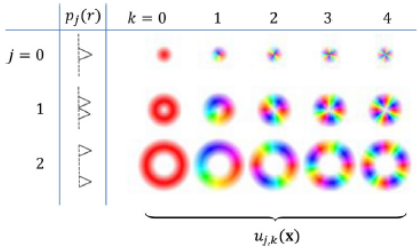


Fig. 5. Set of profiles and basis functions with $J = 3$, $K = 4$ (only $k \geq 0$ displayed). The saturation and hue represent the complex magnitude and argument respectively (Bridge and Noble, 2015).

basis on the angular coordinate, θ (Liu et al., 2014; Bridge and Noble, 2015):

$$b_{j,k}(r,\theta) = p_j(r)e^{ik\theta} \tag{6}$$

for $0 \leq r < R$, $0 \leq \theta < 2\pi$, where j indexes a set of different *radial profiles*, $p_j(r)$. Notice that, while the form of the radial profile is general, the angular part of the separable form of the basis function must be the Fourier basis in order to achieve the desired rotation invariance. Fig. 5 illustrates a set of basis functions.

In order to use the framework with a vector-valued image representation (such as the gradient), $v(\mathbf{p})$, we must first express the orientation of the vectors in a Fourier orientation histogram. This represents an orientation histogram as a truncated set of M Fourier series coefficients, rather than a set of discrete bins. The mth coefficient at image position $\mathbf{p}$ is:

$$c_m(\mathbf{p}) = \|v(\mathbf{p})\| e^{-im\arg(v(\mathbf{p}))}, \quad m = 0, 1, \ldots, M. \tag{7}$$

When working with discrete images, we sample the basis functions on a rectangular grid and use them as a filter kernel on the Fourier histogram images. One feature with parameters j, k, m describing the window centred at position $(\mathbf{x})$ is therefore given by,

$$D_{j,k,m}(\mathbf{p}) = b_{j,k}(\mathbf{p}) * c_m(\mathbf{p}), \tag{8}$$

and a complete description of a window is built up by using a number of such basis functions. In our experiments, parameters j, k, m are empirically set to 6,4,4 respectively. As a result of the shift property of the Fourier series, the complex magnitude of the resulting features are analytically invariant to the orientation of the underlying image.

2.3.2. Support vector classification

For classification of each window as heart or non-heart we use a linear SVM classifier with the rotation invariant features from Section 2.3.1 as input. At test time, each pixel in each frame is assigned a classification score as the output of the SVM classifier, reflecting the probability of belonging to a heart. For each image location, we simply sum these scores across frames to get a total score for each pixel, and choose the pixel with the highest score to be the location of the centre of the heart.

Note that the location only needs to be approximate as the next step uses ROIs around the estimated location for heartbeat detection and the accuracy of location is not the critical factor.

2.4. Step B.2 - Detecting the fetal heartbeat

Once a minimum of 30 consecutive video frames of the fetal heart are identified and the heart is localized using the procedures described in Section 2.3, they are compiled together to form a short video sub-sequence. Our goal is to derive a model of a heartbeat in terms of the intensity patterns in this video subsequence. Moreover, we investigate the accuracy of the framework when learning the dynamics on heart ROI compared to the entire image. The positive training examples used are short video sequences of a fetal heartbeat and the negative training sequences are short video sequences that do not contain a heartbeat, randomly extracted from the videos in dataset. Therefore, the classifier is trained to perform a binary classification to identify whether any given sequence, during test time, contains the correct dynamics and motion that corresponds to a fetal heartbeat.

Specifically, the feature trajectories (dynamics) of the sequences of frames, $\{y_t\}_{t=1}^{T}$, are modelled as the output of a linear dynamical system (LDS). We follow Doretto et al. (2003) for the system identification of the model, which models pixel intensities in each frame as the output of a LDS. However as opposed to using the raw pixel intensities, we instead use the output of frames filtered by a feature symmetry filter (Rajpoot et al., 2009), which produces a contrast invariant representation of structures on each frame. In this model, the appearance of each video frame is determined through the observed variable and the motion and dynamics in the video over a given time is determined through the hidden-state variables, which are sampled from a Gauss–Markov process. Furthermore, the observed frame at any given time can be constructed from a linear combination of the hidden state variables. Therefore, given an ultrasound sequence $\mathbf{S}$ of T video frames, let $\mathbf{S} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$, where $\mathbf{y}_t \in \mathbb{R}^d$ refers to the frame observed at time t. It is assumed that at each time instant t, a noisy version of the image can be measured, $\mathbf{y}(t) = \mathbf{S}(t) + \mathbf{w}(t)$, where $\mathbf{w}(t) \in \mathbb{R}^d$ is an independent and identically distributed (i.i.d.) sequence drawn from a known distribution, resulting in a positive measured sequence $\mathbf{y}(t) \in \mathbb{R}^d$ for $t = 1, \ldots, T$. The evolution of an LDS can be modelled as:

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t \\ \mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{w}_t \end{cases} \tag{9}$$

Here $\mathbf{x}_t \in \mathbb{R}^n$ is the state of the LDS and $\mathbf{y}_t \in \mathbb{R}^d$ are the observed pixel intensities at time t. Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the state transition matrix that describes the dynamics of the state evolution and $\mathbf{C} \in \mathbb{R}^{d\times n}$ is the output matrix.

In a linear system such as Eq. 9, the output matrix $\mathbf{C}$ can be estimated via singular value decomposition of the observation matrix $\mathbf{S}$, where $\mathbf{C}$ can be restricted to the N largest eigenvalues. However, here a non-linear model known as a Kernel Dynamic Texture (KDT) (Chan and Vasconcelos, 2007; Kwitt et al., 2013) is used where the evolution of the hidden states of the model are kept linear. In order to capture the dynamics of the video the output matrix $\mathbf{C}$ is replaced by a non-linear observation function $C : \mathbb{R}^n \to \mathbb{R}^d$. Therefore given the same ordered ultrasound sequence $\mathbf{S} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ and a kernel function $k(\mathbf{y}_1, \mathbf{y}_2)$ with associated feature transformation $< \phi(\mathbf{y}_1), \phi(\mathbf{y}_2) >$, the cth eigenvector $\mathbf{kv}_c$ can be used to obtain the cth kernel principal component in the feature space:

$$\mathbf{kv}_c = \sum_{i=1}^{T} \alpha_{i,c}\phi(\mathbf{y}_i) \tag{10}$$

where $\alpha_{i,c}$ represents the ith component of the cth weight vector and $\boldsymbol{\alpha}_c = \frac{1}{\sqrt{\lambda_c}}\mathbf{kv}_c$, assuming the eigenvectors are sorted in descending order of the eigenvalues $\{\lambda_c\}_{c=1}^{T}$. Here λ_c and $\mathbf{kv}_c$ are the cth largest eigenvalue and eigenvector of the kernel matrix $\mathbf{K}$ respectively. Finally the sequence of hidden states $\mathbf{X}$ and the state transition matrix $\mathbf{A}$ can be estimated as

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\alpha}^{\intercal}\mathbf{K} \\ \mathbf{A} &= [\mathbf{x}_1, \ldots, \mathbf{x}_{T-1}][\mathbf{x}_0, \ldots, \mathbf{x}_{T-2}]^{\dagger} \end{aligned} \tag{11}$$

2.4.1. Distance metrics

Given a KDT model estimate for each sub-sequence, a suitable metric now needs to be defined to assess similarity between any two sub-sequence models. Prior work on comparison metrics of LDSs range from metrics based on subspace angles between the observability subspaces of the systems (De Cock and De Moor, 2000) to metrics based on the Binet–Cauchy kernels (Vishwanathan et al., 2007; Bissacco et al., 2007) and finally metrics based on the KL-divergence of the probability distributions of the stochastic processes (Chan and Vasconcelos, 2005). A full comparison of these classes of metrics is outside the scope of this paper. However, Chaudhry and Vidal (2009) illustrated on a number of applications that the similarity metrics based on the Martin Distance and Binet–Cauchy maximum singular values kernel produced the best results. Furthermore, we have previously shown (Maraci et al., 2014b) that the Binet–Cauchy maximum singular values kernel produced superior results on medical ultrasound data.

The Binet–Cauchy (BC) singular value kernel (Vishwanathan et al., 2007) used in this paper can be explained as an extension of the BC trace kernel. Given two LDS models $\mathbf{M}_1$ and $\mathbf{M}_2$ (represented by their model parameters), with corresponding sequences $\{y_t^{M_i}\}_{t=1}^{T}$ that have the same underlying noise process, the trace kernel for the two non-linear dynamical systems (NLDS) is as follows:

$$K_{NLDS}(\mathbf{M}_1, \mathbf{M}_2) := \mathbb{E}_{v,w}\left[\sum_{t=0}^{\infty} \lambda_t k(\mathbf{y}_t^1, \mathbf{y}_t^2)\right], \tag{12}$$

where λ is a weight factor between 0 and 1 and $\mathbb{E}$ is the expected value of the infinite sum of inner products with respect to the joint probability distribution of v_t and w_t. Thus the BC trace kernel for NLDS is defined as

$$K_{NLDS}(\mathbf{M}_1, \mathbf{M}_2) = \mathbf{x}_0^{\intercal}\tilde{\mathbf{P}}\mathbf{x}_0^{\intercal} + \frac{\lambda}{1-\lambda} trace(\mathbf{Q}\tilde{\mathbf{P}} + \mathbf{R}) \tag{13}$$

where $\mathbf{x}_0$ is the initial state of the system, $\tilde{\mathbf{P}} = \sum_{t=0}^{\infty} \lambda_t (\mathbf{A}_t^1)^{\intercal}\mathbf{F}\mathbf{A}_t^2$, $\mathbf{F}$ is the inner product matrix between all the Kernel PCA (KPCA) components and $\mathbf{Q}$ and $\mathbf{R}$ are the state and output covariance matrices. To remove the dependency on the initial state and the noise process, Chaudhry et al. (2009) proposed the BC maximum singular value kernel for NLDSs as $K_{NLDS}^{\sigma} = \max \sigma(\tilde{\mathbf{P}})$, where σ represents the singular values kernel, to take into account only the dynamics of the NLDS. Thus a normalized kernel of the similarity values can be constructed such that $K(\mathbf{M}_1, \mathbf{M}_2) = 1$ if $\mathbf{M}_1 = \mathbf{M}_2$ as

$$K(\mathbf{M}_1, \mathbf{M}_2) = \frac{K(\mathbf{M}_1, \mathbf{M}_2)}{\sqrt{K(\mathbf{M}_1, \mathbf{M}_1), K(\mathbf{M}_2, \mathbf{M}_2)}} \tag{14}$$

A distance between two sequences with LDS parameters $\mathbf{M}_1$ and $\mathbf{M}_2$ can now be computed as $d(\mathbf{M}_1, \mathbf{M}_2) = 2(1 - K(\mathbf{M}_1, \mathbf{M}_2))$. This distance is then used as the kernel in an SVM classification framework to identify the presence or absence of a fetal heartbeat in the sequence.

3. Results and discussion

Experiments were designed to evaluate the accuracy of the proposed framework. The first experiment evaluated the accuracy of the frame classification task, including the use of different low-level features and SVM kernels. The second experiment compared detecting heartbeats on full images with first localizing a region of interest (ROI) around the heart and only detecting the heartbeat from analysis of heartbeat ROIs.

Table 1
Mean classification accuracies. The most accurate configurations for the different features and encoding strategies, over the number of words. Breakdown plots are shown in Appendix A.

No. Words ↓	$SIFT_{L1}$	$SIFT_{L5}$	$rootSIFT_{L1}$	$rootSIFT_{L5}$	$SURF_{L5}$
10	74.9	79.5	72	78.4	72.5
20	77.4	80.3	78	79.1	74.8
40	81.5	81.7	80.3	81	77.7
60	82.2	83	81	82.3	77.5
80	80.7	82	81.8	82.8	77.9
100	81.5	82.5	82.7	83	78.5

3.1. Classifying video frames

In order to ensure training and test data are independent, a five-fold cross validation procedure was implemented for training the classifier. At each training step, the model was trained on four fifths of the videos (260 videos) and tested on the unseen one fifth (65 videos). RootSIFT, SIFT, and SURF descriptors were calculated on each 240 × 320 image with a stride of 4 pixels. Moreover, SIFT and rootSIFT descriptors were calculated at 9 different scales with a scaling factor of $\sqrt{2}$. As the ultrasound data is only visible within the ultrasound fan (field of view), all feature descriptors were only computed within the bounding box around this region to avoid calculating redundant information. The number of words (GMM clusters) was varied from 10 clusters to 100 and the three feature encoding techniques (BoVW, VLAD, and FV) were utilized to encode each image before classification. Furthermore, to investigate the effect of SIFT feature dimensionality reduction on classification accuracy, on the experiments in which the number of words exceed 60, SIFT features were decorated and reduced in dimensions from 128D to 40D and 20D, as suggested in Chatfield et al. (2011). The effect of subdividing the data into 1 × 1 and 2 × 2 spatial subdivisions were also investigated. Here, for each tile, the corresponding features were computed and stacked as one. In addition, the effect of using larger SIFT descriptor patches was investigated, by varying the SIFT patch size (8 × 8, 16 × 16, 32 × 32, and 64 × 64 pixels). Finally, the accuracy of using different SVM kernels, namely the linear kernel, Hellinger kernel and the χ^2 kernel was investigated. Fig. 6summarizes the classification accuracies for each of the four classes, where the number of words vary from 10 to 100. The experiments were repeated using the BoVW, illustrated using black colour, VLAD illustrated using blue, and FV encoding illustrated using cyan. For the experiments where PCA is used to reduce feature dimensions, the classification accuracy is illustrated using a single point on the plot, indicated by the same colour and pointer shape. Finally, $L1$ indicates no spatial subdivisions and $L5$ indicates the additional 4 spatial subdivisions. As can be seen, generally, increasing the number of words up to 80, improves classification results but a further increase to 100 does not show any substantial improvement to the classification accuracies. Regardless of the use of spatial subdivision, the skull and "other" classes have the best performance and fetal heart is the class that performs the worst. Moreover, Figs A.11, A.12, and A.13 show the mean classification accuracies where the number of GMM clusters have varied between 10 and 100 utilizing different features (SIFT, rootSIFT, and SURF), feature encoding techniques (BoVW, VLAD, FV), and SVM kernels (linear, Hellinger, χ^2). Similarly, Figs A.14, A.15, and A.16 show the mean average precision for the same experiments. A summary of the most accurate configurations are illustrated in Tables 1 and 2.

As can be seen from Figs A.11 and A.12, the gain in accuracy is only marginal when the number of GMMs is extended beyond eighty clusters. Moreover, when the SIFT and rootSIFT features are used, the χ^2 SVM kernel results in the worst performance com-

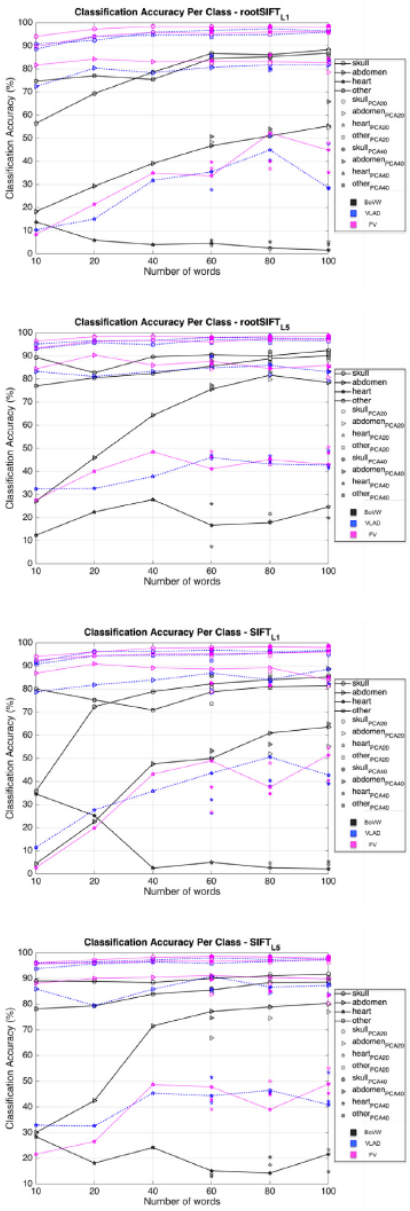


Fig. 6. Classification accuracies for skull, abdomen, heart, and other structures. Individual class accuracies are reported for SIFT and rootSIFT features, while varying the encoder type (BoVW, VLAD, FV) and number of words. A SVM classifier with Hellinger kernel is utilized.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Mean average precision. The most accurate configurations for the different features and encoding strategies, over the number of words. Breakdown plots are shown in Appendix A.

No. Words ↓	SIFT$_{L1}$	SIFT$_{L5}$	rootSIFT$_{L1}$	rootSIFT$_{L5}$	SURF$_{L5}$
10	82.4	88.5	81.2	88.1	81.6
20	87.3	90.3	87.9	90.1	82.8
40	89.6	91.8	89.2	90.9	85.8
60	90.9	92.5	90.4	92.3	86.3
80	91.2	92.8	90.9	92.7	86
100	92	93.3	92.2	93.4	87.2

pared to using the linear or the Hellinger's kernel. Generally the results for the other two kernels are very similar with minor improvements when the Hellinger kernel is used. For the SURF features, the choice of the kernel does not have a dramatic effect on the accuracy.

As for the different feature encodings, FV encoding demonstrates a small gain in accuracy across the experiments compared to using VLAD or BovW. PCA dimensionality reduction can also provide a small boost to the accuracy when 80 or more words are used. It is interesting to note that the use of spatial subdivision boosts the classification results as smaller structures can be better learned when the feature descriptor is augmented by spatial subdivision. Figs. A.11, A.12, A.13, A.14, A.15, and A.16 show the mean classification and mean average precisions results. Moreover, the most accurate cofiguration in these figures are summarized in Tables 1 and 2.

It is worth noting that using PCA to reduce the feature dimensionality to 20 reduces the classification performance results in all experiments. However, for rootSIFT descriptors, using PCA to reduce the feature dimensionality to 40 improves the performance when spatial subdivisions are used, but reduces the performance when spatial subdivisions are not used. This can be explained by the fact that rootSIFT$_{L1}$ descriptors capture less information compared to rootSIFT$_{L5}$, and thus reducing their dimensions even further results in loss of vital discriminative information. It is interesting to note that SIFT$_{L1}$ and SIFT$_{L5}$ features illustrate a similar effect when PCA is applied to reduce feature dimensionality, whereby a decreased classification accuracy is observed. Figs. A.13 and A.16 show plots of the mean classification accuracy and mean average precision that have been obtained using the SURF feature descriptor. Similar to the previous experiments, FV encoding results in better performance compared to the other encoding techniques. In addition, the fetal skull and "other" classes have the best classification performance and the fetal heart is shown to be the most challenging class. In order to better understand the effect of the three encoding techniques, the SVM kernels, and PCA dimensionality reduction on the classification accuracy of each individual class, an experiment was conduction where the number of GMM clusters was fixed to 80. The results are shown in Fig. 7. It is interesting to note that FV and VLAD encoding mainly boost up the classification performance for *skull* and *other* class. Their performance for these two classes are very similar. FV encoding results in slightly better accuracy for the *abdomen* class.

To investigate the effect of various rootSIFT descriptor patch sizes, the number of words was fixed to 80 and PCA was used to reduce the feature dimensions to 40. The mean accuracy and mean average precision (mAP) have been calculated for this experimented and are summarized in Table 3 (bold indicates best results). As can be seen, larger patch sizes improve classification accuracy, especially the results for the fetal heart. This is an intuitive finding as the fetal heart is a small structure and larger descriptors can capture a better representation of structures of interest in relation to other anatomical structures. Moreover, applying

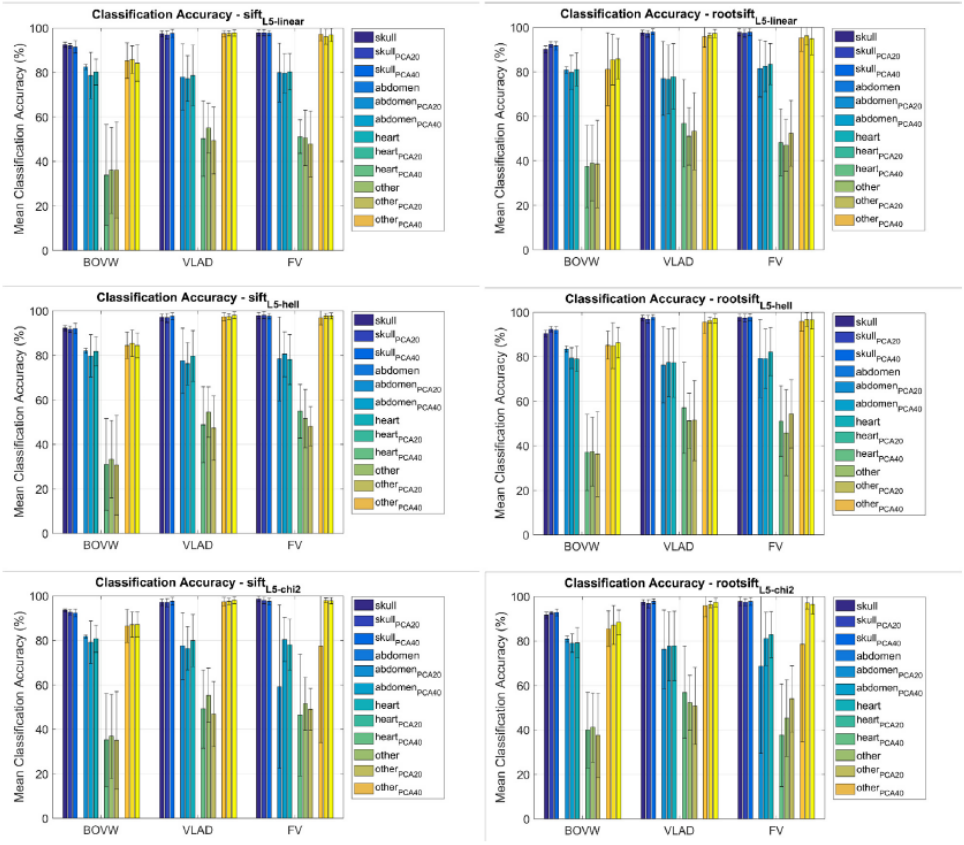


Fig. 7. Mean classification accuracy for all four classes individually. The number of GMM is set to 80 clusters to allow for a performance comparison on each class, using the three encoding techniques, with and without PCA dimensionality reduction.

Table 3
Video frame classification results (Step A). The effect of increasing the rootSIFT descriptor size from an 8 × 8 to a 16 × 16 patch is shown, where the number of the GMM is set to 80 modes and the PCA is used to reduce feature dimensions to 40.

rootSIFT Patch Size	Skull class. Accuracy (%)	Abdomen class. Accuracy (%)	Heart class. Accuracy (%)	Other class. Accuracy (%)	Mean Accuracy (%)	Mean ave. Precision (%)
8 × 8	94.38	89.17	35.21	94.58	78.33	90.01
16 × 16	95.83	91.04	50.42	97.71	83.75	93.37
32 × 32	96.46	92.08	60.63	97.92	86.77	94.75
64 × 64	96.25	86.13	72.92	97.92	**87.55**	95.25

the CRF model to the classification scores regularizes the results and eliminates sudden peaks. This is illustrated in Fig. 8, where the top bar illustrates the raw classification scores. As can be seen, there are a number of frames that have been incorrectly classified as *other* and *abdomen* but applying the CRF regularizes the results as illustrated on the bottom bar. The results show that CRF regularization makes the choice of rootSIFT and SIFT features less significant because it levels their accuracy to a similar level. However, it cannot washout the differences between SURF and rootSIFT or SIFT. This is because the accuracy obtained using the SIFT and rootSIFT features are close, but the SURF features result in a significantly lower accuracy. From a total of 129 unseen videos in the test dataset, 41 videos missed either the skull or abdomen structures as assessed by visual inspection of videos. Unfortunately, keeping the sweeps so simple increases the chance of missing key anatomical structures. Therefore, automatic detection of fetal presentation would not be possible in such scenarios. From the remaining 88 videos, the presentation was correctly identified in 76

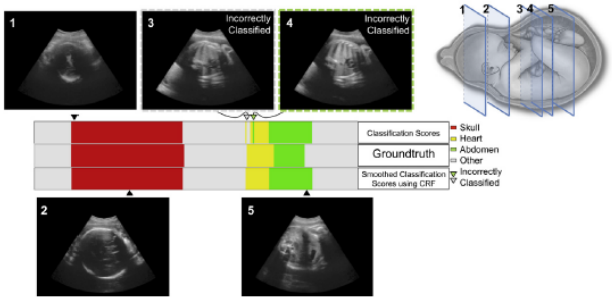


Fig. 8. Classification scores for a test video: Raw classification scores are shown on the top bar and regularized scores on the bottom bar. The red colour represents the frames that have been classified as fetal skull, and similarly yellow, green and grey represent the fetal heart, abdomen and other structures, receptively. As can be seen, the misclassified frames have been relabelled correctly based on their neighbouring frames through the regularization process. Moreover, the slices labelled 1 – 5 on the left, correspond to approximate locations of the five sample frames illustrated on the right. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Accuracy of fetal heart localization. The algorithm described in Section 2.3 is used to localize the fetal heart in each sequence and crop the frames around the located heart. Accuracy is reported in terms of the euclidean distance between predicted heart centre point and the groundtruth (GT).

Accuracy indication	%
Within GT diameter	82.4
Within GT radius	65.5
Within half GT radius	55.6

videos sweeps (83.4%). One of the main challenges with free-hand sweeps is to ensure correct anatomical structures are present and displayed appropriately. Inspecting some of the failure cases suggests that *unusual* appearance of the fetal skull or abdomen has contributed to mis-detection or failure to detect the presentation. These include views of the skull or abdomen that had not been seen in the training set and can be addressed in the future studies through larger and more comprehensive datasets.

Generally, the fetal skull and abdomen have significant distinguishing characteristics such as their outer boundaries and inner texture structures. In addition, they both occupy a substantial portion of the image on each frame. However, in our dataset this is not the case for the fetal heart. Due to the simplified scanning protocol, it is easy for fetal heart views to be very similar to those of the fetal abdomen. Moreover, the fetal heart is contained within a very small portion of the image, in comparison to the skull or abdomen. Indeed it may not even be captured as part of sweep at all. Such factors make fetal heart detection and characterization highly challenging in our dataset.

3.2. Localizing the fetal heart

136 short video sequences of a fetal heartbeat each of 30 frames long were extracted from the dataset. The method described in Section 2.3 was applied to find the approximate location of the fetal heart. The Euclidean distance between the predicted centre point of the fetal heart and the ground truth (GT) was calculated. Furthermore, a histogram of the distances is shown in Fig. 9 and the localization accuracy is shown in Table 4. As only the approximate location of the heart is required, the accuracy of this step was evaluated in terms of the Euclidean distance between the

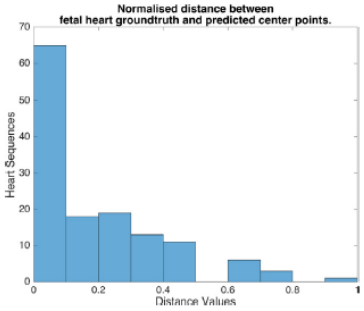


Fig. 9. Normalized Euclidean distance between the centre points of the predicted fetal heart and the groundtruth. A histogram of the normalized euclidean distances between the groundtruth points and the predicted centre points. The histogram is skewed towards lower distance points.

predicted and GT centre points of the fetal heart. As can be seen, in 82% of the cases, the distance between the GT and predicted centre point is less than the diameter of the detected fetal heart. This is the maximum permitted distance for an approximate localization of the fetal heart. Moreover, in more than 55% of the sequences the fetal heart has been localized almost perfectly, where the distance between the predicted and GT is less than half the radius of the fetal heart.

3.3. Analysing the fetal heartbeat

136 sequences of the fetal heart were used as positive fetal heartbeat examples. In addition, another 136 short sequences of the same duration were extracted randomly from dataset, where no fetal heart was present. This formed the negative samples. The dataset was split such that 70% was used for training and the remaining sequences were used for evaluating the accuracy of the system. Two experiments were conducted to analyse the dynamics of these subsequences, using the method described in Section 2.4 to identify whether a fetal heartbeat could be detected. The first experiment used the entire ultrasound image, whereas in the second experiment the fetal heart was initially localized following the method described in Section 2.3 and the video frames

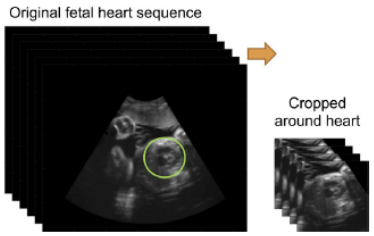


Fig. 10. Cropping around the detected fetal heart. Fetal heart dynamics are analysed once using the original ultrasound sequence (left) and once on the cropped sequence around the detected fetal heart (right).(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Classification results for detecting the fetal heartbeat without localising the fetal heart (Step B). The sequence dimensions are 240 × 320 × 30. Here n indicates the number of KDT model states and σ signifies the filter's centre-wavelength.

	Model n=3	Model n=4	Model n=5	Model n=6	Model n=7
σ =30	80.46	81.61	83.91	83.91	79.31
σ =35	80.46	81.61	83.91	83.91	83.91
σ =40	52.87	82.76	51.72	52.87	45.98
σ =45	83.91	**88.51**	85.06	57.47	68.97
σ =50	**88.51**	86.21	86.21	83.91	52.87
σ =55	55.17	52.87	57.47	52.87	55.17
σ =60	85.06	56.32	56.32	62.07	52.87

Table 6
Classification results for detecting the fetal heartbeat after cropping the frames around the localized fetal heart (Step B). The sequence dimensions are 120 × 120 × 30, cropped around the detected fetal heart. Here n indicates the number of KDT model states and σ signifies the filter's centre-wavelength.

	Model n=3	Model n=4	Model n=5	Model n=6	Model n=7
σ =30	81.61	80.46	80.46	83.91	54.02
σ =35	87.36	88.51	63.22	87.36	86.21
σ =40	49.43	52.87	79.31	78.16	52.87
σ =45	89.66	85.06	52.87	52.87	64.37
σ =50	**93.10**	50.57	89.66	70.11	54.02
σ =55	78.16	51.72	52.87	51.72	74.71
σ =60	63.22	57.47	55.17	64.37	42.53

were cropped around the detected fetal heart as illustrated in Fig. 10.

Recall that this algorithm is only run on frames that have been classified as fetal heart. Fig. 10 shows the heart detection boundary using a green circle. Moreover, as the main application is not accurate heart segmentation, a rectangular area defined by twice the radius of the detection circle plus an offset is empirically set to be the potential area of interest that would contain fetal heart motion. The accuracy values presented in Table 5 are for heartbeat detection without localizing the fetal heart and the accuracy values presented in Table 6 are for the combined localization and heartbeat detection pipeline. The purpose of this step is to assess the accuracy of the motion classification. To elaborate, a dedicated classifier for detecting the fetal hearts was not specifically trained using the sweep data. Instead the best trained model from Bridge and Noble (2015) was applied to the short sequence that have been short-listed in Step A of Fig. 4.

As shown in Table 5 without fetal heart detection, the best results were achieved with a 3-state model and $\sigma_{feat.symm} = 50$ for the signed feature symmetry filter (detection accuracy of 88.5%). In general, the classification accuracy was higher when the heart was first localized and cropped out of the video sequence. This reflects the fact that the full image contains a lot of irrelevant information and motion due to probe movements and fetal movement that can confound the heartbeat detection. When the frames are cropped around the detected fetal heart, a 3-state model and $\sigma_{feat.symm} = 50$ for the signed feature symmetry filter (detection accuracy of 93.1%) produced the highest results. In general, increasing the number of states leads to a decrease in performance. This can be explained by the fact that when KPCA is used, the main dynamics of the video are best described using the first 3 or 4 eigenvalues. Additional eigenvalues capture a very small portion of the variation in the feature space, thus resulting in noisier KDT model parameter estimates. Moreover, the duration of the heart sequences in this experiment are considerably short, thus larger increase in the number of states beyond reported does not improve the results.

One of the main challenges in modelling the dynamics of the fetal heart is the quality of the positive and negative samples used to train the dynamical model. Although the positive examples contain motions of beating fetal hearts, our negative dataset does not contain any examples of non-viable (non-beating) fetal heart. Instead the negative dataset consists of short sequences of anything but a fetal heart motion.

4. Conclusions

In this study we have looked at the problem of automatically locating anatomical features in fetal ultrasound video specifically motivated by a real world global health application of low-cost ultrasound for identification of breech presentation and fetal viability. Breech delivery can significantly increase the risk for neonates (Hannah et al., 2000). However, planned vaginal breech deliveries where antenatal ultrasound is available can be associated with a better outcome than reported in randomized trials (Goffinet et al., 2006).

Ultrasound requires a high degree of skill to perform well, and there is a lack of experienced sonographers in many developing world healthcare settings. The image analysis framework we have developed was directly developed to address the need to empower less experienced or well-trained users of ultrasound, or users new to ultrasound to effectively identify structures of interest and interpret the images with high confidence. Further, computer memory requirements for analysis are not large. The solution is amenable to use within a low-cost free-hand ultrasound system platform (where today USB and wireless transducers are of the order of $7k or less).

The implementation reported in the paper was for proof-of-principle and not optimized for real-time use. We have added the processing times but as no attempt was made to optimize them they are not really meaningful from which to infer potential real-time performance. Computation time are shown in Table 7. The experiment was carried out using *Matlab2016a* on a PC with 32GB of RAM, restricting the machine to use only a single core.

This framework assumes that a consecutive sequence of fetal skulls and abdomens are present in any given sweep, in order to identify the fetal presentation. From a total of 129 unseen videos in the test dataset, 41 videos did not contain either the skull or abdomen structures, which are necessary for automatic detection of fetal presentation. From the remaining 88 videos, the presentation was correctly identified in 76 videos sweeps (83.4%). Furthermore,

Table 7
Computation time for encoding SIFT, rootSIFT, and SURF fea- tures using BoVW, VLAD, and FV encoding. The duration for encoding the features for one image, in seconds.

	BoVW	VLAD	FV
SIFT	1.452	0.297	0.165
rootSIFT	1.218	0.286	0.184
SURF	0.267	0.085	0.051

for the detection of the heartbeat an overall classification accuracy of 93.1% was achieved. In the 12 videos where the presentation was not identified correctly, although the fetal and abdomen were present in the ultrasound video sweep, these structures were not captured fully and visually looked different to the majority of those in the training set.

On our choice of classifier, SVM is a classical learning algorithm, which has demonstrated excellent performance in many applications, including our previous work and work of others. For example Lei et al. (2015) recently proposed the use of root-SIFT features with an SVM classification framework for detecting fetal faces in ultrasound scans. We found it gave good results (as evidenced in the article) and did not see the value to move on to consider more sophisticated hand-crafted feature classifiers (for instance, random forests). Convolutional neural networks (CNNs) have very recently become popular in medical image analysis. Popularity of CNNs coincided with the later stages of the work reported here. Current CNN architectures generally require larger datasets than were available for this research, and work best with balanced label datasets (ours is unbalanced). You can use CNNs to partition ultrasound video, as described in recent preliminary research of our group (Gao et al., 2016), and other on-going research in our laboratory. The accuracy is slightly better. Going forward, it will be interesting to investigate whether CNN architectures can be designed to offer significant advantages over other methods for ultrasound video analysis.

In practice obstetricians may repeat an acquisition multiple times before they obtain satisfactory results. In this study, we have used only a single sweep. Initially the aim was to analyse what can be achieved from analysis of an extremely simple linear sweep (a minimal sweep). Given that the results are so promising, a logical next step is to extend the analysis to multiple sweeps which poses interesting research questions about how to fuse information obtained from multiple sweeps for clinical decision-making. This is the subject of some of our on-going work that we hope to report on in a future publication.

The data used in this study was obtained from the INTERGROWTH-21ST project (Sarris et al., 2013; Papageorghiou et al., 2014), which contains mothers at different gestational ages and with diverse body mass indices. Therefore the positive sub-sequences extracted from this data include a variety of representations of the fetal skull with different sizes and shadowing. This provides a set of rich features for the dataset of the positive sequences however a larger dataset of ultrasound sweeps might be required to build a robust classifier for more general populations.

Acknowledgements

M.A. Maraci acknowledges the support of RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation) and the Oxford EPSRC Impact Acceleration Award fund. C.P. Bridge support via a UK EPSRC DTA studentship. A. Noble acknowledgesthe support of the EPSRC Programme Grant Seebibyte. The authors also acknowledges that the data was acquired as part of the INTERGROWTH-21ST project. The authors gratefully acknowledge the help of Dr. Tess Norris, Dr. Sikolia Wanyonyi, Dr. Malid Molloholli, Dr. Christina Aye, and Miss Fenella Roseman, Research Midwife for acquiring the ultrasound data used in this project. All the in-vivo experiments were in accordance with the ethical standards of the institutional and national research committees.

Data statement

The images and image annotations used in this paper cannot be made freely available for reasons of ethical sensitivity. Data related to the tables will be made available from the Oxford University Research Archive (ORA-Data) on paper acceptance.

Appendix A. Breakdown of results

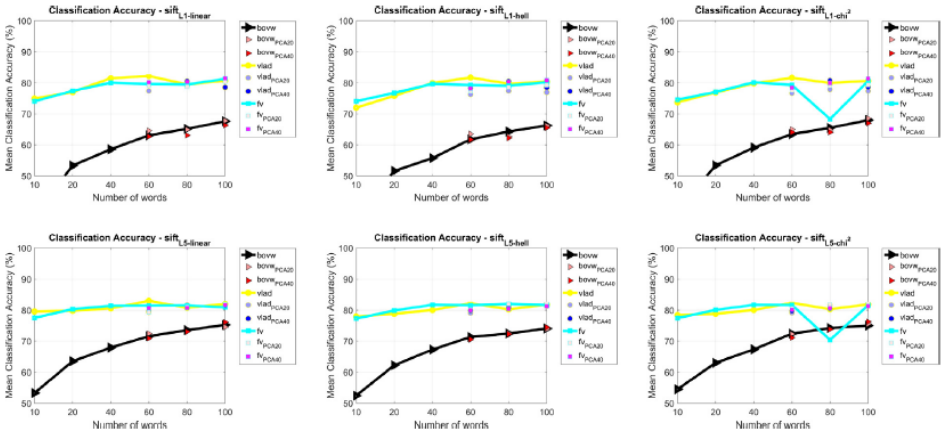


Fig. A.11. Mean classification accuracies for SIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

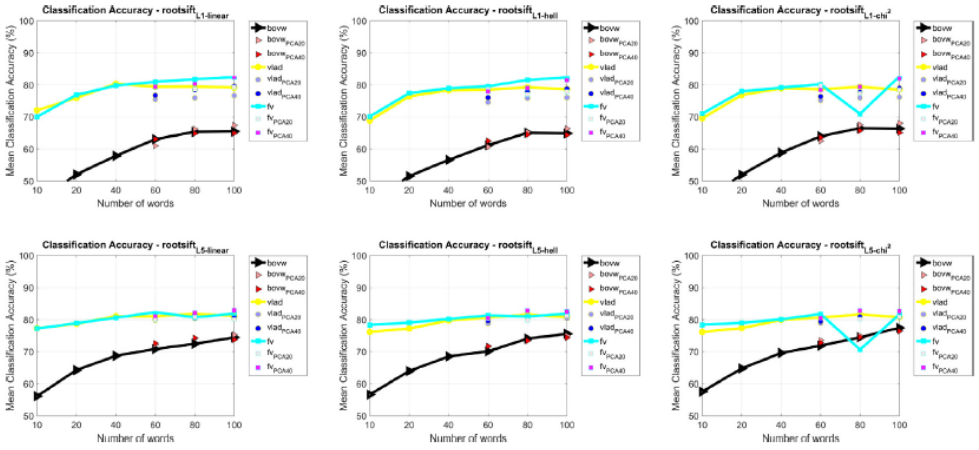


Fig. A.12. Mean classification accuracies for rootSIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

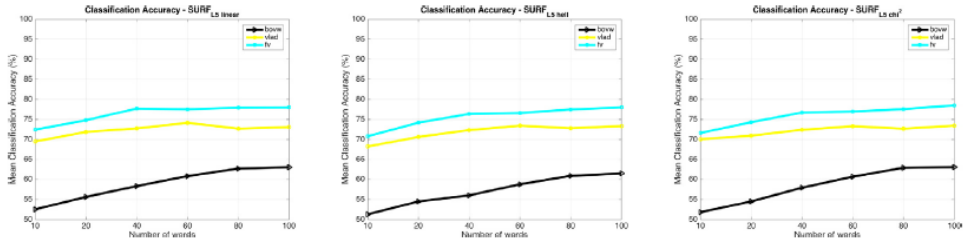


Fig. A.13. Mean classification accuracies for SURF feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

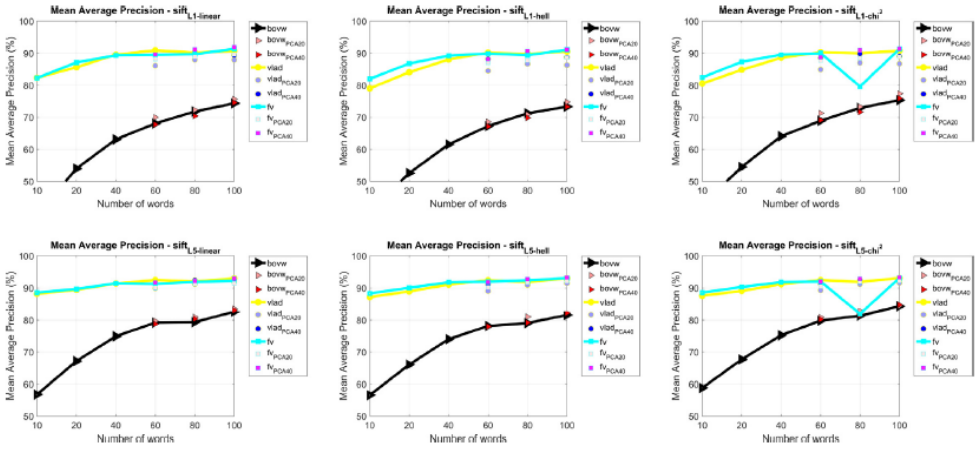


Fig. A.14. Mean average precision for SIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

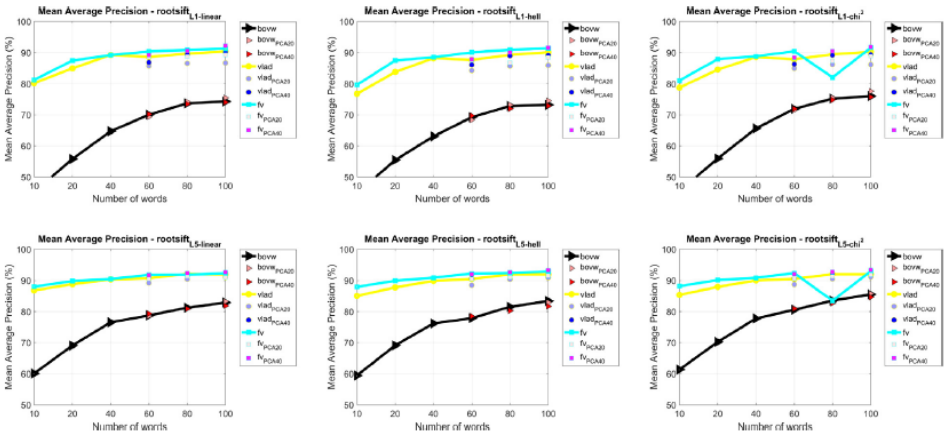


Fig. A.15. Mean average precision for rootSIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

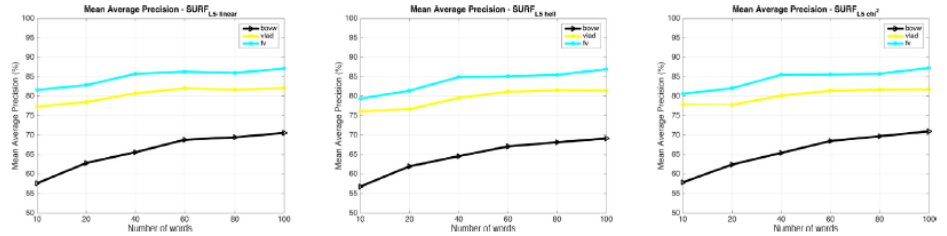


Fig. A.16. Mean average precision for rootSIFT feature descriptors. Feature encoding is carried out using the FV, VLAD, BoVW.

References

Anto, E.A., Amoah, B., Crimi, A., 2015. Segmentation of ultrasound images of fetal anatomic structures using random forest for low-cost settings. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, pp. 793–796. doi:10.1109/EMBC.2015.7318481.

Arandjelović, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Washington, DC, USA, pp. 2911–2918.

Bauer, S., Nolte, L.-P., Reyes, M., 2011. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part III. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 354–361. chapter Fully Automatic Segmentation of Brain Tumor Images Using Support Vector Machine Classification in Combination with Hierarchical Conditional Random Field Regularization. doi: 10.1007/978-3-642-23626-6_44.

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Smith, S., Kainz, B., Rueckert, D., 2016. Real-Time Standard Scan Plane Detection and Localisation in Fetal Ultrasound Using Fully Convolutional Neural Networks. Springer International Publishing, Cham, pp. 203–211. doi:10.1007/978-3-319-46723-8_24.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: speeded up robust features. In: European conference on computer vision. Springer, pp. 404–417.

Bissacco, A., Chiuso, A., Soatto, S., 2007. Classification and recognition of dynamical models: the role of phase, independent components, kernels and optimal transport. Pattern Anal. Mach. Intell. IEEE Trans. 29 (11), 1958–1972.

Bridge, C., Ioannou, C., Noble, J., 2017. Automated annotation and quantitative description of ultrasound videos of the fetal heart. Medical Image Analysis 36, 147–161. doi:10.1016/j.media.2016.11.006.

Bridge, C.P., Noble, J.A., 2015. Object localisation in fetal ultrasound images using invariant features. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, pp. 156–159. doi:10.1109/ISBI.2015.7163839.

Brown, P.G., Alsousou, J., Cooper, A., Thompson, M.S., Noble, J.A., 2013. The autoqual ultrasound elastography method for quantitative assessment of lateral strain in post-rupture achilles tendons. J. Biomech. 46 (15), 2695–2700. doi:10.1016/j.jbiomech.2013.07.044.

Carneiro, G., Georgescu, B., Good, S., Comaniciu, D., 2008. Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree.. IEEE Trans. Med. Imaging 27 (9), 1342–1355. doi:10.1109/TMI.2008.928917.

Chan, A.B., Vasconcelos, N., 2005. Probabilistic kernels for the classification of auto-regressive visual processes. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 1. IEEE, pp. 846–851.

Chan, A.B., Vasconcelos, N., 2007. Classifying video with kernel dynamic textures. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, pp. 1–6.

Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A., 2011. The devil is in the details: an evaluation of recent feature encoding methods.. In: BMVC, 2, p. 8.

Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R., 2009. Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, pp. 1932–1939.

Chaudhry, R., Vidal, R., 2009. Recognition of Visual Dynamical Processes: Theory, Kernels and Experimental Evaluation. Technical Report09–01.

Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., Heng, P., 2015. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. Biomed Health Inf. IEEE J. 19 (5), 1627–1636. doi:10.1109/JBHI.2015.2425041.

Chykeyuk, K., Yaqub, M., Alison Noble, J., 2014. Class-specific regression random forest for accurate extraction of standard planes from 3d echocardiography. In: Menze, B., Langs, G., Montillo, A., Kelm, M., Miller, H., Tu, Z. (Eds.), Medical Computer Vision. Large Data in Medical Imaging. In: Lecture Notes in Computer Science, 8331. Springer International Publishing, pp. 53–62. doi:10.1007/978-3-319-05530-5_6.

De Cock, K., De Moor, B., 2000. Subspace angles and distances between arma models. In: Proc. of the Intl. Symp. of Math. Theory of networks and systems, 1. Citeseer.

Doretto, G., Chiuso, A., Wu, Y.N., Soatto, S., 2003. Dynamic textures. Int. J. Comput. Vis. 51 (2), 91–109.

Forney Jr, G.D., 1973. The viterbi algorithm. Proc. IEEE 61 (3), 268–278.

Gao, Y., Maraci, M., Noble, J., 2016. Describing ultrasound video content using deep convolutional neural networks. Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on.

Goffinet, F., Carayol, M., Foidart, J.-M., Alexander, S., Uzan, S., Subtil, D., Brart, G., 2006. Is planned vaginal delivery for breech presentation at term still an option? results of an observational prospective survey in france and belgium. Am. J. Obstet. Gynecol. 194 (4), 1002–1011. doi:10.1016/j.ajog.2005.10.817.

Hannah, M.E., Hannah, W.J., Hewson, S.A., Hodnett, E.D., Saigal, S., Willan, A.R., 2000. Planned caesarean section versus planned vaginal birth for breech presentation at term: a randomised multicentre trial. term breech trial collaborative group.. Lancet 356 (9239), 1375–1383.

Imaduddin, Z., Akbar, M.A., Tawakal, H., Satwika, P., Saroyo, Y., 2015. Automatic detection and measurement of fetal biometrics to determine the gestational age. In: Information and Communication Technology (ICoICT), 2015 3rd International Conference on, pp. 608–612. doi:10.1109/ICoICT.2015.7231495.

Jegou, H., Douze, M., Schmid, C., Perez, P., 2010. Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 3304–3311. doi:10.1109/CVPR.2010.5540039.

Joachims, T., 2006. Training linear svms in linear time. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 217–226.

Kadour, M.J., Adams, R., English, R., Parulekar, V., Christopher, S., Noble, J.A., 2010. Slip imaging: reducing ambiguity in breast lesion assessment. Ultrasound Med. Biol. 36 (12), 2027–2035.

Kadour, M.J., Noble, J.A., 2009. Assisted-freehand ultrasound elasticity imaging. Ultrason. Ferroelectr. Freq. Control IEEE Trans. 56 (1), 36–43.

Kumar, A., Shriram, K., 2015. Automated scoring of fetal abdomen ultrasound scan–planes for biometry. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on. IEEE, pp. 862–865.

Kwitt, R., Vasconcelos, N., Razzaque, S., Aylward, S., 2013. Localizing target structures in ultrasound video - a phantom study. Med. Image Anal. 17 (7), 712–722. doi:10.1016/j.media.2013.05.003. Special Issue on the 2012 Conference on Medical Image Computing and Computer Assisted Intervention

Lafferty, J. D., McCallum, A., Pereira, F. C. N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 282–289.

Lei, B., Tan, E.-L., Chen, S., Zhuo, L., Li, S., Ni, D., Wang, T., 2015. Automatic recognition of fetal facial standard plane in ultrasound image via fisher vector. PLoS ONE 10 (5), e0121838. doi:10.1371/journal.pone.0121838.

Liu, K., Skibbe, H., et al., 2014. Rotation-invariant HOG descriptors using fourier analysis in polar and spherical coordinates. IJCV 106 (3), 342–364.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60 (2), 91–110.

Maraci, M., Napolitano, R., Papageorghiou, A., Noble, J., 2015. Fisher vector encoding for detecting objects of interest in ultrasound videos. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on. IEEE, pp. 651–654.

Maraci, M.A., Napolitano, R., Papageorghiou, A., Noble, J.A., 2014. Object classification in an ultrasound video using lp-sift features. In: Menze, B., Langs, G., Montillo, A., Kelm, M., Miller, H., Zhang, S., Cai, W.T., Metaxas, D. (Eds.), Medical Computer Vision: Algorithms for Big Data. In: Lecture Notes in Computer Science, 8848. Springer International Publishing, pp. 71–81. doi:10.1007/978-3-319-13972-2_7.

Maraci, M.A., Napolitano, R., Papageorghiou, A., Noble, J.A., 2014. Searching for structures of interest in an ultrasound video sequence. In: Wu, G., Zhang, D., Zhou, L. (Eds.), Machine Learning in Medical Imaging. In: Lecture Notes in Computer Science, 8679. Springer International Publishing, pp. 133–140. doi:10.1007/978-3-319-10581-9_17.

McIntosh, C., Svistoun, I., Purdie, T.G., 2013. Groupwise conditional random forests for automatic shape classification and contour quality assessment in radiotherapy planning. IEEE Trans. Med. Imaging 32 (6), 1043–1057. doi:10.1109/TMI.2013.2251421.

Namburete, A.I., Stebbing, R.V., Kemp, B., Yaqub, M., Papageorghiou, A.T., Noble, J.A., 2015. Learning-based prediction of gestational age from ultrasound images of the fetal brain. Med. Image Anal. 21 (1), 72–86.

Noble, J.A., 2016. Reflections on ultrasound image analysis. Med. Image Anal. 33, 33–37. 20th anniversary of the Medical Image Analysis journal (MedIA). doi:10.1016/j.media.2016.06.015

Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., Kohli, P., 2011. Decision tree fields. In: Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, pp. 1668–1675.

Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Ismail, L.C., Lambert, A., Jaffer, Y.A., Bertino, E., Gravett, M.G., Purwar, M., Noble, J.A., Pang, R., Victora, C.G., Barros, F.C., Carvalho, M., Salomon, L.J., Bhutta, Z.A., Kennedy, S.H., Villar, J., 2014. International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the intergrowth-21st project. Lancet 384 (9946), 869–879. doi:10.1016/S0140-6736(14)61490-2.

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the fisher kernel for large-scale image classification. In: ECCV. Springer Berlin Heidelberg, pp. 143–156.

Ponomarev, G., Gelfand, M., Kazanov, M., 2012. A multilevel thresholding combined with edge detection and shape-based recognition for segmentation of fetal ultrasound images. In: Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012, pp. 17–19.

Rahmatullah, B., Papageorghiou, A., Noble, J., 2011. Automated selection of standardized planes from ultrasound volume. In: Suzuki, K., Wang, F., Shen, D., Yan, P. (Eds.), Machine Learning in Medical Imaging. In: Lecture Notes in Computer Science, 7009. Springer Berlin/Heidelberg, pp. 35–42. doi: 10.1007/978-3-642-24319-6_5

Rahmatullah, B., Sarris, I., Papageorghiou, A., Noble, J.A., 2011. Quality control of fetal ultrasound images: detection of abdomen anatomical landmarks using adaboost. In: Proc. IEEE Int Biomedical Imaging: From Nano to Macro Symp, pp. 6–9.

Rajpoot, K., Grau, V., Noble, J., 2009. Local-phase based 3d boundary detection using monogenic signal and its application to real-time 3-d echocardiography images. In: Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on, pp. 783–786. doi:10.1109/ISBI.2009.5193166.

Rijken, M.J., Lee, S.J., Boel, M.E., Papageorghiou, A.T., Visser, G.H.A., Dwell, S.L.M., Kennedy, S.H., Singhasivanon, P., White, N.J., Nosten, F., McGready, R., 2009. Obstetric ultrasound scanning by local health workers in a refugee camp on the thaiburmese border. Ultrasound Obstetrics Gynecol. 34 (4), 395–403. doi:10.1002/uog.7350.

Rueda, S., Fathima, S., Knight, C., Yaqub, M., Papageorghiou, A., Rahmatullah, B., Foi, A., Maggioni, M., Pepe, A., Tohka, J., Stebbing, R., McManigle, J., Ciurte, A., Bresson, X., Cuadra, M., Sun, C., Ponomarev, G., Gelfand, M., Kazanov, M., wei Wang, C., Chen, H.-C., Peng, C.-W., Hung, C.-M., Noble, J., 2014. Evaluation and comparison of current fetal ultrasound image segmentation methods for biometric measurements: A grand challenge. Med. Imaging IEEE Trans. 33 (4), 797–813. doi:10.1109/TMI.2013.2276943.

Salomon, L.J., Alfirevic, Z., Berghella, V., Bilardo, C., Hernandez-Andrade, E., Johnsen, S.L., Kalache, K., Leung, K.-Y., Malinger, G., Munoz, H., Prefumo, F., Toi, A., Lee, W., on behalf of the ISUOG Clinical Standards Committee, 2011. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound Obstetrics Gynecol. 37 (1), 116–126. doi:10.1002/uog.8831.

Sarris, I., Ioannou, C., Dighe, M., Mitidieri, A., Oberto, M., Qingqing, W., Shah, J., Sohoni, S., Al Zidjali, W., Hoch, L., Altman, D.G., Papageorghiou, A.T., I. F., for the 21st Century, N.G.C., 2011. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements.. Ultrasound Obstetrics Gynecol. 38 (6), 681–687.

Sarris, I., Ioannou, C., Ohuma, E., Altman, D., Hoch, L., Cosgrove, C., Fathima, S., Salomon, L., Papageorghiou, A., for the International Fetal, for the 21st Century (INTERGROWTH-21st), N.G.C., 2013. Standardisation and quality control of ultrasound measurements taken in the intergrowth-21st project. BJOG: Int. J. Obstetrics Gynaecol. 120, 33–37. doi:10.1111/1471-0528.12315.

Sun, C., 2012. Automatic fetal head measurements from ultrasound images using circular shortest paths. In: Proceedings of Challenge US: Biometric Measurements from Fetal Ultrasound Images, ISBI 2012, pp. 13–15.

Tiran, D., 2005. Nice guideline on antenatal care: routine care for the healthy pregnant womanrecommendations on the use of complementary therapies do not promote clinical excellence. Complementary Ther. Clin. Pract. 11 (2), 127–129.

Villar, J., Ismail, L.C., Victora, C.G., Ohuma, E.O., Bertino, E., Altman, D.G., Lambert, A., Papageorghiou, A.T., Carvalho, M., Jaffer, Y.A., Gravett, M.G., Purwar, M., Frederick, I.O., Noble, A.J., Pang, R., Barros, F.C., Chumlea, C., Bhutta, Z.A., Kennedy, S.H., 2014. International standards for newborn weight, length, and head circumference by gestational age and sex: the newborn cross-sectional study of the intergrowth-21st project. Lancet 384 (9946), 857–868. doi:10.1016/S0140-6736(14)60932-6.

Vishwanathan, S.V., Smola, A.J., Vidal, R., 2007. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. Int. J. Comput. Vision 73 (1), 95–119. doi:10.1007/s11263-006-9352-0.

Wanyonyi, S.Z., Napolitano, R., Ohuma, E.O., Salomon, L.J., Papageorghiou, A.T., 2014. Image-scoring system for crownrump length measurement. Ultrasound Obstetrics Gynecol. 44 (6), 649–654. doi:10.1002/uog.13376.

Yaqub, M., Javaid, M.K., Cooper, C., Noble, J.A., 2011. Improving the Classification Accuracy of the Classic Rf Method by Intelligent Feature Selection and Weighted Voting of Trees with Application to Medical Image Segmentation. In: Machine Learning in Medical Imaging. Springer, pp. 184–192.

Yaqub, M., Napolitano, R., Ioannou, C., Papageorghiou, A.T., Noble, J.A., 2012. Automatic detection of local fetal brain structures in ultrasound images. In: Proc. 9th IEEE Int Biomedical Imaging (ISBI) Symp, pp. 1555–1558.

Yaqub, M., Rueda, S., Kopuri, A., Melo, P., Papageorghiou, A., Sullivan, P., McCormick, K., Noble, J., 2015. Plane localization in 3d fetal neurosonography for longitudinal analysis of the developing brain. Biomed. Health Inf. IEEE J. PP (99), 1–1. doi: 10.1109/JBHI.2015.2435651.

Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C., 2007. Local features and kernels for classification of texture and object categories: a comprehensive study. Int. J. Comput. Vision 73 (2), 213–238.

Appendix 5: Learning from redundant but inconsistent reference data: Anatomical views and measurements for fetal brain screening

SPIE, Medical Imaging 2016: Image Processing, 97841A

Learning from redundant but inconsistent reference data: Anatomical views and measurements for fetal brain screening

I. Waechter-Stehle[a], T. Klinder[a], J.-M. Rouet[b], D. Roundhill[c], G. Andrews[c], A. Cavallaro[d], M. Molloholli[d], T. Norris[d], R. Napolitano[d], A. Papageorghiou[d] and C. Lorenz[a]

[a]Philips Research Hamburg, Germany,
[b]Philips Research Paris, France,
[c] Philips Ultrasound, Bothell, US,
[d] Nuffield Department of Obstetrics and Gynaecology, University of Oxford, UK

ABSTRACT

In a fetal brain screening examination, a standardized set of anatomical views is inspected and certain biometric measurements are taken in these views. Acquisition of recommended planes requires a certain level of operator expertise. 3D ultrasound has the potential to reduce the manual task to only capture a volume containing the head and to subsequently determine the standard 2D views and measurements automatically. For this purpose, a segmentation model of the fetal brain was created and trained with expert annotations. It was found that the annotations show a considerable intra- and inter-observer variability. To handle the variability, we propose a method to train the model with redundant but inconsistent reference data from many expert users. If the outlier-cleaned average of all reference annotations is considered as ground truth, errors of the automatic view detection are lower than the errors of all individual users and errors of the measurements are in the same range as user error. The resulting functionality allows the completely automated estimation of views and measurements in 3D fetal ultrasound images.

Keywords: Model-based segmentation, ultrasound, fetal screening, fetal brain, automatic measurements

1. INTRODUCTION

To detect fetal brain anomalies, it is recommended to perform an ultrasound screening at the gestational age of 18 to 24 weeks. According to the International Society of Ultrasound in Obstetrics and Gynaecology (ISUOG) guidelines, an examination consists of visual inspection of the brain in several anatomical views and of biometric measurements.[1] The most important views are trans-ventricular (TV), trans-thalamic (TT) and trans-cerebellar (TC). Important measurements are bi-parietal diameter (BPD), occipital-frontal diameter (OFD), and trans-cerebellar diameter (TCD). According to the guidelines, the three views are defined by the brain structures which should be visible within the view. This geometrically somewhat fuzzy definition, limited image quality, and the fact that the views are oblique, lead to substantial variation in how experts place the views. The acquisition of recommended planes requires a certain level of operator expertise, standardisation and quality control.[2] In particular for less experienced users, the task is quite difficult. As the measurements are taken in the anatomical views, variation in view definition also leads to variations in measurements.[3] Currently, mainly 2D ultrasound is used for screening. However, it can be expected that 3D ultrasound will increasingly be used, facilitating automated image analysis.

In this paper, we propose a learning-based solution to automatically determine anatomical views and biometric measurements using a model-based segmentation approach. The model establishes anatomical correspondence between data sets and thus serves as a reference structure. In this way, views and measurements can be learned during the training phase from manual annotations and encoded in the model. Sofka et al.[4] presented the work closest to ours, where an integrated detection network was proposed. In contrast to that, we explicitly segment the underlying anatomy by means of an anatomical model in order to be more robust especially against poor image quality and to also offer volumetric measurements in the future. Cuingnet et al.[5] and Chen et al.[6] presented a method to detect anatomical planes without allowing to do detailed measurements like the TCD.

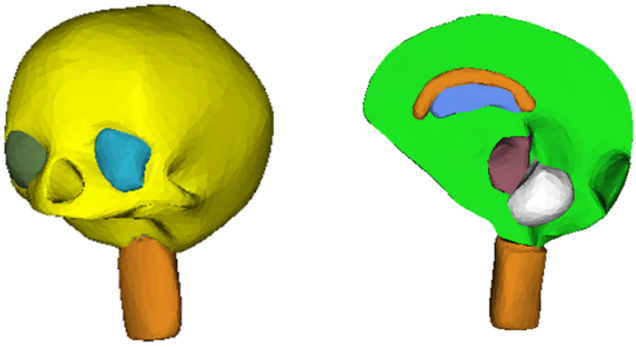


Figure 1. Fetal skull and brain model. Left: skull and eyes. Right: internal brain structures: cerebellum, thalamus, cavum septum pellucidum, corpus callosum, septum.

2. METHODS

As anatomical views are defined based on the structures that need to be examined in the view, a detailed segmentation of these structures facilitates estimation of the view planes. In addition, it enables detailed measurements like the TCD. The segmentation is achieved using a model-based approach. Structures covered by the model are skull, eyes, cerebellum, thalamus, cavum septum pellucidum, corpus callosum and septum (see Fig. 1). To create the model, the following steps were carried out: The initial mesh of the skull was determined from a CT scan of a fetal skeleton phantom. This was adapted manually in a deformable way to an ultrasound image of good image quality where in addition the inner brain structures were delineated manually. This initial mesh was then adapted manually to 12 additional ultrasound images. The resulting meshes with identical mesh topology were used to compute a mean shape model and to learn optimal boundary detection features.[7]

Fetal brain ultrasound images are generally acquired from the right or left side of the fetal head. Depending on the fetal lie, the orientation can still vary considerably. Therefore, in a first step, the location and orientation of the fetal skull is determined.[5] Here, a spheroidal template-matching delivers a localization estimate which is combined with a Hough Transform based estimation of the mid-sagittal plane and a Random Forest based localization of the eye orbits. Using the input from the previous step, the model is posed in the image. Subsequently, the model is adapted to the image in a hierarchical coarse to fine manner.[8]

The main goal of this work is to learn how to detect certain anatomical views and measurements from manual reference annotations and encode them into the fetal head model. As fetal heads differ in size, position, and orientation, a simple approach such as averaging in image space is not possible. Instead, we use the fetal head model as reference to learn the annotations. All models have the same topology and the triangle locations are anatomically well defined, meaning that they end up at very similar anatomical positions after adaptation. This allows to encode anatomical views and measurements into the model mesh by labeling the relevant triangles. During the collection of manual annotations a high inconsistency became apparent. To avoid, or at least reduce the error of individual experts and individual examples, annotations of many experts (up to 10 annotations per case) on several example datasets were used. Before using them for training, outliers were removed.

2.1 Learning of a measurement

In order to encode a distance measurement, the endpoints are encoded as landmarks in the model. Let D be the set of training data, $U_{d,l}$ be the set of users that provided an annotation for dataset $d \in D$ and landmark l and $\vec{x}_{u,d,l}$ the according reference annotation. To reduce the influence of individual sub-optimal annotations, outliers are determined per dataset

$$O(U_{d,l}) = \{u \in U_{d,l} | \, \|\vec{x}_{u,d,l} - \text{Mean}\,\{\vec{x}_{v,d,l} | v \in U_{d,l} \backslash u\}\| > h\} \tag{1}$$

with a threshold h. Landmarks are then averaged according to

$$\vec{x}_{d,l} = \text{Mean}\left\{\vec{x}_{u,d,l} \,|\, u \in U_{d,l} \wedge u \notin O(U_{d,l})\right\} \tag{2}$$

To learn across different datasets, the model mesh is used as support. Each landmark annotation is related to its corresponding (individualized) mesh. Let $M_d = (T_d, V_d)$ be the mesh of the segmentation of dataset d, consisting of triangles T_d and vertices V_d, where $M_{MM} = (T_{MM}, V_{MM})$ is the mean mesh of the model and $\vec{c}_t$ is the center of gravity of a triangle t. The nearest triangle to a point $\vec{x}$ in a mesh M is given by

$$NT(\vec{x}, M_d) = \underset{t \in T_d}{\arg\min} \left\|\vec{c}_t - \vec{x}\right\|. \tag{3}$$

The outlier-cleaned average landmark position is given by

$$\vec{x}_l = \sum_{t \in T_{MM}} w_l(t)\vec{c}_t, \tag{4}$$

where the weighting function $w_l(t)$ contains information about all landmark annotations in the set. The weighting function per triangle is defined by

$$w_l(t) = \frac{1}{|D|}\left|\{d \in D | NT(\vec{x}_{d,l}, M_d) = t\}\right|. \tag{5}$$

The triangle that receives the final encoding label is given by $t_{l,enc} = NT(\vec{x}_l, M_{MM})$.

2.2 Learning of planes

Similar to the measurements, a view plane is encoded as a set of labeled triangles in the model. To decode the view after model individualization, the regression plane of the respective triangles define the plane normal and offset. To ensure defined in-plane orientation of the view, two additional landmarks are encoded in the model, denoting the x- and y-direction.

A plane $P = (\vec{x}, \vec{n})$ is defined by a point $\vec{x}$ and its normal $\vec{n}$. Let $P_{u,d,p}$ be the reference annotation of plane p by user u, for dataset d. To learn a plane across different datasets, the annotation is again related to its corresponding mesh. The plane P cuts the mesh M in the set of triangles that is defined as $\text{Cut}(P, M)$. For averaging planes across different datasets, the regression plane (Reg) is determined for a weighted set of triangles

$$P_p = \text{Reg}(\{w_p(t)\vec{c}_t | t \in T_{MM}\}) \tag{6}$$

where the weighting function $w_p(t)$ contains the information about all planes in the set. The weighting function per tringle t is defined by

$$w_p(t) = \left|\{d \in D | t \in \text{Cut}(P_{d,p}, M_d)\}\right|. \tag{7}$$

For outlier removal, the same is done per case and all triangles that are only contained in one annotation get a weight of zero, giving $P_{d,p}$. The triangles that should get the final encoding labels are given by $t_{p,enc} = \text{Cut}(P_p, M_{MM})$.

3. EXPERIMENTS AND RESULTS

The fetal head localization was trained, optimized and validated separately (see[5]). For creation and validation of the fetal head model, manual view and measurement annotations were done on 27 datasets by up to 10 users per dataset. The users consisted of three technical experts (expert level (EL) 1), four general sonographers (EL2), and three clinicians specialized in fetal medicine (EL3). All were briefed to annotate according to the ISUOG definition.[1] 13 data sets were selected to train the model and the other 14 data sets were used for testing. An example segmentation result of the created fetal head model is shown in Figure 2.

The reference annotations vary considerably, as can be seen in Fig. 3. For quantitative evaluation, the outlier-cleaned average per case with outlier removal was considered as ground truth. Each individual user and

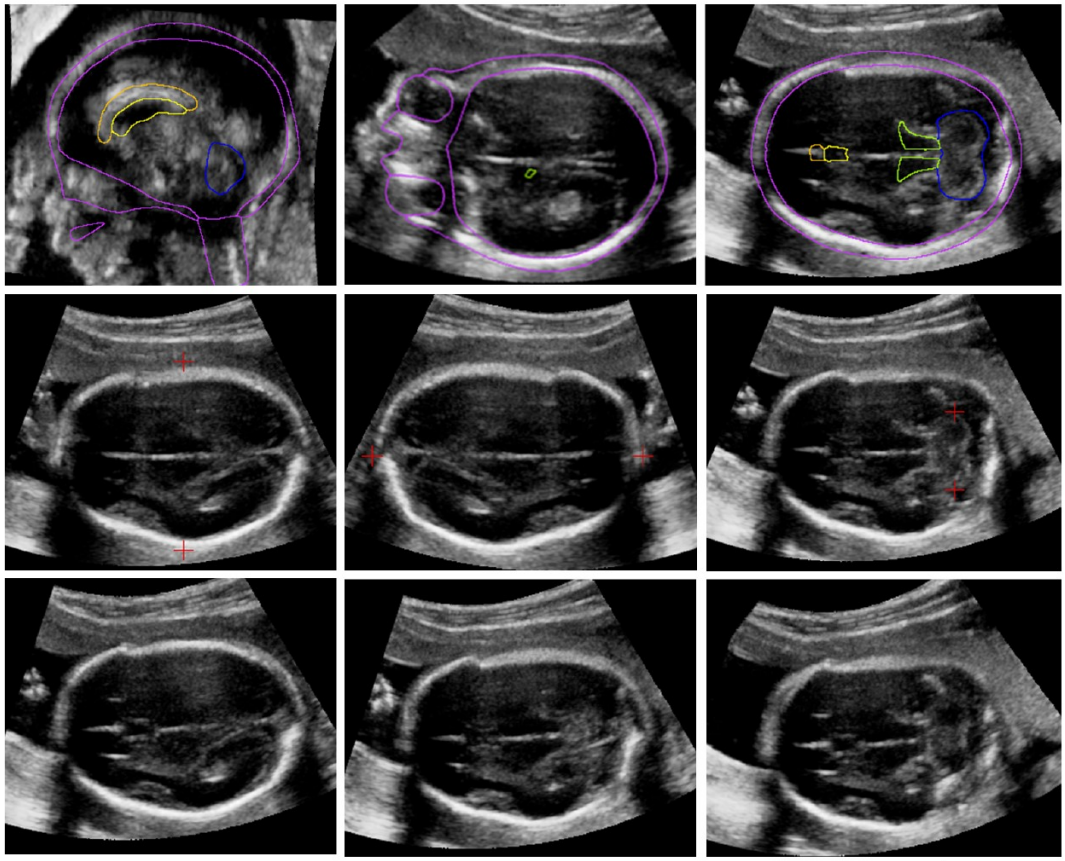


Figure 2. Top: Example of segmentation result in different automatically determined views. Middle: Automatically determined biometric measurements (BPD, OFD, TCD). Bottom: Automatically determined planes (TV, TT, TC).

	TC		TCD	OFD
	Angle [°]	Offset [mm]	Error [mm]	Error [mm]
Auto (trained on all)	**4.71 ± 2.18**	**1.20 ± 0.68**	**0.72 ± 0.66**	**1.02 ± 1.16**
EL1 [min,max]	[6.33, 11.09]	[1.33, 1.77]	[0.95, 1.68]	[0.80, 2.15]
EL1, mean ± std	8.06 ± 4.22	1.52 ± 1.43	1.28 ± 1.37	1.24 ± 1.01
EL2 [min,max]	[7.89, 15.26]	[2.54, 3.63]	[0.98, 2.06]	[0.78, 0.93]
EL2, mean ± std	11.72 ± 7.67	3.11 ± 2.43	1.54 ± 0.89	0.86 ± 0.58
EL3 [min,max]	[8.91, 10.89]	[2.11, 2.64]	[0.64, 1.66]	[0.54, 0.98]
EL3, mean ± std	9.80 ± 6.12	2.34 ± 2.31	1.16 ± 0.96	0.83 ± 0.86

Table 1. Quantitative evaluation of view generation and measurements - ground truth: average of all users with outlier removal, evaluation set: test datasets.

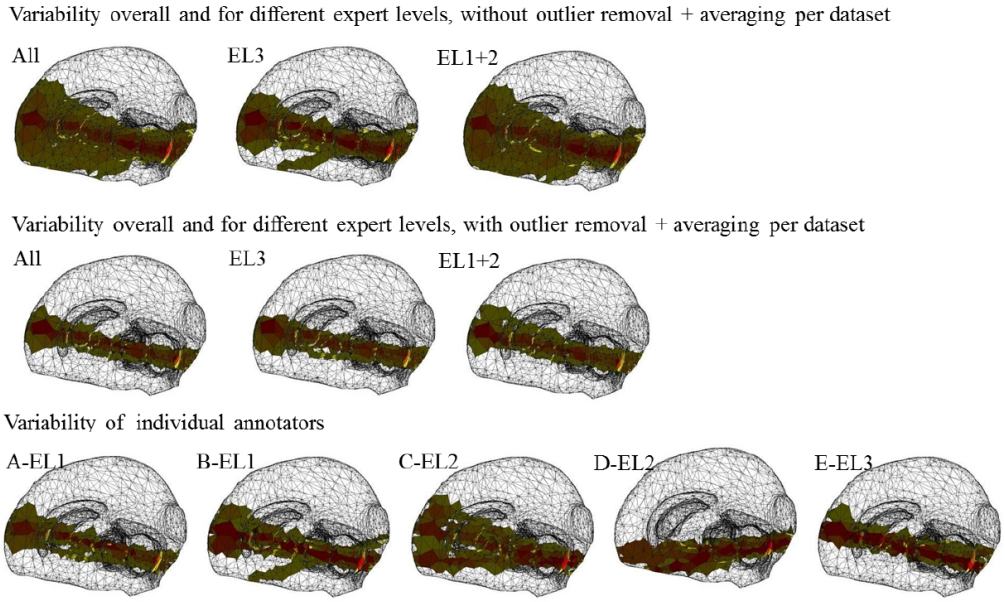


Figure 3. The weighting function (see Eqn. 7) used during training of planes can be utilized to visualize variability. The overall variability is very high, outlier removal and averaging per case drastically reduces it.

	TC Angle [°]	Offset [mm]	TCD Error [mm]	OFD Error [mm]
Auto (train EL3)	**4.74 ± 2.00**	**0.88 ± 0.49**	**1.03 ± 0.91**	**1.47 ± 1.48**
Auto (train EL1+2)	**5.48 ± 2.11**	**1.02 ± 0.80**	**1.04 ± 0.91**	**1.50 ± 1.51**
EL1, mean ± std	8.83 ± 5.71	1.70 ± 1.84	1.58 ± 1.55	1.59 ± 1.24
EL2, mean ± std	11.88 ± 7.95	3.14 ± 2.67	2.11 ± 1.25	1.13 ± 0.86
EL3, mean ± std	8.69 ± 5.80	1.65 ± 1.72	0.86 ± 0.74	0.93 ± 0.63

Table 2. Quantitative evaluation of view generation and measurements - ground truth: average of EL3 users, evalution set: test datasets.

	TC Angle [°]	Offset [mm]	TCD Error [mm]	OFD Error [mm]
Auto (trained on all)	**3.83 ± 1.73**	**0.97 ± 1.06**	**0.65 ± 0.44**	**1.01 ± 0.67**
EL1, mean ± std	8.08 ± 6.71	1.81 ± 2.38	1.60 ± 1.86	1.09 ± 0.91
EL2, mean ± std	14.01 ± 8.15	3.70 ± 3.78	1.29 ± 0.97	1.42 ± 1.01
EL3, mean ± std	5.70 ± 3.74	1.81 ± 1.62	0.92 ± 0.77	0.90 ± 0.85

Table 3. Quantitative evaluation of view generation and measurements - ground truth: average of all users with outlier removal, evaluation set: training datasets with manual segmentation.

the automatic results are then compared to the ground truth. For the distance measurement evaluation, only the error in the distance measurement is considered. For the view evaluation, the position difference between central plane points (centers of the skull cut-contours) and the angle difference between the plane normals are used.

Based on a given ground truth, the mean error was determined for each user, each user group (EL1+2+3), and the automated estimation. In a first experiment, annotations from all observers (EL1+2+3) were used for ground truth generation. Resulting errors can be found in Tab. 1, where for each user group collective mean, std-dev. and the min/max of the individual mean errors are reported.

In a second experiment, focusing on the best annotations available, only the annotations of the EL3 users were used for ground truth generation and the evaluation was repeated (see Tab. 2). In a third experiment, the evaluation was repeated using not the test datasets but the training datasets with manually corrected segmentations (see Tab. 3). For all experiments, the view plane results are shown exemplarily for the TC plane. For the other planes similar results were obtained (not reported). The measurement results are shown for OFD and TCD. BPD measurements could not be evaluated in the same way, as different users used different definitions. To qualitatively evaluate the variability and to determine the causes of variability, the weighting function $w_p(t)$ can be inspected visually for different sets: overall with and without averaging and outlier removal and per user. The result can be seen in Fig. 3.

4. DISCUSSION AND CONCLUSION

In this paper, we proposed a way to learn view definitions and measurements. If the outlier-cleaned average of all manual annotations is considered as ground truth, the errors of the automatic TC plane detection (angle 4.71°, offset 1.20 mm) are lower than the errors of all individual users (EL1+2+3, user with lowest error: angle 6.33°, offset 1.33 mm) and much lower than the average error of all expert groups (see Tab. 1). The errors of the automatic TCD measurement are lower than the errors of most users, but the errors of the automatic OFD measurement are slightly higher than the error of most users.

To analyze in how far the users of lower expertise diminish the results of the experts, the ground truth was set as the average of the EL3 users. Using this ground truth, best automatic results are obtained when the training uses only EL3 annotations. However, the model trained on EL1+2 performed only slightly worse. When analyzing the different users, it can be seen that EL3 errors decrease slightly whereas EL1 and EL2 errors increase slightly, as could be expected.

To analyze the influence of segmentation errors, the analysis was repeated on the training data with manually corrected meshes. The errors of the TC plane are quite similar. The errors of automatic TDC and OFD measurement decrease by approx. 0.4 mm, indicating small segmentation inaccuracies when using the test data. Due to the small difference between results based on the automatic segmentation and on the manually corrected segmentation, it can be deduced that the segmentation accuracy is not a limiting factor for the automated estimation of standard views and biometrical measurements.

The reference annotations can show three types of variability: User variability, anatomical variability and technically induced variability (for example by segmentation errors). The visual inspection of the color coded weighting functions of learning the TC view in Fig. 3 allows some conclusions about the causes of the variability. As most variability can be removed by outlier removal and averaging per dataset, the largest source of variation seems to be user variability. Variability is much larger for EL1+2 compared to EL3. However, the weighting function is sharpest when outlier removal and averaging is done over all annotations. The planes determined from the three sets are very similar. When inspecting the visualization per user, it can be seen that there are different sources of user variability: Users A and E annotated consistently approximatly the same plane. User B determined the same plane but with an outlier, User C annotated a plane through the cerebellum, but always at a different angle. User D annotated consistently but overall at a different angle.

The segmentation model (in particular its mesh) is used in three different ways in this work. The first way is the actual segmentation, the second way is providing a frame of reference of averaging across different datasets and the third way is the encoding and decoding of views and measurements. The segmentation model itself was not validated clinically, instead the clinically relevant output was validated. This facilitated the collection of

manual reference annotations a lot and enabled high redundancy in input data. Visualization of the segmentation results enables a quick verification of success and the model could enable volumetric measurements if they become clinically accepted.

In this paper, we proposed a method to learn anatomical views and measurements from redundant but inconsistent data. We showed that learning from annotations of seven less experienced users gave a similar result as on the three most experienced users. The approach enables "'crowd-learning"' as the model can become better than individual users and several less experienced users can obtain a similar result as the most experienced users. The above automatic solution can be used for quality control, standardization and training purposes in fetal ultrasound screening. Further studies are needed to implement such a system in clinical practice.

We thank R. Cuingnet, R. Ardon, F. Roseman, D. Strassner, S. Heller, L. Pumphrey, L. Johnson for their contributions.

REFERENCES

1. International Society of Ultrasound in Obstetrics and Gynecology Education Committee: Sonographic examination of the fetal central nervous system: guidelines for performing the basic examination and the fetal neurosonogram. Ultrasound in Obstetrics and Gynecology **29**(1) (2007) 109-116
2. I. Sarris and C. Ioannou and E.O. Ohuma and D.G. Altman and L. Hoch and C. Cosgrove and S. Fathima and L.J. Salomon and A.T. Papageorghiou: International Fetal and Newborn Growth Consortium for the 21st Century. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. BJOG. (2013) Sep; 120 Suppl 2 33–7
3. M. Molloholli and S. Wanyonyi and V. Donadono and B. Kemp and T. Norris and F. Roseman and K. Edwards and R. Napolitano and A.T. Papageorghiou: Reproducibility of fetal brain measurements using 3D ultrasound. Ultrasound in Obstetrics and Gynecology **44**(S1) (2014) 10
4. M. Sofka and J. Zhang and S. Good and S. K. Zhou and D. Comaniciu: Automatic detection and measurement of structures in fetal head ultrasound volumes using sequential estimation and Integrated Detection Network (IDN). IEEE TMI **33**(5) (2014) 1054–70
5. R. Cuingnet and O. Somphone and B. Mory and R. Prevost and M. Yaqub and R. Napolitano and A. Papageorghiou and D. Roundhill and J.A. Noble and R. Ardon: Where is my baby? A fast fetal head auto-alignment in 3D-ultrasound. Biomedical Imaging (ISBI) (2013) 768–771
6. H. Chen and Q. Dou and D. Ni and J. Cheng and J. Qin and S. Li and P. Heng: Automatic Fetal Ultrasound Standard Plane Detection Using Knowledge Transferred Recurrent Neural Networks. Medical Image Computing and Computer-Assisted Intervention (2015), 9349 of the series Lecture Notes in Computer Science 507–514
7. J. Peters and O. Ecabert and C. Meyer and R. Kneser R and J. Weese: Optimizing boundary detection via simulated search with applications to multi-modal heart segmentation. Medical Image Analysis **14**(1) (2010) 70–84
8. O. Ecabert and J. Peters and H. Schramm and C. Lorenz and J. von Berg and M.J. Walker and M. Vembar and M.E. Olszewski and K. Subramanyan and G. Lavi and J. Weese: Automatic model-based Segmentation of the heart in CT images. IEEE TMI **27**(9) (2008) 1189–1201

QUALITY CONTROL IN FETAL ULTRASOUND

Quality control in fetal ultrasound has been demonstrated to improve measurements variability and therefore is essential in assessing the growth of fetal structures and creating fetal biometry charts.[41] Appropriate quality control strategies should include image storage, image reviewing and reproducibility assessment.[42] Despite years of practice the use of a comprehensive quality control strategy is lacking in most of the studies aimed to create fetal charts.[36-38] The review and judgment of ultrasound images is part of a standardisation process which contributes to the quality control strategy. The subjective assessment of an image ('poor or good') can affect the reproducibility between different observers. Furthermore, if not appropriate reproducibility tests are used results are not comparable. Different strategies have been tested in order to test reproducibility. The use of intraclass correlation coefficients is not the most appropriate method as it is a measure of agreement between different observers or different measurements methods rather than reproducibility. Two methods are the most appropriate to assess qualitative and quantitative reproducibility: the kappa coefficient (k) and the Bland- Altman plots respectively.[80, 81]

The introduction of an objective assessment has been demonstrated to be more reproducible than subjective assessment in assessing fetal biometry images,[82] CRL images (Appendix 6)[83] and pulsed wave Doppler images

(Appendix 7). CRL measurement images were assessed in 125 fetuses recruited into the FGLS of the INTERGOWTH-21st Project. Images were acquired according with specific criteria.[84] Images were assessed by two observers using either a subjective evaluation consisting of rating an image as acceptable or unacceptable, or using an objective evaluation based on six criteria. Overall agreement between the observers was higher for objective evaluation (95.2%, adjusted k: 0.904), than for subjective evaluation (77.6%, adjusted k: 0.552) (Appendix 6). A similar approach was used to assess the reproducibility of a proposed six points scoring criteria for pulsed wave Doppler images. 120 umbilical and uterine artery Doppler ultrasonographic images selected from the INTERBIO-21st Fetal Study (a multicentre, multiethinic study aimed to recruit high and low risk women to identify best predictors of abnormal pregnancy outcome) were assessed either using a subjective evaluation consisting of rating an image as acceptable or unacceptable or using the proposed objective evaluation. Overall agreement between assessors for the objective evaluation was higher (85%, adjusted k: 0.70), than for the subjective evaluation (73%, adjusted k: 0.47) (Appendix 7).

The assessment of reproducibility provides useful in creating fetal biometry standards. In fetal ultrasound assessment this is particularly important as accuracy of a measurement methods cannot be ascertained (comparing ultrasound measurements against a 'gold standard'). Pathology studies, comparison with other health technologies like magnetic resonance imaging or the use of phantoms are not reliable. The

solution highlighted in literature is to consider the most appropriate the method as that one with best reproducibility results.[85] We performed a study aimed to identify the best method for fetal head circumference (HC) biometry measurements by ultrasound using the approach described above (Appendix 8).[75] Different methods of calipers placement (BPD "outer to outer", BPD "outer to inner", OFD, HC using the ellipse facility) onto two different planes of acquisition (TT and TV) have been tested and reproducibility reported. 208 women recruited in the FGLS underwent extra measurements. More than 4400 measurements were taken. No major differences in reproducibility were observed with a standardised approach. The mean intraobserver and interobserver mean differences were < 1% (2.26 mm) and the 95% limits of agreement were < 8% (14.45 mm) for all fetal head measurements obtained in TV and TT planes. As a conclusion BPD "outer to outer", BPD "outer to inner", OFD, and HC using the ellipse facility can be acquired both in the TT or the TV plane. The use of BPD "outer to outer", OFD, and HC using the ellipse facility should be preferred as it allows fetal HC to be measured and compared with neonatal HC. TT plane is preferable as international standards in this plane are available; however, measurements in the TV plane can be plotted on the same standards.

Appendix 6: Image-scoring system for crown–rump length measurement

Ultrasound Obstet Gynecol 2014; **44**: 649–654
Published online in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.13376

Image-scoring system for crown–rump length measurement

S. Z. WANYONYI*, R. NAPOLITANO*, E. O. OHUMA*, L. J. SALOMON†‡ and A. T. PAPAGEORGHIOU*

**Nuffield Department of Obstetrics & Gynaecology, University of Oxford, John Radcliffe Hospital, Oxford, UK; †Maternité Necker-Enfants Malades, AP-HP, Université Paris Descartes, Paris, France; ‡Société Française pour l'Amélioration des Pratiques Echographiques (SFAPE), Paris, France*

KEYWORDS: crown–rump length; image scoring; objective evaluation; reproducibility; subjective evaluation

ABSTRACT

Objective *To develop and evaluate an objective image-scoring system for crown–rump length (CRL) measurements and to determine how this compares with subjective assessment.*

Methods *A total of 125 CRL ultrasound images were selected from the database of the International Fetal and Newborn Growth Consortium for the 21st Century study group. Two reviewers, who were blinded to the operators' and to each others' results, evaluated all images both subjectively and objectively. Subjective evaluation consisted of rating an image as acceptable or unacceptable, while objective evaluation was based on six criteria. Reviewer differences for both the subjective and objective evaluations were compared using percentage of agreement and adjusted kappa values.*

Results *The distribution of individual scores and differences between subjective and objective evaluation for the two reviewers was similar. Overall agreement between the reviewers was higher for objective evaluation (95.2%; adjusted κ, 0.904), than for subjective evaluation (77.6%; adjusted κ, 0.552). There was a high level of agreement for horizontal position (κ = 0.951), magnification (κ = 0.919), visualization of crown and rump (κ = 0.806) and caliper placement (κ = 0.756), while agreement for mid-sagittal section (κ = 0.629) and neutral position (κ = 0.565) were moderate and poor, respectively.*

Conclusion *The proposed six-point scoring system for CRL image rating is more reproducible than is subjective evaluation and should be considered as a method of quality assessment and audit. Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.*

INTRODUCTION

Sonographic measurement of fetal crown–rump length (CRL) is an important aspect of pregnancy care and is used for accurate estimation of gestational age[1]. This not only reduces the incidence of labor induction in prolonged pregnancy[2], but also guides decisions regarding other obstetric interventions such as prenatal testing, growth assessment and timing of delivery[3]. Accurate CRL measurement is also necessary for correct interpretation of the variables used in first-trimester screening for chromosomal abnormalities, namely nuchal translucency thickness (NT), pregnancy-associated plasma protein-A and free beta-human chorionic gonadotropin[4]. Consequently, measurement of CRL has become routine in most developed countries and constitutes one of the commonest imaging investigations carried out. It is, therefore, somewhat surprising that quality assessment for this parameter is lacking. While some criteria for an optimum CRL measurement have been described, we have been unable to find any previous studies assessing the impact of such scoring systems in clinical practice. Similarly, in the research setting it is important to note that substantial heterogeneity exists regarding the criteria for accurate image acquisition when CRL measurements are used for creating charts for pregnancy dating[5].

Standard teaching suggests that a midline sagittal section of the entire fetus should be obtained with optimum magnification. The crown and rump should be clearly visible for correct placement of the electronic calipers at both ends and the fetus should be in a neutral position (not flexed or hyperextended). The fetus should be horizontal with the line connecting the crown and rump positioned at about 90° to the ultrasound beam[4,6,7]. However, like any other sonographic measurement, these criteria are prone to variation, depending on the skill of the sonographer and the fetal position. To mitigate this, guidelines have been proposed to improve the consistency and reproducibility of fetal images, and these are mainly based on standard taught methods. For instance, the National Health Services Fetal Screening Programme (NHS-FASP) in the UK has developed an algorithm

Correspondence to: Dr A. T. Papageorghiou, Nuffield Department of Obstetrics & Gynaecology, University of Oxford, John Radcliffe Hospital, Oxford, UK (e-mail: aris.papageorghiou@obs-gyn.ox.ac.uk)
Accepted: 18 March 2014

 ORIGINAL PAPER

for CRL and NT measurements as part of the Down syndrome screening program[6,8].

The use of image scoring for fetal biometry in the second trimester has been evaluated and found to be both feasible and reproducible[9]. Other studies have also evaluated quality control and image-scoring systems for NT and nasal bone measurements in the first trimester as part of trisomy 21 screening[4,10–14]. The reliability of these image-scoring systems has seen their validation and use in research, quality control and clinical audits. The value of CRL image scoring has not been formally assessed, although this measurement is more common in clinical practice. This means that subjective assessment is used to determine the suitability of images. There is therefore a need to develop and validate a similar scoring system for CRL.

The aim of this study was to develop and evaluate an image-scoring system for CRL measurement and to determine how it compares with subjective assessment.

METHODS

Static images of the crown–rump area were obtained using transabdominal ultrasound, by trained sonographers in eight countries that are part of the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st)[15]. The sonographers had undergone rigorous training in CRL measurement and submitted images as part of quality control[16]. From the image database, stored electronic images were selected using a computerized random number generator; in addition, in order to increase the number of less-than-optimal images, a manual search was performed and these were subjectively added by a third observer (A.P.), who did not take part in the evaluation.

All images were evaluated both subjectively and objectively by two independent reviewers (A and B). This was done using similar methodology to that used for studies assessing second-trimester biometry and first-trimester NT and nasal bone measurements[9–14]. For subjective evaluation, the reviewers were required to rate an image as either acceptable or unacceptable when the image was presented to them. Objective evaluation consisted of criterion-based scoring that was developed based on established standards for CRL measurement[4,6,7] (Table 1, Figure 1). Components of the six-point criteria (Table 1) were derived from practice guidelines developed by the International Society of Ultrasound in Obstetrics and Gynecology, The Fetal Medicine Foundation and the UK National Health Service Fetal Screening Programme[6,7,17]. Each correct component of the image scored one point and a score of zero was given if the criterion was not met. All criteria were accorded equal weight, thus the maximum possible score was 6.

Statistical analysis

A total of 125 images would be needed to detect a 10% difference between two reviewers with 90% power, assuming an interobserver agreement rate of 80%, based on a similar study by Salomon *et al.*[9] for second-trimester fetal biometry.

For subjective evaluation, we compared the number and proportion of images considered acceptable or unacceptable by the two reviewers. For objective evaluation a cut-off of ≥ 4 was used to define an acceptable image, while any image with a score ≤ 3 was considered unacceptable. Therefore, agreement between reviewers was calculated by grouping the images into two score ranges (0–3 and 4–6). The distribution of the objective scores for each reviewer was determined and the differences compared using the Wilcoxon signed-rank test.

Intra- and inter-reviewer agreement between the objective score and subjective assessment were determined using prevalence-adjusted, bias-adjusted kappa (PABAK) coefficients[18]. The reproducibility of each independent criterion was also tested. We used Cronbach's alpha to assess the inter-item reliability for each of the six criteria and evaluated the impact each individual criterion had on the overall scale.

RESULTS

A total of 125 images were evaluated, with a mean CRL measurement of 60.5 ± 9.9 mm. Both subjective and objective evaluation was possible in all cases by both reviewers. Differences in agreement between the two reviewers were less than 10% for both subjective and objective ratings.

On subjective evaluation, 22 (17.6%) and 18 (14.4%) images were found unacceptable by Reviewers A and

Table 1 Image-scoring criteria for crown–rump length (CRL) measurement

Criterion	*Description*
Mid-sagittal section	Midline facial profile, fetal spine and rump should all be visible in one complete image
Neutral position	There should be fluid visible between the chin and the chest of the fetus and the 'profile line'* should form an acute angle with the CRL line before the rump
Horizontal orientation	Fetus should be horizontal with line connecting crown and rump positioned between 75° and 105° to ultrasound beam
Crown and rump clearly visible	Crown and rump should both be clearly visible
Correct caliper placement	Intersection of calipers should be on outer border of skin covering skull and outer border of skin covering rump
Good magnification	Fetus should fill more than two-thirds of image, clearly showing crown and rump

*Profile line is a line connecting brow and tip of nose (Figure 1c).

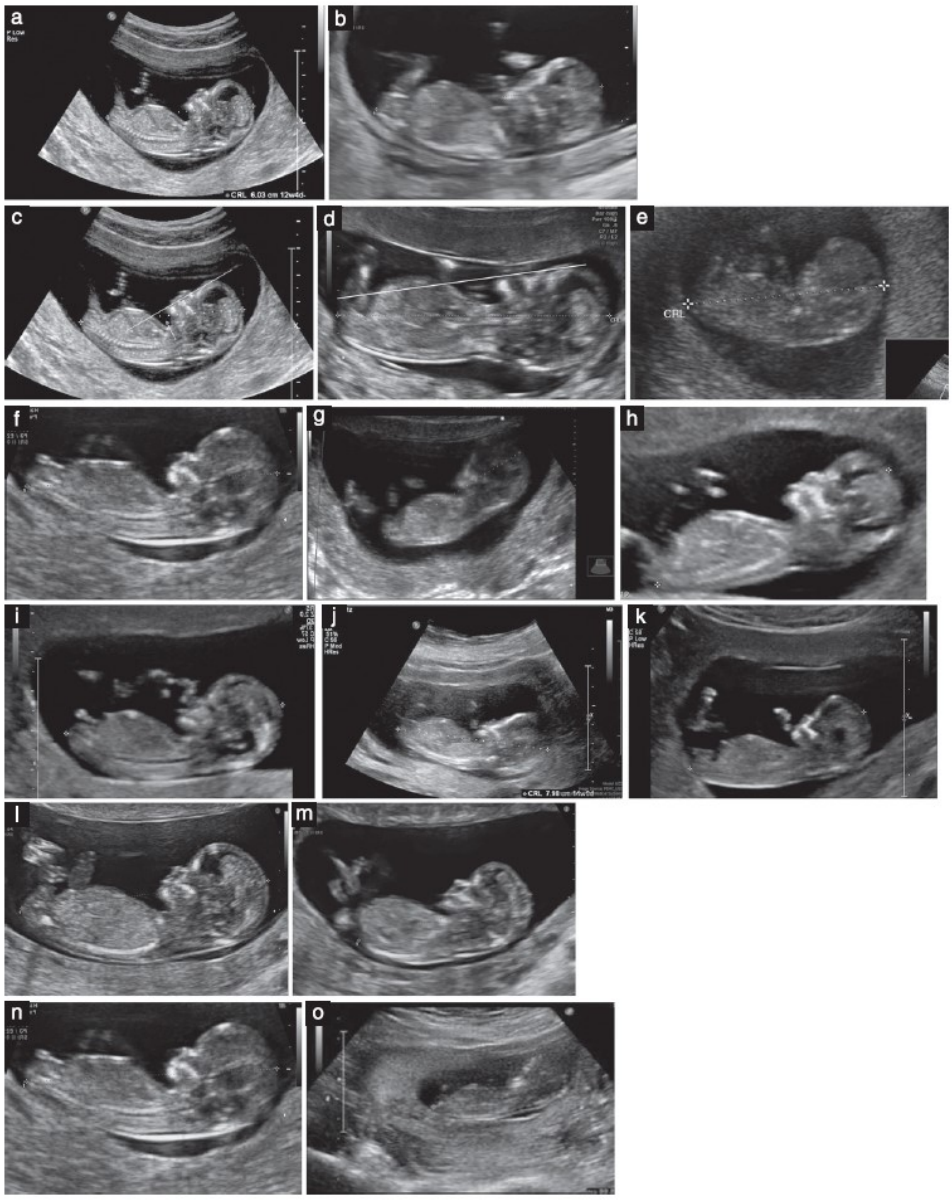


Figure 1 Examples illustrating image-scoring criteria for objective assessment of fetal crown–rump length (CRL) measurement. (a) A precise mid-sagittal section, showing that facial profile, spine, crown and rump are clearly visible. (b) Image not mid-sagittal; although crown, rump and spine are visible, facial profile was not captured correctly. (c) Neutral position, showing that profile line (solid line) forms an acute angle with CRL line (dotted line) and there is fluid between the chest and chin (arrow). (d) Hyperextension (profile line (solid line) runs almost parallel to CRL line (dotted line)). (e) Hyperflexion, showing very little fluid between chin and chest. (f) Horizontal position, showing spine at almost 90° to ultrasound beam. (g,h) Images showing spine at ~45° (g) and ~15° (h), with respect to ultrasound beam. (i) Image showing crown and rump clearly visible with fluid at either end. (j) Image showing edge of crown that is not clearly delineated. (k) Image showing that there is no fluid/space between the rump and uterine wall, making precise caliper placement difficult. (l) Image showing calipers placed correctly on crown and rump. (m) Image showing caliper placement at rump that is lower than it should be. (n) Image showing good magnification. (o) Image showing fetus occupying less than two-thirds of screen.

B, respectively. The overall inter-reviewer agreement for subjective evaluation was 77.6% (adjusted κ, 0.55 (95% CI, 0.41–0.70)).

The images rated subjectively as acceptable by both reviewers predominately had a score of 6 (Table 2). The distribution of the individual objective score for images subjectively rated unacceptable (Table 2) and the intra- and inter-reviewer agreement (Table 3) were similar for both reviewers. There was no significant difference in the distribution of objective scores between the two reviewers (median score 6 (range, 2–6); $P = 0.54$) (Figure 2). The inter-reviewer agreement for objective image rating was 95.2% (adjusted κ, 0.90 (95% CI, 0.83–0.98)).

For objective assessment, the degree of agreement for the individual criteria between the two reviewers was determined. The inter-reviewer agreement was highest for horizontal orientation (97.6%) and good magnification (95.9%) and lowest for neutral position of the image (78.2%) (Table 4). The scale derived from our chosen items estimated correlation between it and the underlying factor it measured as 0.7521 (Cronbach's $\alpha = 0.5656$) and 0.6210 (Cronbach's $\alpha = 0.3857$) for Reviewers A and B, respectively (Table 5).

DISCUSSION

Subjective CRL image rating is commonly used in clinical practice, but this study has shown that, unlike objective evaluation, it has a low rate of inter-reviewer reproducibility. Conversely, a criterion-based scoring system, like the one used in this study, has been shown to be more reliable and reproducible. Similar findings have been demonstrated in second-trimester fetal biometry[9]. Studies on NT and nasal bone measurements have also found objective assessment to be more reproducible than is subjective judgment[10–14].

Table 2 Distribution of objective image score for each subjective image rating for crown–rump length measurement for Reviewers A and B

	Objective image score					
Subjective image evaluation	*1*	*2*	*3*	*4*	*5*	*6*
Unacceptable A	—	2 (1.6)	—	18 (14.4)	2 (1.6)	—
Acceptable A	—	—	—	16 (12.8)	13 (10.4)	74 (59.2)
Unacceptable B	—	—	3 (2.4)	10 (8.0)	3 (2.4)	2 (1.6)
Acceptable B	—	—	1 (0.8)	13 (10.4)	24 (19.2)	69 (55.2)
Unacceptable by both A and B	—	—	—	2 (1.6)	—	—
Acceptable by both A and B	—	—	—	3 (2.4)	7 (5.6)	47 (37.6)

Data are given as *n* (%).

Table 3 Intra- and inter-reviewer agreement between objective and subjective evaluation of image rating for crown–rump length measurement

Comparison	*Agreement (%)*	*Adjusted kappa (95% CI)*
Objective A /Objective B	95.2	0.90 (0.83–0.98)
Subjective A /Subjective B	77.6	0.55 (0.41–0.70)
Subjective A /Objective A	84.8	0.70 (0.57–0.82)
Subjective B /Objective B	88.0	0.76 (0.64–0.87)
Subjective A /Objective B	83.2	0.66 (0.53–0.80)
Subjective B /Objective A	86.4	0.72 (0.61–0.85)

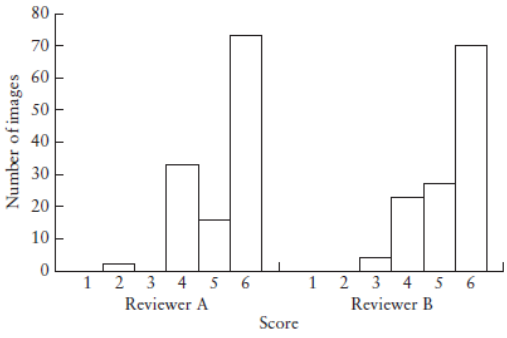


Figure 2 Distribution of objective scores for assessment of fetal crown–rump length measurement by Reviewers A and B.

Table 4 Adjusted kappa and percentage of agreement for each individual criterion of image rating for crown–rump length measurement

Criterion	*Agreement (%)*	*Adjusted kappa (95% CI)*
Mid-sagittal section	81.5	0.629 (0.492–0.766)
Neutral position	78.2	0.565 (0.419–0.710)
Horizontal orientation	97.6	0.951 (0.897–1.00)
Crown and rump clearly visible	90.3	0.806 (0.702–0.911)
Correct caliper placement	87.8	0.756 (0.640–0.872)
Good magnification	95.9	0.919 (0.849–0.988)

Table 5 Inter-item reliability tests for image scoring criteria for crown–rump length measurement

	Cronbach's alpha for excluding each item individually	
Parameter	*Reviewer A*	*Reviewer B*
Mid-sagittal section	0.2514	0.0272
Neutral position	0.2559	0.0592
Horizontal orientation	0.5860	0.4169
Crown and rump visible	0.5565	0.4350
Correct caliper placement	0.5565	0.3905
Good magnification	0.5908	0.4481
Alpha coefficient for all six items	0.5656	0.3857

There was a good level of agreement for assessing horizontal orientation, correct caliper placement, visibility of the crown and rump and good magnification of images (PABAK > 0.7). However, agreement was only moderate for the correct detection of midline sagittal section, and poor for the criterion of neutral position. According to the criteria used in this study, for an image to be regarded as neutral, the reviewer was required to draw a line connecting the brow and tip of the nose ('profile line') and demonstrate that this formed an acute angle with the CRL line before the rump (Figure 1c), besides visualizing fluid between chin and chest. Hyperextension was defined by a line that runs almost parallel to the CRL line, meaning the profile line does not cross the CRL before the rump. This was to ensure objectivity in the determination of a neutral position as opposed to mere observation, which is prone to bias. Despite these efforts, there was significant variation in the scores and a rather poor level of agreement for this criterion. However, criteria such as horizontal orientation and magnification, which were easier to determine, were associated with high levels of agreement.

Some of the parameters used in the study may be interrelated, and this could have influenced the distribution of the final scores. For instance, a correct mid-sagittal plane requires the rump to be clearly visible and the same is required for correct positioning of the calipers; therefore the absence of either the crown or rump in an image would mean a score of zero for mid-sagittal section, visualization of crown and rump and caliper placement. Despite this limitation, the scoring system used in this study was simple and objective, based on recognized criteria[6,7,17], and easily applicable for audit and quality control of CRL images. We would therefore recommend it for audit and quality control of CRL images. Images were rated as either acceptable or unacceptable, deliberately avoiding an intermediate score, since such a category often introduces uncertainty.

Accurate CRL measurement has significant implications on subsequent management of the pregnancy[2,3], despite the controversies in studies comparing the use of CRL and obstetric outcomes[19,20]. Absence of a quality-control policy could have affected the interpretation of previous results and may be one explanation for the degree of variability in gestational-age estimation using different dating equations[5]. We have previously demonstrated that the policy of adopting strict standardization and quality-control procedures in fetal biometry can lead to a measurable improvement in interobserver variability in the setting of a multicenter study, and it is likely that such an approach would also be relevant in the measurement of fetal CRL[21].

One perceived limitation of our study may be the fact that equal weight was given to each of the six measures. The main drawback of using this approach is that it can result in some images being erroneously labeled as unacceptable: for example, an image with poorly placed calipers and a non-visible rump scored 3 and was therefore labeled unacceptable, while it may be rated as acceptable if more weight were placed on the acquisition of a horizontal and mid-sagittal plane and less on correct caliper placement. Conversely, an image may not be in a perfect mid-sagittal plane (for example, Figure 1b), but it may meet all the other criteria. Such an image would score 4 and be rated acceptable, but if the mid-sagittal section had a stronger rating then it may be regarded as unacceptable. This is a recognized shortcoming in studies of a similar nature whereby lower rating in one criterion may be masked by the final score. This can be overcome by a weighted scoring system, which could be incorporated into audit programs and compared with the proposed scoring system to determine its feasibility for use in clinical practice. Similarly, we acknowledge the fact that the cut-offs we set in the study were arbitrary, and based on what we thought was reasonable. This may have resulted in many images passing the test, despite being found to be unacceptable on subjective evaluation (Table 2, Figure 2). There may be a need to assess different cut-offs in future, especially considering the varying degree of experience among sonographers. Our study was also done in a purely research setting with trained sonographers who were operating to similar standards; whether these results would be applicable on a larger scale or in a clinical setting is not clear. Nevertheless, we believe that quality-control processes in ultrasonography are an essential requirement not just for research studies on fetal biometry but also in clinical practice.

This study was drawn from a large, multicenter population, adding credibility to our findings. Image acquisition was done by experienced and well-trained sonographers and the whole process was standardized. The study was also well powered to detect a significant difference and the reviewers were blinded to the identity of the operators, hence minimizing bias. Since kappa statistics have been shown to give paradoxical results that are dependent on the prevalence of the condition and are prone to sampling bias[22], we used the PABAK as described by Bennett *et al.*[18] to give a more reliable degree of agreement for both the subjective and objective evaluation. Efforts were made to set tighter criteria; for example we checked for fetal neutrality more objectively, although this could have introduced a degree of complexity and increased the time spent on assessments. It is also worth noting that the scale derived from our chosen items was reasonable as determined by Cronbachs' alpha. This additional effort needs further evaluation against the improvements made in screening and audit programs.

In conclusion, we propose the six-point objective quality-scoring system for the assessment of ultrasound images of fetal CRL. This objective method of image scoring is more reliable and reproducible than is subjective evaluation and should be the preferred method used in quality control and training, as the introduction of such scoring criteria has the potential to improve the quality of ultrasound. Larger-scale studies in a clinical setting are needed to assess the impact and the possible need of a weighted scoring system.

ACKNOWLEDGMENTS

We are grateful to the pregnant women who participated in this study, and to the sonographers at the eight study sites for submitting their logbooks and images for evaluation. This study is part of the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) project, which is supported by a grant from The Bill & Melinda Gates Foundation to the University of Oxford, for which we are grateful.

REFERENCES

1. Robinson HP. A critical evaluation of sonar crown–rump length measurements. *Br J Obstet Gynaecol* 1975; **82**: 702–710.
2. National Collaborating Centre for Women's and Children's Health. *Antenatal care: routine care for the healthy pregnant woman*. www.nice.org.uk/nicemedia [Accessed 22 February 2013].
3. Taipale P, Hiilesmaa V. Predicting delivery date by ultrasound and last menstrual period in early gestation. *Obstet Gynecol* 2001; **97**: 189–194.
4. Snijders RJM, Noble P, Sebire N, Souka A, Nicolaides KH. UK multicentre project on assessment of risk of trisomy 21 by maternal age and fetal nuchal-translucency thickness at 10–14 weeks of gestation. Fetal Medicine Foundation First Trimester Screening Group. *Lancet* 1998; **352**: 343–346.
5. Napolitano R, Dhami J, Ohuma EO, Ioannou C, Conde-Agudelo A, Kennedy SH, Villar J, Papageorghiou AT. Pregnancy dating by fetal crown–rump length: a systematic review of charts. *BJOG* 2014; **121**: 556–565.
6. NHS Fetal Screening Programme (NHSFASP). Recommended criteria for measurement of fetal crown rump length (CRL) as part of combined screening for Trisomy 21 within the NHS in England, 2012. www.fetalanomaly.screening.nhs.uk/standardsandpolicies [Accessed 27 February 2013].
7. Salomon LJ, Alfirevic Z, Bilardo CM, Chalouhi GE, Ghi T, Kagan KO, Lau TK, Papageorghiou AT, Raine-Fenning NJ, Stirnemann J, Suresh S, Tabor A, Timor-Tritsch IE, Toi A, Yeo G. ISUOG Practice Guidelines: performance of first-trimester fetal ultrasound scan. *Ultrasound Obstet Gynecol* 2013; **41**: 102–113.
8. Salomon LJ, Ville Y. Quality control of prenatal ultrasound. *Ultrasound Rev Obstet Gynecol* 2005; **5**: 297–303.
9. Salomon LJ, Bernard JP, Duyme M, Doris B, Mas N, Ville Y. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. *Ultrasound Obstet Gynecol* 2006; **27**: 34–40.
10. Herman A, Maymon R, Dreazen E, Caspi E, Bukovsky I, Weinraub Z. Nuchal translucency audit: a novel image-scoring method. *Ultrasound Obstet Gynecol* 1998; **12**: 398–403.
11. Snijders RJ, Thom EA, Zachary JM, Platt LD, Greene N, Jackson LG, Sabbagha RE, Filkins K, Silver RK, Hogge WA, Ginsberg NA, Beverly S, Morgan P, Blum K, Chilis P, Hill LM, Hecker J, Wapner RJ. First-trimester trisomy screening: nuchal translucency measurement training and quality assurance to correct and unify technique. *Ultrasound Obstet Gynecol* 2002; **19**: 353–359.
12. McLennan A, Schlutter PJ, Pincham V, Hyett J. First-trimester fetal nasal bone audit: evaluation of a novel method of image assessment. *Ultrasound Obstet Gynecol* 2009; **34**: 623–628.
13. Thia EW, Wei X, Tan DT, Lai XH, Zhang XJ, Oo SY, Yeo GS. Evaluation of an objective method of image assessment for first-trimester nasal bone. *Ultrasound Obstet Gynecol* 2011; **38**: 533–537.
14. Staboulidou I, Wüstemann M, Vaske B, Scharf A, Hillemanns P, Schmidt P. Interobserver variability of the measurement of fetal nasal bone length between 11 + 0 and 13 + 6 gestation weeks among experienced and inexperienced sonographers. *Ultraschall Med* 2009; **30**: 42–46.
15. The International Fetal and Newborn Growth Consortium. INTERGROWTH 21st http://www.intergrowth21.org.uk [Accessed 22 February 2013].
16. Sarris I, Ioannou C, Ohuma EO, Altman DG, Hoch L, Cosgrove C, Fathima S, Salomon LJ, Papageorghiou AT; International Fetal and Newborn Growth Consortium for the 21st Century. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. *BJOG* 2013; **120** (Suppl 2): 33–37.
17. The Fetal Medicine Foundation 11–13 weeks on line education aid http://www.fetalmedicine.com/fmf/online-education/01-11-136-week-scan/ [Accessed 12 August 2012].
18. Bennett E, Albert R, Goldstein A. Communications through limited-response questioning. *Public Opinion Q* 1954; **18**: 303–308.
19. Harrington D, MacKenzie IZ, Thompson K, Fleminger M, Greenwood C. Does a first trimester dating scan using crown rump length measurement reduce the rate of induction of labour for prolonged pregnancy? An uncompleted randomised controlled trial of 463 women. *BJOG* 2006; **113**: 171–176.
20. Bennett KA, Crane JM, O'Shea P, Lacelle J, Hutchens D, Copel JA. First trimester ultrasound screening is effective in reducing postterm labor induction rates: a randomized controlled trial. *Am J Obstet Gynecol* 2004; **190**: 1077–1081.
21. Sarris I, Ioannou C, Dighe M, Mitidieri A, Oberto M, Qingqing W, Shah J, Sohoni S, Al Zidjali W, Hoch L, Altman DG, Papageorghiou AT; International Fetal and Newborn Growth Consortium for the 21st Century. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. *Ultrasound Obstet Gynecol* 2011; **38**: 681–687.
22. Banerjee M, Fielding J. Interpreting kappa values for two-observer nursing diagnosis data. *Res Nurs Health* 1997; **20**: 465–470.

Appendix 7: Image-scoring system for umbilical and uterine artery pulsed wave Doppler measurement

This is study is under the review of the Scientific Steering Committee of the INTERGORWTH-21st Project and it might undergo substantial review of the data before the publication in the journal peer reviewed process.

ABSTRACT

Objective: To develop an objective, image scoring system for pulsed wave Doppler measurement in obstetrics based upon six predefined criteria and evaluate how the system compares with subjective assessment.

Methods: A total of 120 umbilical and uterine artery Doppler ultrasonographic images were randomly selected from the INTERBIO-21st Study database. Two assessors retrospectively evaluated the images objectively using the six-point image-scoring system and subjectively, in a blinded fashion. Subjective assessment consisted of classifying an image as acceptable or unacceptable. The percentage of agreement and a Kappa statistic of the two assessors were compared.

Results: Overall agreement between assessors for the objective evaluation was higher (agreement=85%, adjusted k=0.70), than for the subjective evaluation (agreement=73%, adjusted k=0.47). The levels of

agreement for the six criteria were: anatomical site (adjusted k=0.97), sweep speed (adjusted k=0.88), magnification (adjusted k=0.77), velocity scale (adjusted k=0.68), image clarity (adjusted k=0.68), and angle of insonation (adjusted k=0.65).

Conclusion:

The proposed six-point image-scoring system for umbilical and uterine artery pulsed wave Doppler measurement is more reliable and reproducible than subjective evaluation. We suggest the system should be the preferred method for quality control, auditing and teaching.

INTRODUCTION

Doppler ultrasonography is a safe and non-invasive way of evaluating blood flow in vivo[1], which plays an important role in identifying and managing pregnancies at greatest risk of preeclampsia, intrauterine growth restriction (IUGR), fetal distress in labour, and neonatal morbidity[2]. In pregnancies with suspected IUGR and/or hypertensive disease Doppler ultrasound is associated with a reduced number of perinatal deaths and unnecessary obstetric interventions[3].

The International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) practice guidelines for the use of Doppler ultrasonography in obstetrics recommend considering a number of factors to minimise measurement errors and improve reproducibility, including fetal breathing and body movements, colour flow mapping, optimal angle of insonation, horizontal sweep speed, gain, and pulsed wave frequency[4].

These factors aim to improve the reproducibility of measurements, and recognize the fact that imaging quality in Doppler is important: for example a change in the angle of insonation of only 10° corresponds to a 2% velocity error whilst a 20° angle corresponds to 6% error[4].

Nevertheless, despite widespread use, objective quality control criteria for the use of Doppler in pregnancy are lacking. Although different techniques and some optimum criteria have been described, we have been unable to find any previous studies on formal scoring systems or objective assessment. The use of scoring systems, for example in fetal biometry[5], nuchal translucency or measurement of crown rump length[6-9], have been demonstrated to be feasible and more reproducible between assessors than subjective assessment. This can play an important role in training, auditing and quality control of sonographers.

The aim of this study was to develop an image-scoring system for Doppler ultrasound and to evaluate how it compares with a subjective assessment.

METHODS

Stored images of uterine and umbilical artery pulsed wave Doppler assessment were obtained by the ultrasound quality control unit of the INTERBIO-21st Fetal Study (a study under the umbrella of the INTERGROWTH-21st Project)[10, 11].

The INTERBIO-21st Fetal Study is a multicentre, multiethnic, project conducted in centres taking part in the INTERGROWTH-21st Project and in other resource-poor setting countries, which aimed to evaluate newborn phenotypes so as to understand better the relationship between the causes of FGR and preterm birth syndromes. Data collected include maternal and neonatal biological samples, fetal ultrasound growth patterns, pregnancy and postnatal outcomes. Inclusion criteria in the INTERBIO-21st Fetal Study are reported elsewhere (www.interbio21.org.uk). Briefly, women were more than 18 years old with BMI < 35 and having natural conception. Pregnancies were dated according with crawn rump length (CRL) measurement between 9+0 and 14+0 weeks. Serial bidimensional ultrasound scans were performed every 5±1 weeks, from 14+0 to 41+6 weeks' gestation, and images were stored for later analysis. Other than measurements obtained as in the FGLS of the INTERGROWTH-21st Project one uterine and umbilical artery pulsed wave Doppler measurement for each scan was obtained according to the corresponding FGLS Protocol and Ultrasound Operations Manual (the

INTERBIO-21st Consortium, INTERBIO-21st Study Protocol, Oxford, October 2012. (www.interbio21.org.uk)).

Women taking part in the INTERBIO-21st Fetal Study had Doppler assessment using commercially available ultrasound machines (Philips HD- 9, Philips Ultrasound, Bothell, WA, USA, equipped with curvilinear abdominal transducers: C5-2, C6-3 and V7-3 and GE Voluson E8, GE Healthcare, Zipf, Austria equipped with RAB 4-8-D probe). Briefly, the protocol recommended using transabdominal ultrasound with appropriate setting adjustment. Assessment of uterine artery blood flow was performed between 19+0 and 23+6 weeks. The artery on each side was identified using color flow mapping at the apparent crossover with the external iliac artery. Pulsed wave Doppler was then used using an appropriate gate size and minimum angle of insonation to obtain 4 - 6 similar waveforms. After angle correction, the Pulsatility index (PI), Resistance index (RI) are measured and presence of an early diastolic "notch" recorded for each vessel (defined as a clearly defined upturn of the flow velocity waveform at the beginning of diastole in all waveforms)[12]. For umbilical artery Doppler, carried out after 24 weeks, the signal was obtained from a free loop of the umbilical cord, ensuring fetal quiescence (absence of significant limb or breathing movements). Color Doppler was used to identify the vessel and the pulsed wave Doppler gate used to obtain 4 - 6 consistent waveforms. PI, RI, Systolic / Diastolic ratio (S/D) were measured via auto tracing of three or more consecutive similar waveforms, from the beginning of the systolic to the end of the diastolic

signal. In case where this was not possible, a manual trace can be used for these calculations. End diastolic flow (EDF) was reported as present, absent or reversed.

The Doppler images were taken by trained sonographers in five different countries (Brazil, Kenya, South Africa, Pakistan and Thailand) who underwent a specific standardisation process similar to that for fetal biometry[13]. For the purposes of this study a pre-specified number of images were selected at random and retrieved from the database.

Subjective and objective evaluation of all images was then performed by two independent assessors (A and B). The assessors were blinded to each other's results and also to the sonographer who took the original image. For subjective evaluation the reviewers were asked to rate the images as either "acceptable" or "unacceptable" based on visual assessment. In the objective evaluation a new six-point image-scoring criterion was developed based on recommended and established standards for Doppler measurements[4, 11, 12], (Table 1). Assessors gave one point to each criterion if it was satisfied and zero points if the criterion was not satisfied (Figure 1-4). Therefore the total maximum score an image could achieve was six. All criteria were accorded equal weight. Of note is that the main components of scoring criteria were a product of well established guidelines[4]. For the purpose of the comparison of the subjective versus the objective score, scores of 4-6 were considered as

“acceptable” while those scoring 3 or less were classed as “unacceptable” (Figure 5).

Statistical analysis

Based on findings from previous studies[5], 14 we determined that a total of 120 images would be needed to detect a 10% difference between two assessors with 90% power, assuming an inter-observer agreement rate of 80%.

Agreement between the two assessors based on subjective and objective results were assessed independently and also compared between them. Prevalence-adjusted, bias-adjusted kappa coefficients were used to determine the intra- and inter-assessor agreement between the objective score and subjective assessment.

RESULTS

A total of 120 umbilical and uterine artery pulsed wave Doppler images were examined; both assessors were able to undertake the subjective and objective evaluation on all images. The percentage of agreement between two assessors was 73.3% for subjective and 85% for objective evaluation.

For subjective assessment, 47 (39%) were classified as unacceptable by assessor A, 23 (19%) by assessor B and as a result 19 (15.8%) images classified as unacceptable by both assessors. This resulted in overall inter-reviewer agreement of 73.3% [adjusted kappa, 0.47 (95% CI, 0.31-0.62)].

In the images rated subjectively as acceptable by both assessors, all scored 3 and above in the respective objective assessment, with the majority scoring 5 and 6 (Table 2). Conversely, none of the images deemed subjectively unacceptable scored 6 in objective assessment and only one scored 5 in the objective assessment of reviewer B (Table 2). The inter-assessor agreement for objective rating was 85% [adjusted kappa, 0.70 (95% CI, 0.58 – 0.83)].

We also evaluated the degree of agreement for the individual criteria between the two assessors in the objective assessment. The interassessor agreement was highest for the anatomic site (98.3%) and sweep speed (94.2%) and lowest for the angle of insonation (82.5%) (Table 3).

DISCUSSION

Pulsed waved Doppler measurement is used widely in clinical practice and of particularly importance in high risk pregnancies. The lack of standardisation and quality control in acquisition of Doppler data can lead to heterogeneous results and methodology bias in fetal ultrasound studies[12, 15]. Furthermore, the absence of a quality control can significantly affect the clinical practice[6].

Quality assessment of Doppler images can be undertaken subjectively, by judging an image to be acceptable or not; or objectively by using criteria that have been derived for this purpose. What we show in this study is that objective, criterion-based scoring has been demonstrated to be more reliable and reproducible than subjective evaluation, with the former associated with substantial agreement, rather than the moderate agreement with subjective evaluation[16]. This is an important finding as better quality assessment could allow better identification of sonographers who could benefit from further training and could allow focused feedback. Poor technique is unlikely to “normalise” raised uterine / umbilical PI (or indeed normalise absent or reversed end diastolic flow), but may make normal blood flow appear less normal. Therefore, the most likely impact of this may be in reducing the number of false positive (falsely abnormal) results. In addition, the adoption of objective quality control is likely to

reduce measurement variability, and this has been shown in other areas of ultrasound[13].

We found that with objective scoring the level of agreement was particularly high for assessment of anatomic site, sweep speed and magnification (adjusted kappa > 0.7); while it was good for the assessment of the angle of insonation, image clarity and velocity scale (adjusted kappa: 0.65, 0.68, 0.68, respectively). This was much lower with subjective scoring (adjusted kappa: 0.47). This is in keeping with previous studies that found objective scoring was more reproducible than subjective assessment in second-trimester fetal biometry, NT, nasal bone and CRL measurement[5, 7, 8, 14, 17, 18]. In the case of fetal biometry, a quality-control and standardisation process led to a measurable improvement in inter-observer variability in the settings of a multicentre study[13].

Previous studies have demonstrated that use of umbilical artery Doppler in the management of pregnancies suspected with intrauterine growth restriction and or hypertensive disease of pregnancy may reduce the number of perinatal deaths and unnecessary obstetric interventions[19]. Similarly, uterine artery Doppler screening is effective at predicting pregnancies at risk of adverse outcome and in selecting cases for more intense surveillance[20, 21]. In view of this, the accurate measurement of pulsed-waved Doppler for umbilical and uterine artery takes a particular importance. We believe that the scoring system used in this study is simple enough for clinical use, and it was derived from well established

guidelines[4]. It can be used easily in ultrasound departments for teaching, auditing and quality control.

Our study was well powered to detect a significant difference. It was from a large, multicentre population and performed by trained sonographers who were blind to each other and to the reviewers. Adjusted kappa was used to minimise bias and to give a more reliable agreement for the criteria used in subjective or objective assessment[22].

There are limitations of this study. In the objective assessment all criteria had the same weight in the final score; it is likely that there are some parameters that are more important than others in ensuring a satisfactory Doppler signal. However, a complicated scale that uses different weighting of criteria must be balanced by the ease of application in daily practice. Another limitation is that there is a possibility that during the time interval between the color flow image being frozen and freezing of the final pulsed Doppler signal, movements might have taken place that could have changed the ultimate angle of insonation.

We believe that in addition to a criterion-based scoring system advice given, to optimise the Doppler measurement, must be adhered to. This includes performing assessments during fetal quiescence; reduction of gate size to avoid sampling adjacent vessels[4].

We also realise that in this study we used cut-offs of continuous variables to dichotomise into acceptable or unacceptable. As an example, the angle of insonation is a continuous variable and quality is related to the angle

with the aim to use the smallest possible angle; yet we divide it by sing an angle above or below a certain threshold. This may have led to disagreement between subjective impression and objective measurement. However, use of such cut-offs was in keeping with the aim to derive a practical system based on accepted guidelines.

As a result, we propose that our six-criterion objective quality-scoring system is used for the assessment of the images of the fetal pulsed wave Doppler measurements. Such objective assessment is more reliable and reproducible than subjective impression and should form the basis for quality control, teaching and auditing.

Acknowledgements

We are very grateful to the pregnant women who participated in this study, and to the sonographers involved for submitting their images for evaluation. This study is part of the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) project, which is supported by a grant from The Bill & Melinda Gates Foundation to the University of Oxford, for which we are grateful.

Table 1: Image-scoring criteria for umbilical and uterine artery Doppler measurement.

Criteria	Description
Magnification	50% of the screen with zoom box and sample gate in the centre of the vessel
Angle of insonation	less than 30^{o}
Sweep speed	4 - 6 waveforms with consistent and similar signal
Clarity of the image	Pulse rate frequency and color gain correction (avoid venous signal)
Anatomic site of the sample	Umbilical artery: free loop Uterine artery: before the bifurcation above the iliac vessels
Velocity scale	75% of the peak systolic velocity

Table 2: Distribution of objective image score for each subjective image rating for pulsed wave Doppler measurement for reviewers A and B.

	Objective image score					
Subjective scoring	1	2	3	4	5	6
Unacceptable A	1 (0.8)	5 (4.2)	16 (13.3)	16 (13.3)	6 (5)	3 (2.5)
Acceptable A	-	1 (0.8)	10 (8.4)	17 (14.2)	29 (24.2)	16 (13.3)
Unacceptable B	-	1 (0.8)	15 (12.5)	6 (5)	1 (0.8)	-
Acceptable B	-	-	11 (9.2)	25 (20.8)	41 (34.2)	20 (16.7)
Unacceptable by both A and B, Objective scoring for A	1 (5.26)	5 (26.32)	9 (47.37)	4 (21.05)	-	-
Unacceptable by both A and B, Objective scoring for B	-	1 (5.26)	11 (57.89)	6 (31.58)	1 (5.26)	-
Acceptable by both A and B, Objective scoring for A	-	-	7 (10.14)	16 (23.19)	29 (42.03)	17 (24.64)
Acceptable by both A and B, Objective scoring for B	-	-	7 (10.14)	14 (20.29)	32 (46.38)	16 (23.19)

A: Operator A, B: Operator B, data given n (%)

Table 3: Adjusted kappa and percentage of agreement for individual criteria of pulsed wave Doppler assessment.

Criterion	Adjusted kappa (95% CI)	Agreement (%)
Magnification	0.77 (0.65-0.88)	88.3%
Angle of insonation	0.65 (0.52-0.78)	82.5%
Sweep speed	0.88 (0.80-0.97)	94.2%
Image clarity	0.68 (0.56-0.81)	84.2%
Anatomic site	0.97 (0.92-1.01)	98.3%
Velocity scale	0.68 (0.55-0.81)	84.2%

Figure 1: Image showing the correct way of measuring the umbilical artery Doppler.

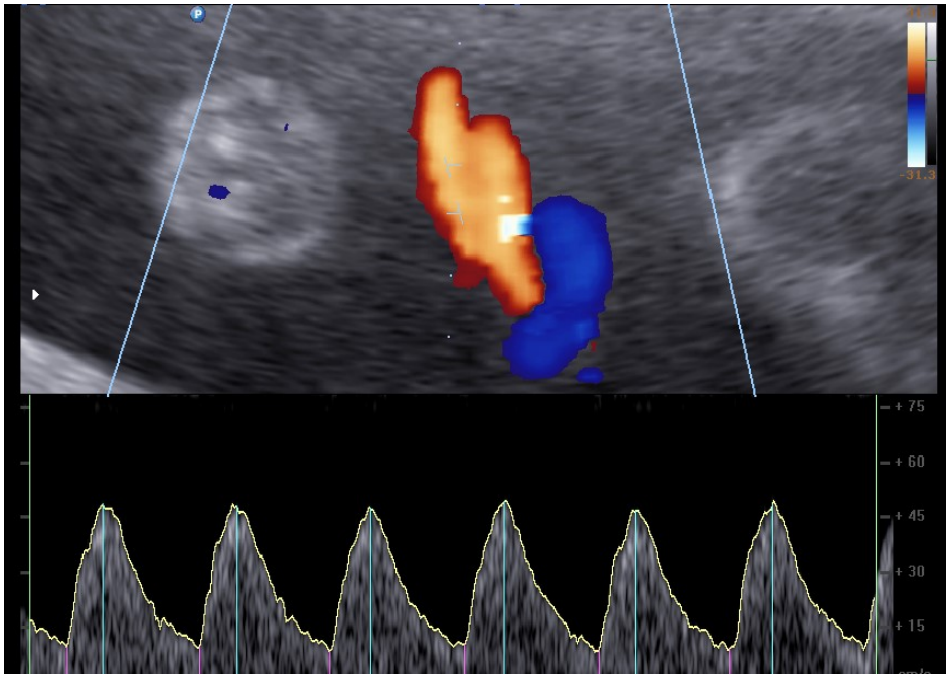


Figure 2: Image of umbilical artery Doppler demonstrating poor magnification; the velocity scale is less than 75% and sweep speed is more than 4-6 waveforms.

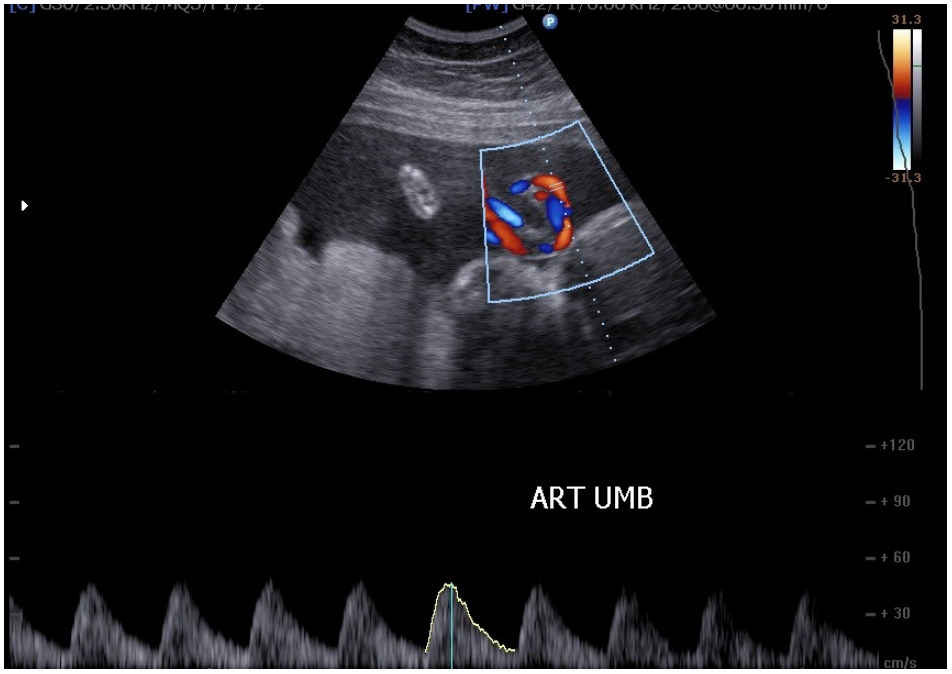


Figure 3: Image showing the measurement of uterine artery Doppler in the correct way. Note that the angle of insonation has been corrected.

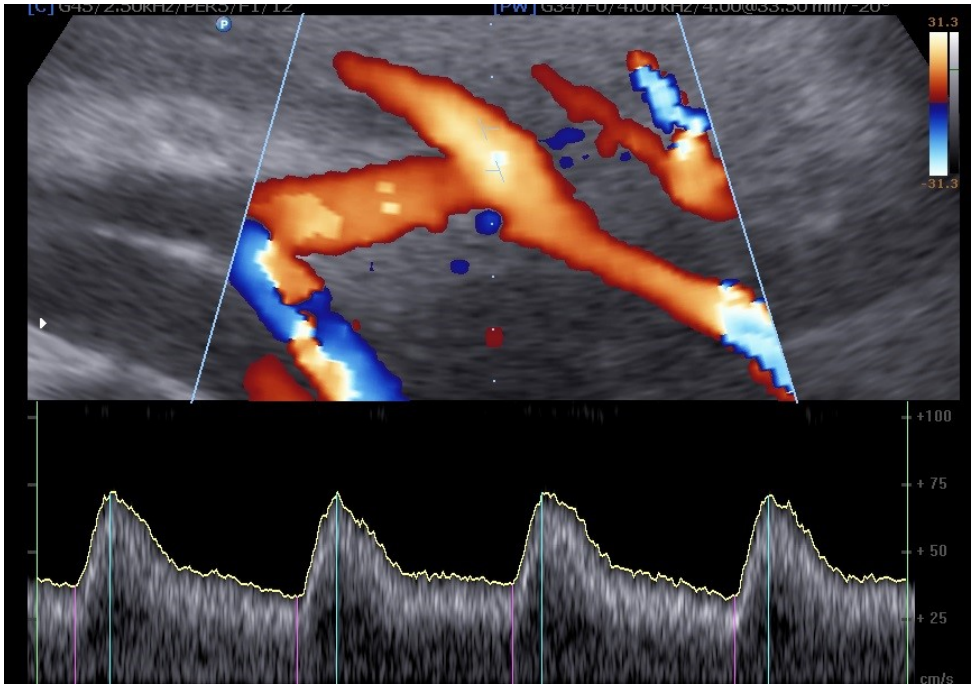


Figure 4: Image showing uterine artery Doppler measurement demonstrating poor magnification and an angle of insonation greater than 30 degrees.

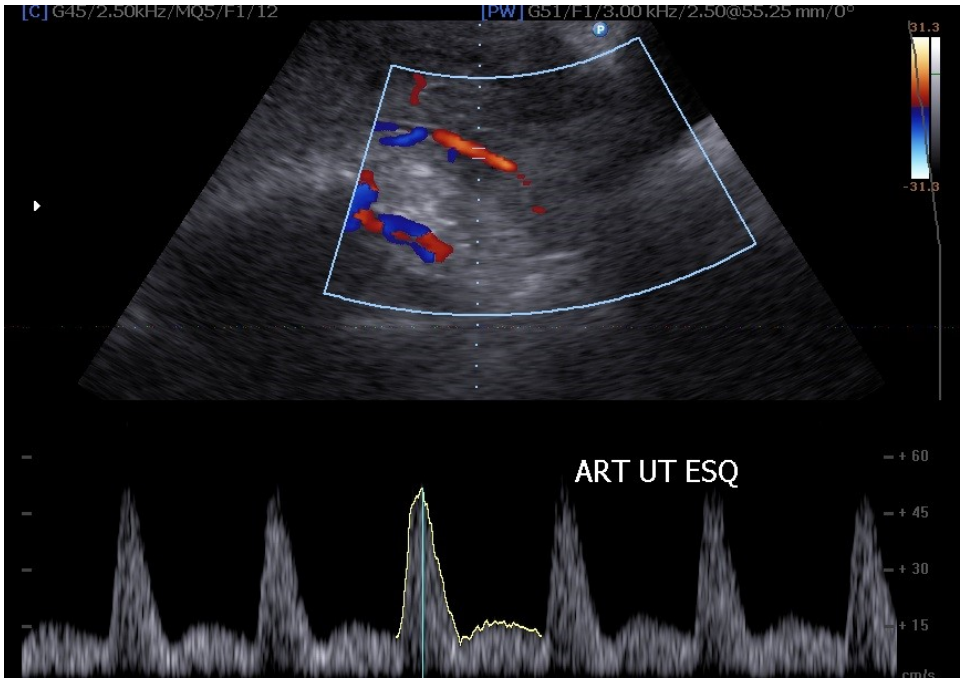
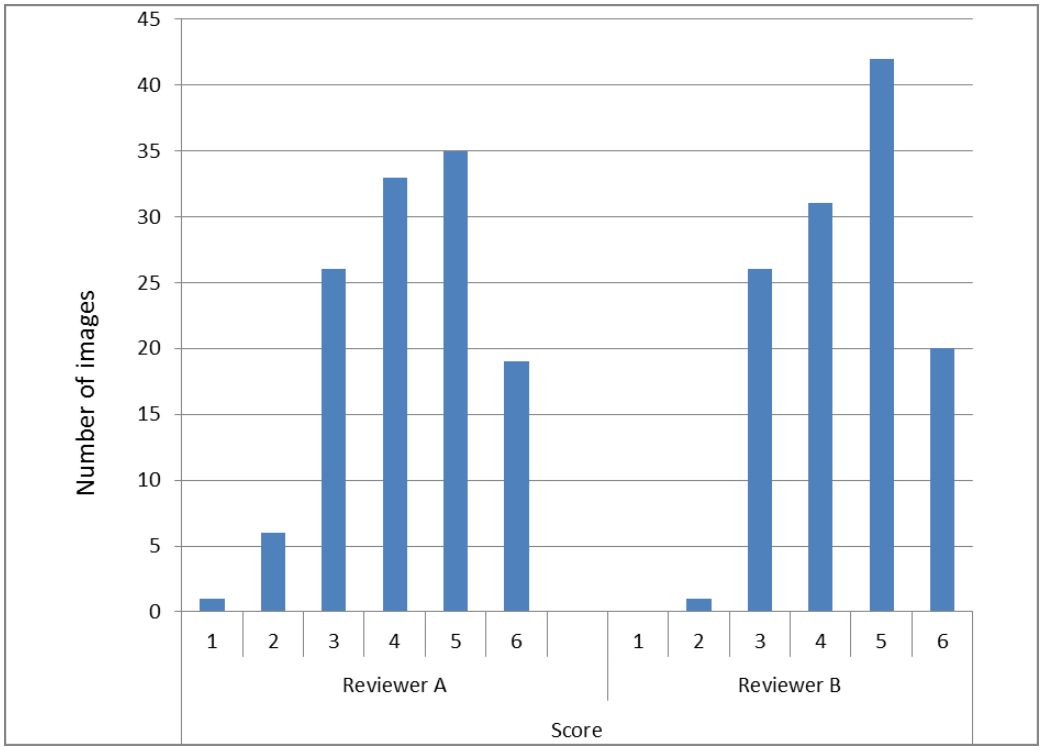


Figure 5: Frequency distribution of objective scoring for the two operators.



REFERENCES

1. Bruner JP, Gabbe SG, Levy DW, Arger PH. Doppler ultrasonography of the umbilical cord in normal pregnancy. South Med J. 1993 Jan;86(1):52-5.

2. Bruner JB, Levy DW, Arger PH. Doppler ultrasonography of the umbilical cord in complicated pregnancies. South Med J. 1993 Apr;86(4):418-22.

3. Westergaard HB, Langhoff-Roos J, Lingman G, Marsal K, Kreiner S. A critical appraisal of the use of umbilical artery Doppler ultrasound in high-risk pregnancies: use of meta-analyses in evidence-based obstetrics. Ultrasound Obstet Gynecol. 2001 Jun;17(6):466-76.

4. Bhide A, Acharya G, Bilardo CM, Brezinka C, Cafici D, Hernandez-Andrade E, et al. ISUOG practice guidelines: use of Doppler ultrasonography in obstetrics. Ultrasound Obstet Gynecol. 2013 Feb;41(2):233-39.

5. Salomon LJ, Bernard JP, Duyme M, Doris B, Mas N, Ville Y. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. Ultrasound Obstet Gynecol. 2006 Jan;27(1):34-40.

6. Snijders RJ, Thom EA, Zachary JM, Platt LD, Greene N, Jackson LG, et al. First-trimester trisomy screening: nuchal translucency measurement training and quality assurance to correct and unify technique. Ultrasound Obstet Gynecol. 2002 Apr;19(4):353-9.

7. Herman A, Maymon R, Dreazen E, Caspi E, Bukovsky I, Weinraub Z. Nuchal translucency audit: a novel image-scoring method. Ultrasound Obstet Gynecol. 1998 Dec;12(6):398-403.

8. Herman A, Maymon R, Dreazen E, Zohav E, Segal O, Segal S, et al. Utilization of the nuchal translucency image-scoring method during training of new examiners. Fetal Diagn Ther. 1999 Aug;14(4):234-9.

9. Wanyonyi SZ, Napolitano R, Ohuma EO, Salomon LJ, Papageorghiou AT. Image-scoring system for crown-rump length measurement. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2014 Dec;44(6):649-54.

10. Villar J, Altman DG, Purwar M, Noble JA, Knight HE, Ruyan P, et al. The objectives, design and implementation of the INTERGROWTH-21st Project. BJOG : an international journal of obstetrics and gynaecology. 2013 Sep;120 Suppl 2:9-26, v.

11. Interbio21.org.uk. Update to the Intergrowth-21st ultrasound operations manual. Ultrasound manual for additional measurements. http://www.interbio21.org.uk/ [31 October 2016].

12. Napolitano R, Melchiorre K, Arcangeli T, Dias T, Bhide A, Thilaganathan B. Screening for pre-eclampsia by using changes in uterine artery Doppler indices with advancing gestation. Prenat Diagn. 2012 Feb;32(2):180-4.

13. Sarris I, Ioannou C, Ohuma E, Altman D, Hoch L, Cosgrove C, et al. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21 Project. BJOG : an international journal of obstetrics and gynaecology. 2013 Jul 11.

14. Wanyonyi SZ, Napolitano R, Ohuma EO, Salomon LJ, Papageorghiou AT. Image-scoring system for crown-rump length measurement. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2014 Dec;44(6):649-54.

15. Ioannou C, Talbot K, Ohuma E, Sarris I, Villar J, Conde-Agudelo A, et al. Systematic review of methodology used in ultrasound studies aimed at creating charts of

fetal size. BJOG : an international journal of obstetrics and gynaecology. 2012;119(12):1425-39.

16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74.

17. McLennan A, Schluter PJ, Pincham V, Hyett J. First-trimester fetal nasal bone audit: evaluation of a novel method of image assessment. Ultrasound Obstet Gynecol. 2009 Dec;34(6):623-8.

18. Thia EW, Wei X, Tan DT, Lai XH, Zhang XJ, Oo SY, et al. Evaluation of an objective method of image assessment for first-trimester nasal bone. Ultrasound Obstet Gynecol. 2011 Nov;38(5):533-7.

19. Alfirevic Z, Stampalija T, Gyte GM. Fetal and umbilical Doppler ultrasound in high-risk pregnancies. Cochrane Database Syst Rev. 2013 Nov 12(11):CD007529.

20. Papageorghiou AT, Yu CK, Erasmus IE, Cuckle HS, Nicolaides KH. Assessment of risk for the development of pre-eclampsia by maternal characteristics and uterine artery Doppler. BJOG : an international journal of obstetrics and gynaecology. 2005 Jun;112(6):703-9.

21. Velauthar L, Plana MN, Kalidindi M, Zamora J, Thilaganathan B, Illanes SE, et al. First-trimester uterine artery Doppler and adverse pregnancy outcome: a meta-analysis involving 55,974 women. Ultrasound Obstet Gynecol. 2014 May;43(5):500-7.

22. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993 May;46(5):423-9.

Appendix 8: Scientific basis for standardization of fetal head measurements by ultrasound: a reproducibility study

Ultrasound Obstet Gynecol 2016; 48: 80–85
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/uog.15956

Scientific basis for standardization of fetal head measurements by ultrasound: a reproducibility study

R. NAPOLITANO*, V. DONADONO*, E. O. OHUMA*†‡, C. L. KNIGHT*, S. Z. WANYONYI*, B. KEMP*, T. NORRIS* and A. T. PAPAGEORGHIOU*†

**Nuffield Department of Obstetrics & Gynaecology, University of Oxford, Oxford, UK; †Oxford Maternal & Perinatal Health Institute, Green Templeton College, University of Oxford, Oxford, UK; ‡Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Oxford, UK*

KEYWORDS: biparietal diameter; fetal head biometry; head circumference; occipitofrontal diameter; reproducibility; transthalamic and transventricular plane; ultrasound; variability

ABSTRACT

Objective To compare the standard methods for ultrasound measurement of fetal head circumference (HC) and biparietal diameter (BPD) (outer-to-outer (BPDoo) vs outer-to-inner (BPDoi) caliper placement), and compare acquisition of these measurements in transthalamic (TT) vs transventricular (TV) planes.

Methods This study utilized ultrasound images acquired from women participating in the Oxford arm of the INTERGROWTH-21st Project. In the first phase of the study, BPDoo and BPDoi were measured on stored images. In the second phase, real-time measurements of BPD, occipitofrontal diameter (OFD) and HC in TT and TV planes were obtained by pairs of sonographers. Reproducibility of measurements made by the same (intraobserver) and by different (interobserver) sonographers, as well as the reproducibility of caliper placement and measurements obtained in different planes, was assessed using Bland–Altman plots.

Results In Phase I, we analyzed ultrasound images of 108 singleton fetuses. The mean intraobserver and interobserver differences were < 2% (1.34 mm) and the 95% limits of agreement were < 5% (3 mm) for both BPDoo and BPDoi. Neither method for measuring BPD showed consistently better reproducibility. In Phase II, we analyzed ultrasound images of 100 different singleton fetuses. The mean intraobserver and interobserver differences were < 1% (2.26 mm) and the 95% limits of agreement were < 8% (14.45 mm) for all fetal head measurements obtained in TV and TT planes. Neither plane for measuring fetal head showed consistently better reproducibility. Measurement of HC using the ellipse facility was as reproducible as HC calculated from BPD and OFD. OFD by itself was the least reproducible of all fetal head measurements.

Conclusions Measurements of BPDoi and BPDoo are equally reproducible; however, we believe BPDoo should be used in clinical practice as it allows fetal HC to be measured and compared with neonatal HC. For all head measurements, TV and TT planes provide equally reproducible values at any gestational age, and HC values are similar in both planes. Fetal head measurement in the TT plane is preferable as international standards in this plane are available; however, measurements in the TV plane can be plotted on the same standards. Copyright © 2016 ISUOG. Published by John Wiley & Sons Ltd.

INTRODUCTION

Fetal head biometry is important for estimation of gestational age in the second trimester and for monitoring fetal growth. Unfortunately, even after decades of clinical practice, guidelines still vary as to how the measurements should be taken, i.e. whether the biparietal diameter (BPD) should be measured by outer-to-outer (BPDoo) or outer-to-inner (BPDoi) caliper placement[1,2]. It is also uncertain whether head circumference (HC) should be calculated from the occipitofrontal diameter (OFD) and BPD ($HC_{calculated}$) or by using the ellipse facility ($HC_{ellipse}$) on the ultrasound machine, and which is the better plane to use, i.e. transthalamic (TT) or transventricular (TV)[1,3]. These issues are important clinically because measurement inconsistencies may affect the management of individual pregnancies, make it difficult to compare data across units and contribute to the heterogeneity of studies describing fetal size[4,5].

Correspondence to: Dr A. T. Papageorghiou, Nuffield Department of Obstetrics & Gynaecology, John Radcliffe Hospital, Oxford, OX3 9DU, UK (e-mail: aris.papageorghiou@obs-gyn.ox.ac.uk)

Accepted: 29 April 2016

ORIGINAL PAPER

In this study, we aimed to compare (i) the standard methods for measuring fetal HC ($HC_{ellipse}$ *vs* $HC_{calculated}$) and BPD (BPDoo *vs* BPDoi caliper placement) on ultrasound and (ii) the effect of acquiring head measurements in TT *vs* TV planes, so as to make recommendations regarding best practice.

SUBJECTS AND METHODS

This study involved women at low risk of adverse pregnancy outcome who were recruited into the Oxford arm of the INTERGROWTH-21st Project (www.intergrowth21.org.uk), a multicenter, multiethnic, population-based project, conducted between 2008 and 2014 across eight countries[6]. The Fetal Growth Longitudinal Study (FGLS) is one of the three main components of the INTERGROWTH-21st Project, which aimed to construct international standards for fetal growth. All women included in our study were part of the FGLS. In the FGLS, serial two-dimensional ultrasound scans were performed every 5 ± 1 weeks, from $14 + 0$ to $41 + 6$ weeks' gestation, and images were stored for later analysis. Inclusion criteria for the FGLS were pregnant women with a known, certain last menstrual period, who had regular menstrual cycles and were not taking hormonal contraceptives or breastfeeding in the 2 months before they conceived naturally. Gestational age was calculated using the last menstrual period, with ultrasound confirmation based on a crown–rump length measurement at $9 + 0$ to $13 + 6$ weeks' gestation that was in agreement by ≤ 7 days[7,8].

All ultrasound scans in the FGLS were performed by sonographers who were trained, standardized and regularly audited[2,8,9]. At each examination, BPDoo, OFD and $HC_{ellipse}$ were acquired in triplicate in the TT plane. The same commercially available ultrasound machine (Philips HD-9, Philips Ultrasound, Bothell, WA, USA) with curvilinear abdominal transducers (C5-2, C6-3 and V7-3) was used at all study sites. For the purposes of the INTERGROWTH-21st Project, the manufacturer reprogrammed the machine's software to ensure that measurement values did not appear on the screen, so as to reduce operator 'expected value' bias[2]. The INTERGROWTH-21st Project was approved by the Oxfordshire Research Ethics Committee 'C' (reference: 08/H0606/139) and all participants gave written informed consent.

Phase I: evaluation of biparietal diameter caliper placement

Using the stored ultrasound images acquired in the FGLS, two sonographers twice measured the BPD using two methods (BPDoo (Figure 1a) and BPDoi (Figure 1b)) on the first of the three images, after the original caliper placements had been removed from the image. The sonographers were blinded to their own and each other's measurements. The intraobserver reproducibility for both methods was calculated for the two sonographers. To calculate the interobserver reproducibility, the first measurements of Sonographer A were compared with those of Sonographer B, and then repeated for the second measurements.

Phase II: evaluation of transthalamic and transventricular planes

From a cohort of participants that was different from that in Phase I, two sonographers obtained real-time measurements of BPDoo, OFD and $HC_{ellipse}$ in the TV (Figure 1a) and TT (Figure 1b) planes in duplicate, providing an additional set of images to those in the FGLS. As no difference was found between BPDoo and BPDoi in Phase I, only BPDoo was measured to reduce scanning time. All measurements were obtained in a blinded fashion and were stored on the ultrasound machine and retrieved after completion of the study.

Each sonographer placed the calipers once on each of the four images acquired per participant (i.e. a total of 12 measurements per sonographer for BPDoo, OFD and $HC_{ellipse}$). Sonographer B repeated the caliper placements on the images acquired by Sonographer A, resulting in a total of 36 measurements. HC was also calculated from BPD and OFD ($HC_{calculated}$) for each image.

Measurement and plane definitions

BPDoo was measured with the intersection of the calipers placed from the outer edge of the proximal calvarial wall to the outer edge of the distal calvarial wall, at the widest part of the skull (Figure 1a). BPDoi was measured with the intersection of the calipers placed from the outer edge of the proximal calvarial wall to the inner edge of the distal calvarial wall (Figure 1b)[10]. OFD was measured with the intersection of the calipers placed from the outer edge of the anterior frontal wall to the outer edge of the distal occipital wall, at the longest part of the skull (Figure 1b). $HC_{ellipse}$ was measured using the ellipse facility, placing the line of the ellipse on the outer border of the skull (Figure 1b)[2]. The TT plane was acquired according to the following conditions: axial view at the level of the thalami with an angle of insonation as close as possible to 90°; the head had to be oval in shape, symmetrical, centrally positioned and filling at least 30% of the monitor; the midline echo (representing the falx cerebri) had to be broken anteriorly, at a third of its length, by the cavum septi pellucidi; and the thalami had to be located symmetrically on either side of the midline (Figure 1b)[2]. The TV plane was acquired including all the standard parameters to obtain a TT plane but visualizing the lateral ventricles rather than the thalami at a more cranial level, with the ventricles located symmetrically on each side of the midline, the anterior and posterior horns both visible, and the posterior ventricle cavity visualized as a hypoechoic structure (Figure 1a)[1].

Statistical analysis

In Phase I, the following analyses were performed: (i) intraobserver reproducibility of caliper placement for

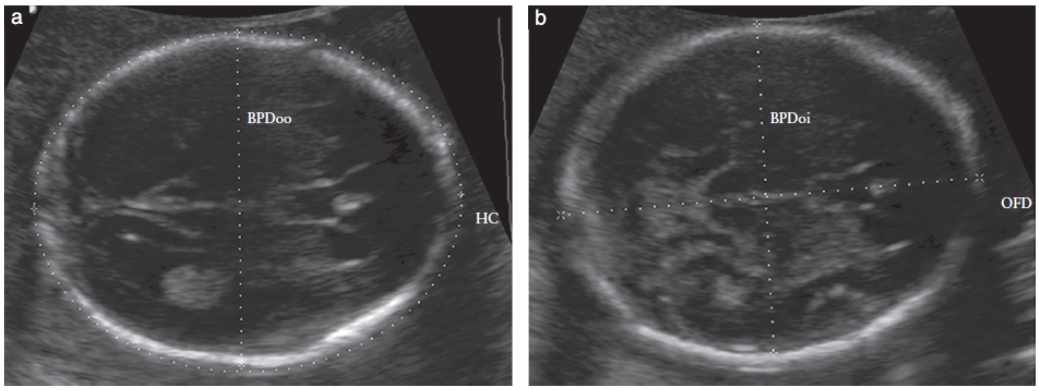


Figure 1 (a) Ultrasound image of biparietal diameter, measured using outer-to-outer caliper placement (BPDoo), and fetal head circumference (HC), measured using the ellipse facility, in the transventricular plane. (b) Ultrasound image of biparietal diameter, measured using outer-to-inner caliper placement (BPDoi), and occipitofrontal diameter (OFD) in the transthalamic plane.

measurement of BPD using the BPDoo and BPDoi method, calculated for Sonographers A and B; and (ii) interobserver reproducibility of caliper placement for measurements of BPD using the BPDoo and BPDoi method, comparing the first measurements of Sonographer A with those of Sonographer B, and the second measurements of Sonographer A with those first obtained by Sonographer B.

In Phase II, the following analyses were performed: (i) intraobserver reproducibility of plane acquisition and caliper placement for TT and TV planes, comparing each sonographer's first and second measurements in the same plane; (ii) interobserver reproducibility of plane acquisition and caliper placement for TT and TV planes, comparing measurements of Sonographers A and B in the same plane; (iii) caliper replacement reproducibility, based on Sonographer B replacing the calipers on the images acquired by Sonographer A in the TT and TV planes (interobserver reproducibility); (iv) intraobserver reproducibility of plane acquisition and caliper placement between TT and TV planes, comparing the measurements of Sonographer A acquired in the TT plane with those acquired by Sonographer A in the TV plane (the same was then calculated for Sonographer B); and (v) interobserver reproducibility for plane acquisition and caliper placement between TT and TV planes, comparing the measurements of Sonographer A acquired in the TT plane with those acquired by Sonographer B in the TV plane, and then the measurements of Sonographer B acquired in the TT plane with those acquired by Sonographer A in the TV plane.

Intraobserver and interobserver variability were expressed as a percentage to account for increasing fetal head size with gestational age. Percentages were calculated as the difference between two measurements divided by the average of the two measurements, multiplied by 100. Reproducibility was assessed using Bland–Altman plots. All plots and analyses were performed using STATA 11 (StataCorp, College Station, TX, USA).

Paired or unpaired *t*-tests, as appropriate, were performed to assess mean differences between measurements obtained by the same sonographer (intraobserver reproducibility) and different sonographers (interobserver reproducibility), and those obtained in two different planes (between-plane reproducibility). A *P*-value of < 0.05 was considered statistically significant.

RESULTS

Four women were included in the study at each gestational week, from 15 to 41 weeks in Phase I (108 women) and from 16 to 40 weeks in Phase II (100 women), resulting in a total of 4464 measurements. The demographic characteristics of the 208 participants are shown in Table 1.

Phase I: evaluation of biparietal diameter caliper placement

A total of 864 measurements were obtained in Phase I. Intraobserver and interobserver reproducibility was very good overall. The mean differences were $< 2\%$ (1.34 mm) and the 95% limits of agreement were $< 5\%$ (3 mm) for both BPDoo and BPDoi (Table 2 and Figures S1 and S2); however, neither method showed consistently better reproducibility. As expected, the 95% limits of agreement for interobserver reproducibility of BPDoo and BPDoi (3.1–4.2%) were slightly wider than for the intraobserver reproducibility (1.3–2.1%).

Phase II: evaluation of transthalamic *vs* transventricular plane

A total of 3600 measurements (1200 for BPD, OFD and $HC_{ellipse}$) were obtained in Phase II. $HC_{ellipse}$ was

Table 1 Demographic characteristics of women with singleton pregnancy recruited into the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project who had retrospective measurement of biparietal diameter (BPD) (Phase I) or real-time measurements of fetal biometry in transthalamic (TT) and transventricular (TV) planes (Phase II)

Characteristic	*Phase I: BPD study (n = 108)*	*Phase II: TT/TV study (n = 100)*
Maternal age (years)	30 ± 4	30 ± 5
BMI (kg/m^2)	23.3 ± 2.7	26.9 ± 3.9
Nulliparous	66 (61)	42 (42)
GA at scan (weeks)	28.1 ± 7.7	28.0 ± 7.2

Data are given as mean ± SD or *n* (%). BMI, body mass index; GA, gestational age.

Table 2 Intra- and interobserver reproducibility of biparietal diameter measurement using outer-to-outer (BPDoo) or outer-to-inner (BPDoi) caliper placement method

	Intraobserver		
Measurement	*Sonographer A*	*Sonographer B*	*Interobserver*
BPDoo	0.01 (2.08)	0.02 (1.28)	1.93 (4.16)
BPDoi	−0.16 (1.63)	−0.15 (1.33)	0.80 (3.10)

Data are given as mean difference (95% limits of agreement (LOA)) in percent. Upper and lower 95% LOA in each case can be calculated as mean difference ± value displayed.

marginally larger, by 0.09% (0.61 mm, $P = 0.034$), when measured in the TV than when measured in the TT plane. However, no such difference was observed for BPD or OFD. In terms of overall reproducibility, the mean differences in fetal head measurements were < 1% (2.26 mm) and the 95% limits of agreement were < 8% (14.45 mm) for both TV and TT planes (Figures S3–S7).

Overall, the reproducibility of caliper placement accounted for 50–60% of the reproducibility of measurements obtained in each plane. For example, the 95% limits of agreement for interobserver reproducibility of $HC_{ellipse}$ in the TV plane was 4.87% (Table 3 and Figure S4) and the respective value for reproducibility of caliper replacements in the same plane was 3.05% (Table 3 and Figure S5), constituting approximately 60% of the total reproducibility.

Neither the TV or TT plane was associated with consistently better reproducibility. In addition, the 95% limits of agreement between sonographers measuring in the same plane (interobserver reproducibility within the same plane) were only slightly wider than the limits of agreement between TV and TT planes acquired and measured by the same sonographer (intraobserver reproducibility between TT and TV planes). This suggests that the effect of two sonographers measuring in the same plane is similar to that of the same sonographer measuring in different planes. The 95% limits of agreement were highest when two sonographers measured in different planes (interobserver reproducibility between TT and TV planes) (Table 3 and Figure S7). Lastly, there was no significant difference between $HC_{ellipse}$ measurements and an equal number of $HC_{calculated}$ measurements.

DISCUSSION

Main findings

The aim of this study was to determine the most reproducible method for performing fetal head biometry for clinical practice and research, such as the production of standards. There are two approaches that could have been used. The first is to assess the accuracy of the ultrasound measurements against a 'gold standard'[11]. However, defining a gold standard for fetal measurements is difficult. For example, magnetic resonance imaging allows clear visualization of the fetus, but estimates are still associated with errors[12]. The use of phantoms has obvious limitations as inanimate structures do not effectively represent the variability of live structures[13]. The second approach is to assess the reproducibility of different methods of measuring fetal head biometry and to use the one with least error and bias[14].

We found no major differences in the reproducibility of caliper placement for measuring BPDoo or BPDoi. Similarly, there was no difference in the reproducibility of measuring HC in the TV or TT planes. Using the ellipse facility ($HC_{ellipse}$) to measure HC was marginally more reproducible than using the two-diameters method ($HC_{calculated}$), with the former having interobserver 95% limits of agreement of just below 5% and the latter having interobserver 95% limits of agreement of just above 5%. This is probably due to the contribution of the OFD, which is the least reproducible head measurement in the two-diameters method.

The BPDoi method was used originally because the inner margin of the fetal skull in the distal field was sharper when using static B scanners[15–18]. However, modern equipment produces a clearer image and so the BPDoi method appears to have no measurable effect on reproducibility (Table 2), even though caliper replacement constitutes up to 60% of the total variability. Therefore, choosing between BPDoo and BPDoi should be for reasons other than trying to reduce error, such as the protocol used (BPDoo) to develop international standards for monitoring fetal growth[19]. Another reason for using BPDoo is that it enables direct comparisons to be made between antenatal and postnatal measurements of HC[20,21].

Lastly, neither the TV nor TT plane was found to be consistently associated with better reproducibility. We did find that biometry in the TV plane yielded a very slightly larger HC than that measured in the TT plane. Although this was statistically significant, it was not clinically relevant (< 0.1%, 0.61 mm). Furthermore, when comparing the reproducibility of measuring HC in the TT and TV planes, the difference between sonographers measuring in the same plane was similar to that of the same sonographer measuring in different planes.

Table 3 Intraobserver and interobserver reproducibility of ultrasound measurements of fetal head biometry and caliper replacement in the same plane and between planes

	Within-plane reproducibility						*Between-plane reproducibility*	
	Intraobserver		*Interobserver*		*Caliper replacement*		*TT vs TV*	
Measurement	*TT*	*TV*	*TT*	*TV*	*TT interobserver*	*TV interobserver*	*Intraobserver*	*Interobserver*
BPDoo	−0.14 (4.05)	−0.02 (3.43)	0.70 (6.65)	0.09 (4.78)	0.30 (3.16)	0.41 (2.69)	0.24 (5.63)	0.24 (5.84)
OFD	−0.31 (6.55)	−0.41 (5.50)	−0.03 (7.98)	−0.13 (7.66)	0.50 (4.63)	0.86 (4.58)	−0.13 (6.69)	−0.14 (8.11)
$HC_{ellipse}$	−0.06 (3.47)	−0.25 (3.32)	−0.48 (4.78)	−0.75 (4.87)	−0.43 (3.14)	0.12 (3.05)	−0.09 (4.53)	−0.10 (5.11)
$HC_{calculated}$	−0.23 (4.13)	−0.24 (3.53)	0.29 (5.54)	0.02 (5.02)	0.43 (2.91)	0.66 (2.92)	0.04 (4.78)	0.03 (5.50)

Data are given as mean difference (95% limits of agreement (LOA)) in percent. Upper and lower 95% LOA in each case can be calculated as mean difference ± value displayed. BPDoo, biparietal diameter measured using outer-to-outer caliper placement; $HC_{calculated}$, head circumference calculated from biparietal diameter and occiptofrontal diameter (OFD); $HC_{ellipse}$, head circumference measured using ellipse facility on ultrasound machine; TT, transthalamic; TV, transventricular.

Limitations and strengths

There are some limitations to our study. It can be argued that the use of six different sonographers working in pairs (rather than one pair) might have had an impact on the results. However, we feel that the study design more accurately reflects clinical practice, as most units have several qualified sonographers[22]. The setting of near-optimal conditions (i.e. experienced sonographers, healthy population and a scientifically rigorous study design) may be seen as creating an artificial setting. However, such conditions were necessary to minimize the contribution of confounding factors so as to define the variability in relation to the research question as purely as possible, which we see as a strength. The other strengths of our study were that reproducibility was assessed throughout pregnancy by recruiting a fixed number of women per week of gestation, and recommended methods[23] were used that have been shown to be the most appropriate for assessing the reproducibility of two measurements[24,25].

Our findings in context with other studies

A literature search was performed to identify all publications reporting reproducibility in the evaluation of fetal head biometry. We searched MEDLINE using the following keywords: biparietal diameter OR BPD OR occipitofrontal diameter OR OFD OR head circumference OR HC AND fetal OR foetal OR fetus OR foetus AND ultrasound OR ultrasonogra* OR ultra-sonogra* OR sonic* OR scan* AND reproducibility OR variability OR repeatability. Restrictions that were applied were studies in humans, in the English language and published after 1970. Additional references were added from an important article[4]. Nineteen relevant studies were identified (Table S1)[15–18,22,26–39]. In most, the primary aim of the study was not to assess reproducibility but to build growth charts. The studies reporting either BPD method did not reveal large differences from our findings (the reported mean differences were < 2% for BPDoi, with limits of agreement of < 5%[15–18,34,36], and there were only two small studies[29,38] on BPDoo showing limits of agreement of 3.8 and 7.4 mm, respectively). In only one study was the reproducibility of BPDoo and BPDoi reported in the same group of fetuses, which showed repeatability coefficients that were similar for both methods[34]. Measurements of $HC_{ellipse}$ were reproducible, with a mean difference of 3.5 mm and limits of agreement of < 12 mm (5%), in line with our results[15–17,22,27–29,34,35,39]. No previous study was found comparing the two different planes of acquisition (TV *vs* TT) in the same population.

In conclusion, using modern ultrasound equipment, measurement of BPD is equally reproducible irrespective of whether calipers are placed BPDoo or BPDoi. However, BPDoo can be used for both BPD and HC measurements and is also the method to measure OFD. It therefore seems simplest to use BPDoo as a conceptually similar methodological approach for all head measurements. BPDoo is also clinically useful (as part of the $HC_{calculated}$) for monitoring growth from the 'womb to the classroom'[40], as it is possible to track head size and growth from the antenatal to postnatal periods[41]. We found that HC measurements using $HC_{ellipse}$ were associated with slightly better interobserver reproducibility than using $HC_{calculated}$, based on BPD and OFD. However, there was no large difference in reproducibility of BPD, OFD or $HC_{ellipse}$ measured in the TV compared with TT plane. The mean difference in head size between these two planes was also minimal (< 1%) at every gestational age.

We therefore recommend that standard fetal head biometry measurements are performed using the BPDoo, OFD and $HC_{ellipse}$, all measured in the TT plane, based on the reproducibility evidence presented in this study and the existence of international standards based on these methods. In centers in which HC is measured in the TV plane, use of the international standards is still appropriate[19].

ACKNOWLEDGMENTS

A.T.P. is the Chief Medical Officer of Intelligent Ultrasound and receives non-financial support from Philips Ultrasound. This project was supported by a generous grant from the Bill & Melinda Gates Foundation to the University of Oxford (Oxford, UK), for which we are very grateful.

REFERENCES

1. Sonographic examination of the fetal central nervous system: guidelines for performing the 'basic examination' and the 'fetal neurosonogram'. *Ultrasound Obstet Gynecol* 2007; **29:** 109–116.
2. Papageorghiou AT, Sarris I, Ioannou C, Todros T, Carvalho M, Pilu G, Salomon LJ. Ultrasound methodology used to construct the fetal growth standards in the INTERGROWTH-21st Project. *BJOG* 2013; **120** Suppl 2: 27–32.
3. Salomon LJ, Alfirevic Z, Berghella V, Bilardo C, Hernandez-Andrade E, Johnsen SL, Kalache K, Leung KY, Malinger G, Munoz H, Prefumo F, Toi A, Lee W. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound Obstet Gynecol* 2011; **37:** 116–126.
4. Ioannou C, Talbot K, Ohuma E, Sarris I, Villar J, Conde-Agudelo A, Papageorghiou A. Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size. *BJOG* 2012; **119:** 1425–1439.
5. Napolitano R, Dhami J, Ohuma EO, Ioannou C, Conde-Agudelo A, Kennedy SH, Villar J, Papageorghiou AT. Pregnancy dating by fetal crown-rump length: a systematic review of charts. *BJOG* 2014; **121:** 556–565.
6. Villar J, Altman DG, Purwar M, Noble JA, Knight HE, Ruyan P, Cheikh Ismail L, Barros FC, Lambert A, Papageorghiou AT, Carvalho M, Jaffer YA, Bertino E, Gravett MG, Bhutta ZA, Kennedy SH. The objectives, design and implementation of the INTERGROWTH-21st Project. *BJOG* 2013; **120** Suppl 2: 9–26.
7. Wanyonyi S, Napolitano R, Ohuma E, Salomon L, Papageorghiou A. Image-scoring system for crown-rump length measurements. *Ultrasound Obstet Gynecol* 2014; **44:** 649–654.
8. Ioannou C, Sarris I, Hoch L, Salomon LJ, Papageorghiou AT. Standardisation of crown-rump length measurement. *BJOG* 2013; **120** Suppl 2: 38–41.
9. Sarris I, Ioannou C, Ohuma E, Altman D, Hoch L, Cosgrove C, Fathima S, Salomon L, Papageorghiou A. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21 Project. *BJOG* 2013; **120** Suppl 2: 33–37.
10. Leung TN, Pang MW, Daljit SS, Leung TY, Poon CF, Wong SM, Lau TK. Fetal biometry in ethnic Chinese: biparietal diameter, head circumference, abdominal circumference and femur length. *Ultrasound Obstet Gynecol* 2008; **31:** 321–327.
11. Ioannou C, Javaid MK, Mahon P, Yaqub MK, Harvey NC, Godfrey KM, Noble JA, Cooper C, Papageorghiou AT. The effect of maternal vitamin D concentration on fetal bone. *J Clin Endocrinol Metab* 2012; **97:** E2070–E2077.
12. Lo Zito L, Kadji C, Cannie M, Kacem Y, Strizek B, Mbonyumutwa M, Wuyts F, Jani J. Determination of fetal body volume measurement at term with magnetic resonance imaging: effect of various factors. *J Matern Fetal Neonatal Med* 2013; **26:** 1254–1258.
13. Ioannou C, Sarris I, Yaqub MK, Noble JA, Javaid MK, Papageorghiou AT. Surface area measurement using rendered three-dimensional ultrasound imaging: an in-vitro phantom study. *Ultrasound Obstet Gynecol* 2011; **38:** 445–449.
14. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Clin Epidemiol* 2009; **62:** 797–806.
15. Al-Meshari AA, Raber H. Fetal biparietal diamater in Saudi Arabia. *Ann Saudi Med* 1987; **7:** 227–233.
16. Chan LW, Fung TY, Leung TY, Sahota DS, Lau TK. Volumetric (3D) imaging reduces inter- and intraobserver variation of fetal biometry measurements. *Ultrasound Obstet Gynecol* 2009; **33:** 447–452.
17. Di Battista E, Bertino E, Benso L, Fabris C, Aicardi G, Pagliano M, Bossi A, De Biasio P, Milani S. Longitudinal distance standards of fetal growth. Intrauterine and Infant Longitudinal Growth Study: IILGS. *Acta Obstet Gynecol Scand* 2000; **79:** 165–173.
18. Gull I, Fait G, Har-Toov J, Kupferminc MJ, Lessing JB, Jaffa AJ, Wolman I. Prediction of fetal weight by ultrasound: the contribution of additional examiners. *Ultrasound Obstet Gynecol* 2002; **20:** 57–60.
19. Papageorghiou AT, Ohuma EO, Altman DG, Todros T, Cheikh Ismail L, Lambert A, Jaffer YA, Bertino E, Gravett MG, Purwar M, Noble JA, Pang R, Victora CG, Barros FC, Carvalho M, Salomon LJ, Bhutta ZA, Kennedy SH, J. Villar J; for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *Lancet* 2014; **384:** 869–879.
20. Cheikh Ismail L, Knight HE, Ohuma EO, Hoch L, Chumlea WC. Anthropometric standardisation and quality control protocols for the construction of new, international, fetal and newborn growth standards: the INTERGROWTH-21st Project. *BJOG* 2013; **120** Suppl 2: 48–55.
21. Villar J, Cheikh Ismail L, Victora CG, Ohuma EO, Bertino E, Altman DG, Lambert A, Papageorghiou AT, Carvalho M, Jaffer YA, Gravett MG, Purwar M, Frederick IO, Noble AJ, Pang R, Barros FC, Chumlea C, Bhutta ZA, Kennedy SH; for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. *Lancet* 2014; **384:** 857–868.
22. Sarris I, Ioannou C, Chamberlain P, Ohuma E, Roseman F, Hoch L, Altman DG, Papageorghiou AT. Intra- and interobserver variability in fetal ultrasound measurements. *Ultrasound Obstet Gynecol* 2012; **39:** 266–273.
23. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1:** 307–310.
24. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304:** 1491–1494.
25. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med* 1990; **20:** 337–340.
26. Bergsjo P, Bakke T, Bjerkedal T. Growth of the fetal skull, with special reference to weight-for-dates of the newborn child. *Acta Obstet Gynecol Scand* 1976; **55:** 53–57.
27. Deter RL, Harrist RB, Hadlock FP, Carpenter RJ. Fetal head and abdominal circumferences: I. Evaluation of measurement errors. *J Clin Ultrasound* 1982; **10:** 357–363.
28. Hadlock FP, Deter RL, Harrist RB, Park SK. Fetal head circumference: relation to menstrual age. *AJR Am J Roentgenol* 1982; **138:** 649–653.
29. Johnsen SL, Wilsgaard T, Rasmussen S, Sollien R, Kiserud T. Longitudinal reference charts for growth of the fetal head, abdomen and femur. *Eur J Obstet Gynecol Reprod Biol* 2006; **127:** 172–185.
30. Krampl E, Lees C, Bland JM, Espinoza Dorado J, Moscoso G, Campbell S. Fetal biometry at 4300 m compared to sea level in Peru. *Ultrasound Obstet Gynecol* 2000; **16:** 9–18.
31. Larsen T, Petersen S, Greisen G, Larsen JF. Normal fetal growth evaluated by longitudinal ultrasound examinations. *Early Hum Dev* 1990; **24:** 37–45.
32. Lima JC, Miyague AH, Filho FM, Nastri CO, Martins WP. Biometry and fetal weight estimation by two-dimensional and three-dimensional ultrasonography: an intraobserver and interobserver reliability and agreement study. *Ultrasound Obstet Gynecol* 2012; **40:** 186–193.
33. Merialdi M, Caulfield LE, Zavaleta N, Figueroa A, Costigan KA, Dominici F, Dipietro JA. Fetal growth in Peru: comparisons with international fetal size charts and implications for fetal growth assessment. *Ultrasound Obstet Gynecol* 2005; **26:** 123–128.
34. Pang MW, Leung TN, Sahota DS, Lau TK, Chang AM. Customizing fetal biometric charts. *Ultrasound Obstet Gynecol* 2003; **22:** 271–276.
35. Perni SC, Chervenak FA, Kalish RB, Magherini-Rothe S, Predanic M, Streltzoff J, Skupski DW. Intraobserver and interobserver reproducibility of fetal biometry. *Ultrasound Obstet Gynecol* 2004; **24:** 654–658.
36. Persson PH, Grennert L, Gennser G, Gullberg B. Normal range curves for the intrauterine growth of the biparietal diameter. *Acta Obstet Gynecol Scand Suppl* 1978; **78:** 15–20.
37. Salpou D, Kiserud T, Rasmussen S, Johnsen SL. Fetal age assessment based on 2nd trimester ultrasound in Africa and the effect of ethnicity. *BMC Pregnancy Childbirth* 2008; **8:** 48.
38. Shepard M, Filly RA. A standardized plane for biparietal diameter measurement. *J Ultrasound Med* 1982; **1:** 145–150.
39. Yang F, Leung KY, Lee YP, Chan HY, Tang MH. Fetal biometry by an inexperienced operator using two- and three-dimensional ultrasound. *Ultrasound Obstet Gynecol* 2010; **35:** 566–571.
40. Villar J, Papageorghiou AT, Pang R, Salomon LJ, Langer A, Victora C, Purwar M, Chumlea C, Qingqing W, Scherjon SA, Barros FC, Carvalho M, Altman DG, Giuliani F, Bertino E, Jaffer YA, Cheikh Ismail L, Ohuma EO, Lambert A, Noble JA, Gravett MG, Bhutta ZA, Kennedy SH. Monitoring human growth and development: a continuum from the womb to the classroom. *Am J Obstet Gynecol* 2015; **213:** 494–499.
41. Villar J, Papageorghiou AT, Pang R, Ohuma EO, Cheikh Ismail L, Barros FC, Lambert A, Carvalho M, Jaffer YA, Bertino E, Gravett MG, Altman DG, Purwar M, Frederick IO, Noble AJ, Victora CG, Bhutta ZA, Kennedy SH; for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). The likeness of fetal growth and newborn size across non-isolated populations in the INTERGROWTH-21(st) Project: the Fetal Growth Longitudinal Study and Newborn Cross-Sectional Study. *Lancet Diabetes Endocrinol* 2014; **2:** 781–792.

SUPPORTING INFORMATION ON THE INTERNET

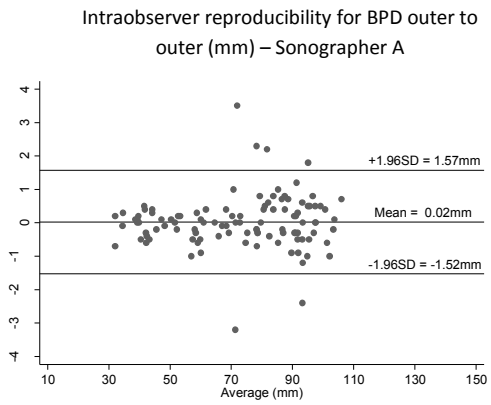
Table S1 and Figures S1–S7 may be found in the online version of this article.

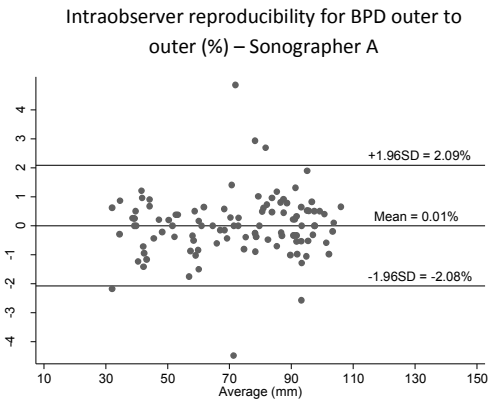
Table S1 Studies reporting on quantitative reproducibility of fetal head biometry identified in literature search

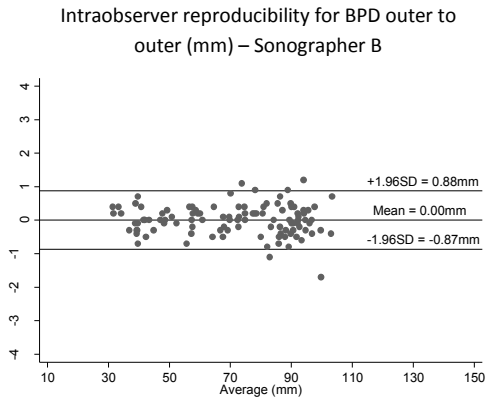
Reference	*N*	GA (weeks)	Biometry measured	Plane	BA plots reported	BPD mean difference mm (%)		BPD 95% CI mm (%))		HC mean difference mm (%)		HC 95% CI mm (%)	
						Intra.	Intra.	Inter.	Inter.	Intra.	Inter.	Intra.	Inter.
Al-Meshari 1987	100	13–36	BPDoi, HC, OFD	TT	No	1.8	1.0	1.0				1.0	1.0
Bergsjo 1976	71	26–44	BPD	NA	No			3.04					
Chan 2009	36	22–30	BPDoi, HC	TT	Yes		1.3	1.7				8.8	8.1
Deter 1982	110	18–34	HC	NA	No					−0.69		2.02	
Di Battista 2000*	20	13–36	BPDoi, HC, OFD	TT	No				1T: 0.67; 3T: 0.76		2T: 2.43; 3T: 3.49		
Gull 2002	39	38–41	BPDoi, HC	TT	No	1.42 (1.88)	2.4 (3.5)					3.4 (4.3)	
Hadlock 1982	26	15–41	HC	NA	No					CP: 0.3 Ellipse: 0.3		CP: 7.8 Ellipse: 11.2	
Johnsen 2006	20	12–31	BPDoo, HC	TT	No		3.8	3.8				11.8	11.8
Krampl 2000	62	14–42	BPDoo, HC, OFD	TV	Yes								
Larsen 1990	5	27–38	BPDoi, HC	TV	No	0.95							
Lima 2012	102	24–40	BPDoi	TT	Yes		2.9 (4.01)	3.4 (4.6)					
Merialdi 2005	NA	24–38	BPDoo, HC	NA	No								
Pang 2003	NA	24–40	BPDoi, BPDoo, HC	TT	No			1.4†; 1.95‡					1.97
Perni 2004	122	15–40	BPD, HC	TT	Yes	0.2	2.5	2.8	0.2	0.6	0.1	9.3	10.9
Persson 1978	30	16–40	BPDoi	NA	No			1.8	Same day: 0.05 Different day: 0.44				
Salpou 2008	200	12–22	BPDoi, BPDoo, HC	TT	No								
Sarris 2012	140	14–41	BPDoo, HC, OFD	TT	Yes							CP: 4.5 (2.4) Ellipse: 7.0 (3.0) Calculated: 7.2 (3.1)	CP: 9.8 (3.7) Ellipse: 12.1 (4.9) Calculated: 12.0 (4.9)
Shepard 1982	18	NA	BPDoo	TT	No			7.4					
Yang 2010	50	17–34	BPD, HC	NA	Yes	Op1: -0.09 Op2: 0.16			-0.09	Op1: -0.27; Op 2: -0.2	0.66		

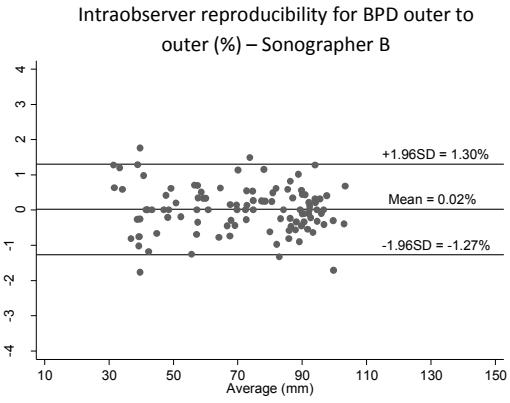
Only the first author of each study is given. *Only study reporting occipitofrontal diameter (OFD) mean difference in intraobserver reproducibility in the second trimester (2T) (1.26 mm) and third trimester (3T) (1.46 mm). †Biparietal diameter outer-to-inner (BPDoi). ‡Biparietal diameter outer-to-outer (BPDoo). 1T, first trimester; BA, Bland–Altman plot; BPD, biparietal diameter; CP, caliper placement; GA, gestational age; HC, head circumference; Inter., interobserver reproducibility; Intra., intraobserver reproducibility; NA, not available; Op, operator; TT, transthalamic plane; TV, transventricular plane.

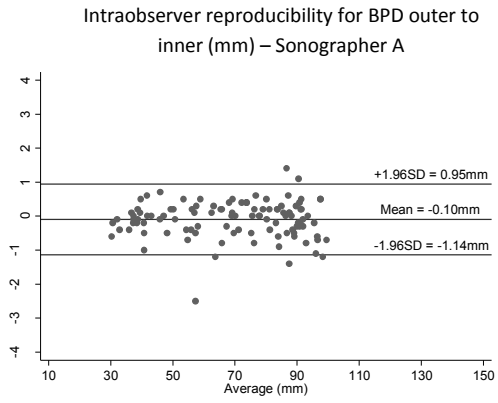
Figure S1 Bland–Altman plots showing intraobserver reproducibility for Sonographers A and B of outer-to-outer and outer-to-inner caliper placement when measuring biparietal diameter (BPD) in the transthalamic plane. Plots on left show absolute difference (in mm) and plots on right show reproducbility as a percentage.

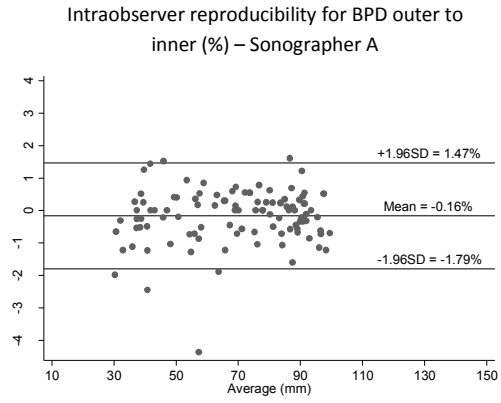




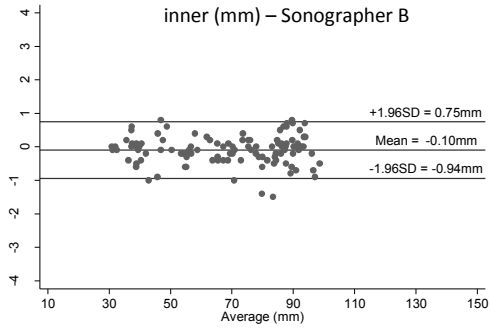








Intraobserver reproducibility for BPD outer to inner (mm) – Sonographer B



Intraobserver reproducibility for BPD outer to inner (%) – Sonographer B

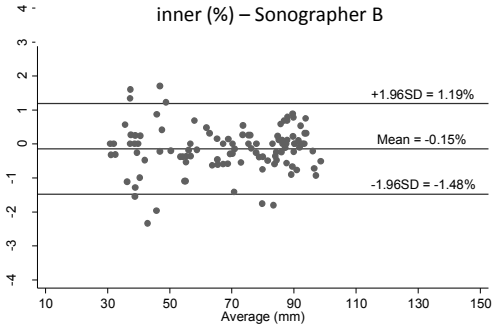


Figure S2 Bland–Altman plots showing interobserver reproducibility of outer-to-outer and outer-to-inner caliper placement when measuring biparietal diameter (BPD) in the transthalamic plane. Plots on left show absolute difference (in mm) and plots on right show reproducbility as a percentage.

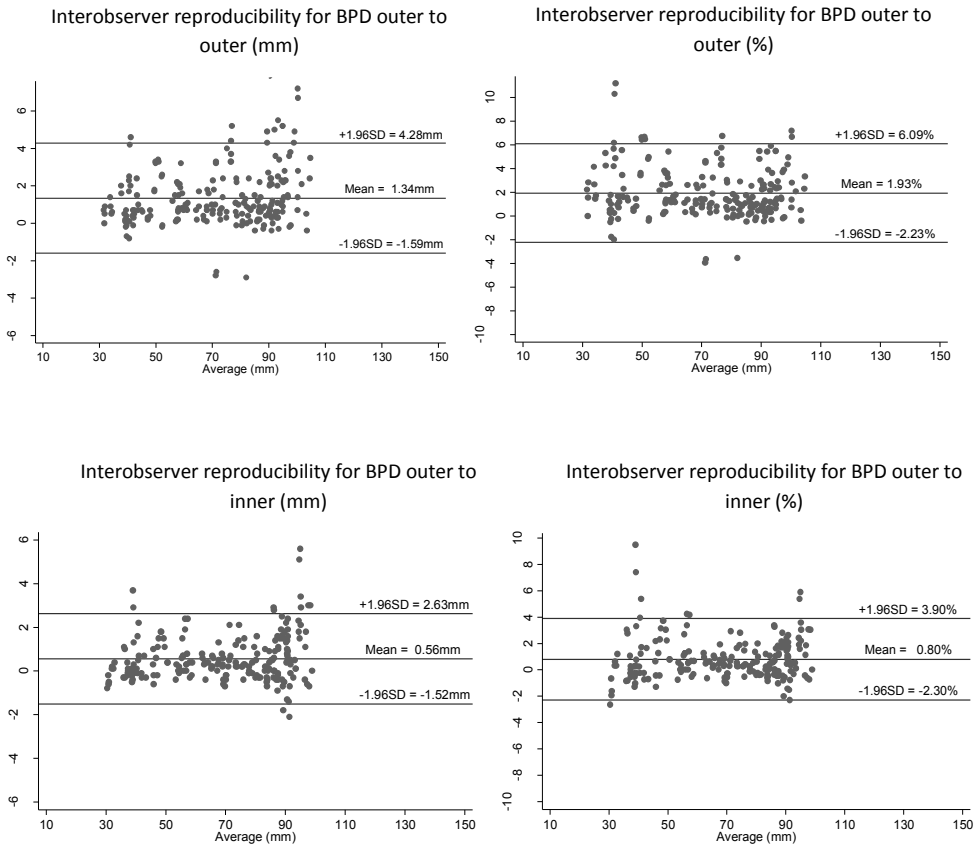
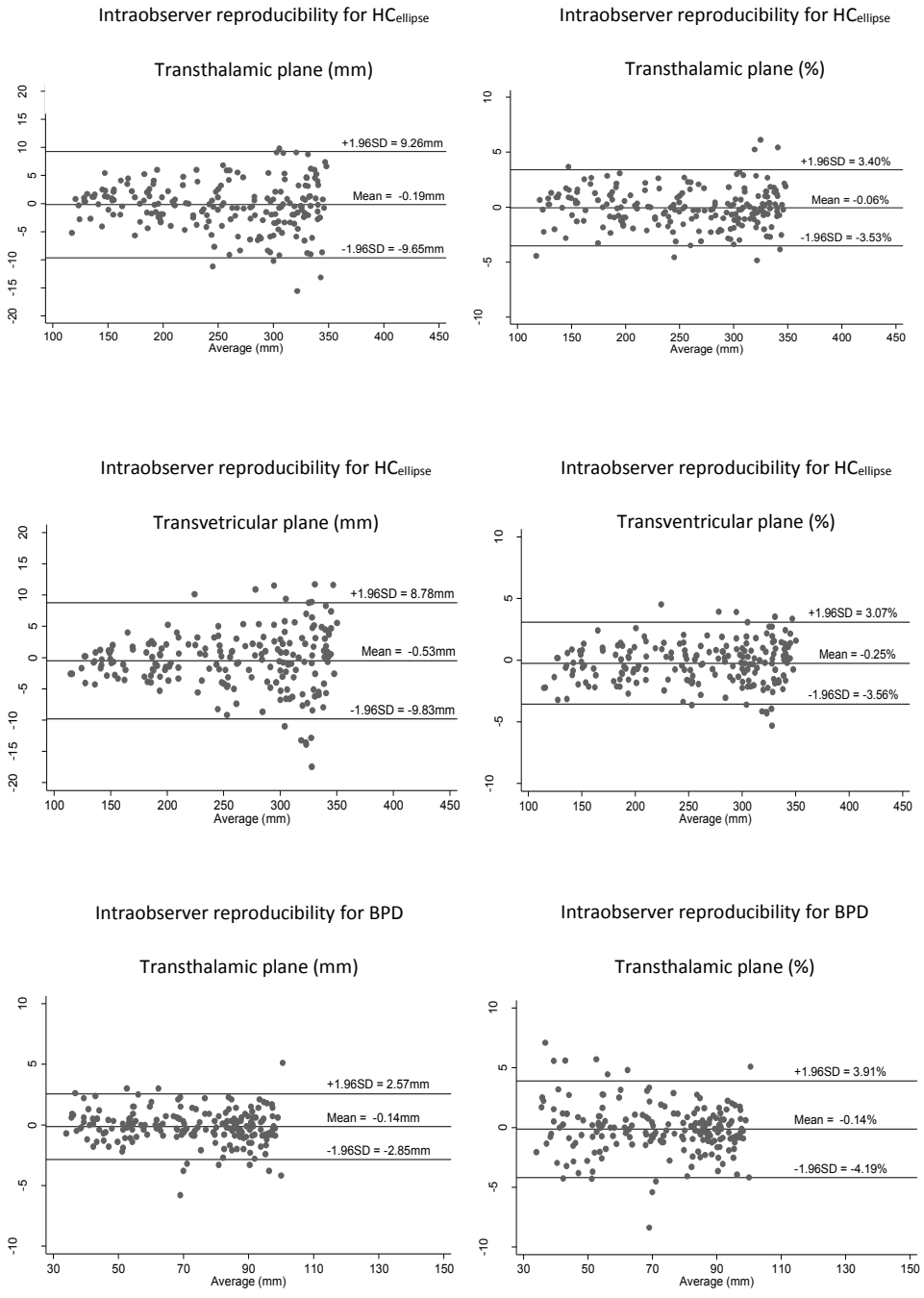
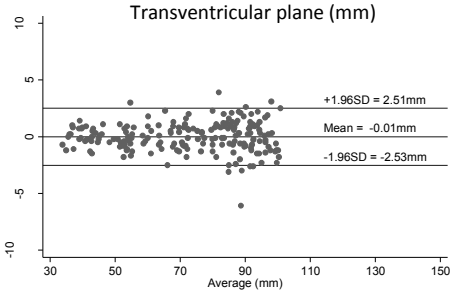


Figure S3: Bland–Altman plots showing intraobserver reproducibility, in the transthalamic and transventricular planes, of acquiring and measuring head circumference using the ellipse facility ($HC_{ellipse}$), biparietal diameter (BPD), occipitofrontal diameter (OFD) and head circumference calculated from the two perpendicular head diameters BPD and OFD ($HC_{calculated}$). Plots on left show absolute difference (in mm) and plots on right show reproducbility as a percentage.



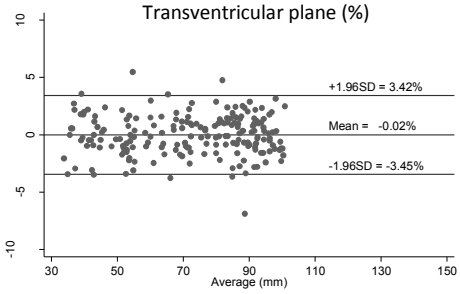
Intraobserver reproducibility for BPD

Transventricular plane (mm)



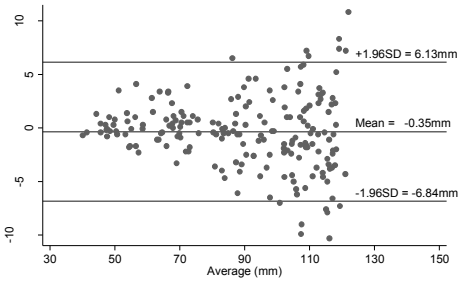
Intraobserver reproducibility for BPD

Transventricular plane (%)



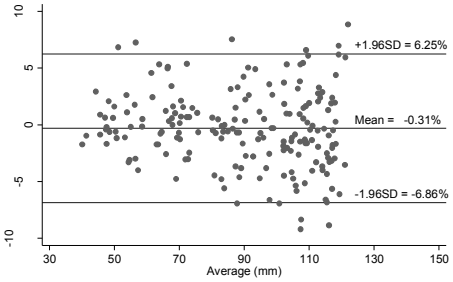
Intraobserver reproducibility for OFD

Transthalamic plane (mm)



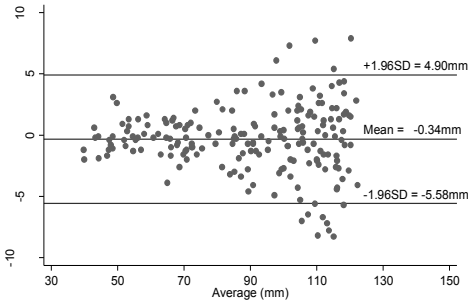
Intraobserver reproducibility for OFD

Transthalamic plane (%)



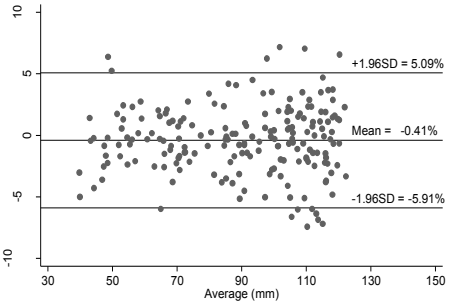
Intraobserver reproducibility for OFD

Transventricular plane (mm)



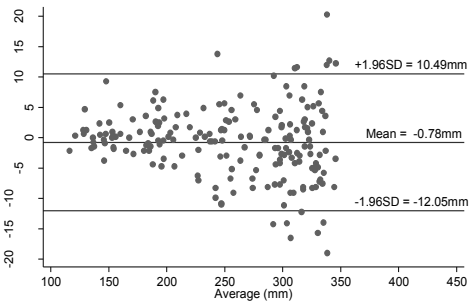
Intraobserver reproducibility for OFD

Transventricular plane (%)



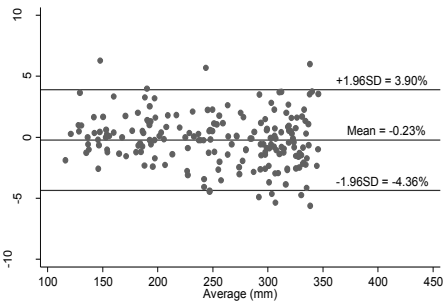
Intraobserver reproducibility for $HC_{calculated}$

Transthalamic plane (mm)



Intraobserver reproducibility for $HC_{calculated}$

Transthalamic plane (%)

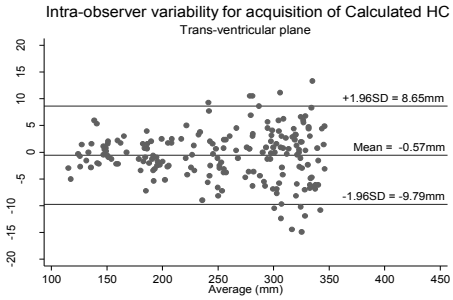


Intraobserver reproducibility for $HC_{calculated}$

Transvetricular plane (mm)

Intraobserver reproducibility for $HC_{calculated}$

Transventricular plane (%)


Intra-observer variability for acquisition of Calculated HC
Trans-ventricular plane
+1.96SD = 8.65mm
Mean = -0.57mm
-1.96SD = -9.79mm
Average (mm)

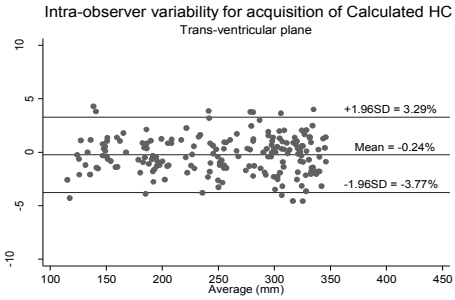
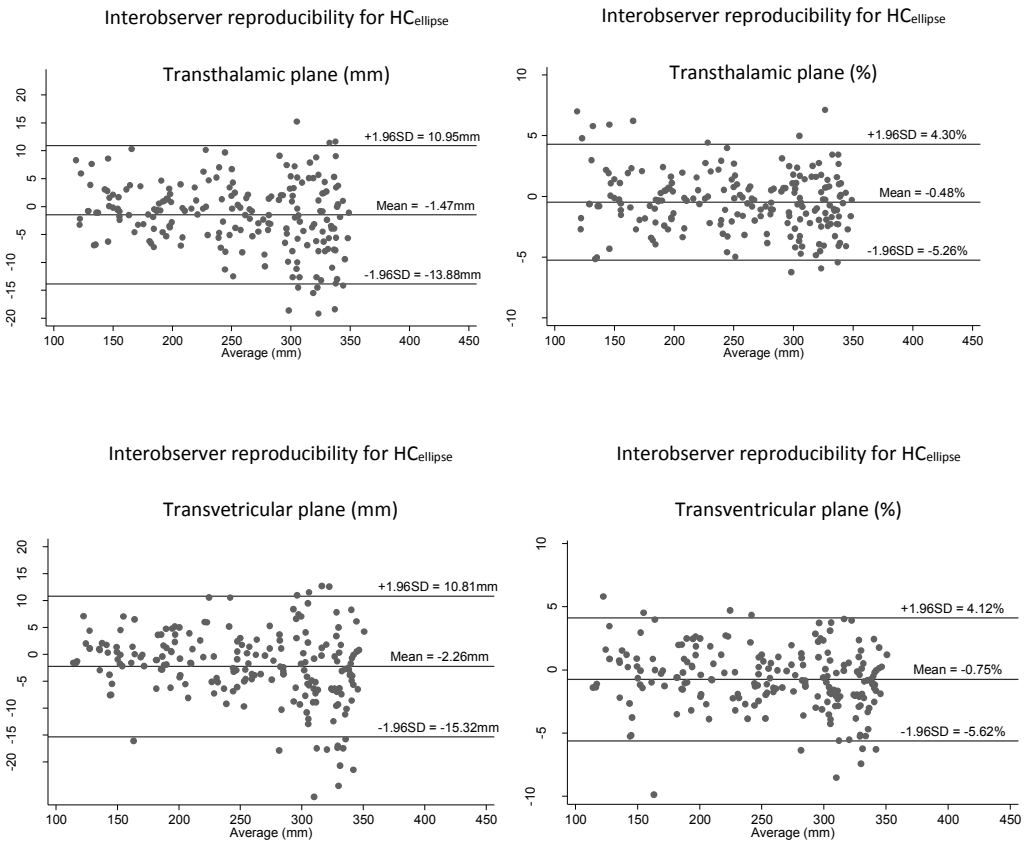
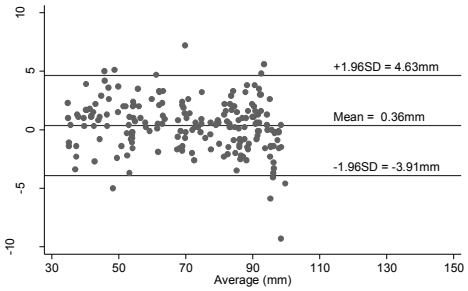

Intra-observer variability for acquisition of Calculated HC
Trans-ventricular plane
+1.96SD = 3.29%
Mean = -0.24%
-1.96SD = -3.77%
Average (mm)

Figure S4: Bland–Altman plots showing interobserver reproducibility, in the transthalamic and transventricular planes, of acquiring and measuring head circumference using the ellipse facility ($HC_{ellipse}$), biparietal diameter (BPD), occipitofrontal diameter (OFD) and head circumference calculated from the two perpendicular head diameters BPD and OFD ($HC_{calculated}$). Plots on left show absolute difference (in mm) and plots on right show reproducbility as a percentage.



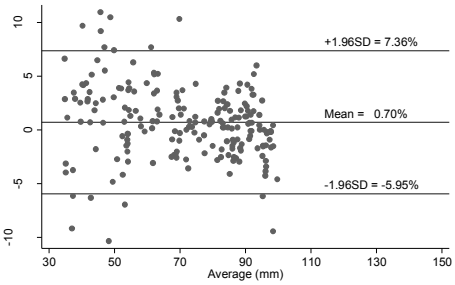
Interobserver reproducibility for BPD

Transthalamic plane (mm)



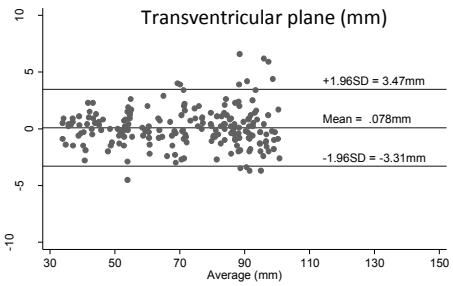
Interobserver reproducibility for BPD

Transthalamic plane (%)



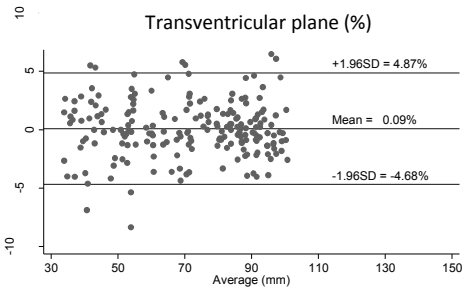
Interobserver reproducibility for BPD

Transventricular plane (mm)



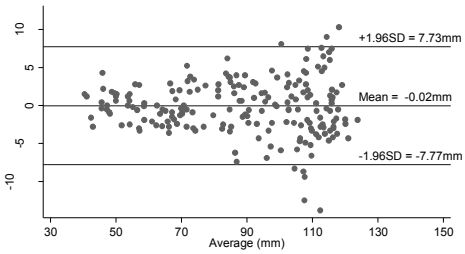
Interobserver reproducibility for BPD

Transventricular plane (%)



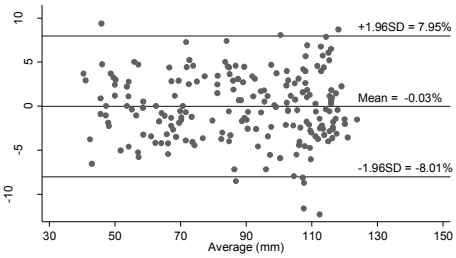
Interobserver reproducibility for OFD

Transthalamic plane (mm)



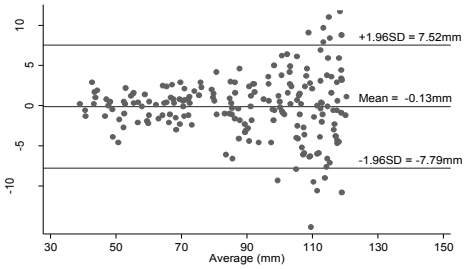
Interobserver reproducibility for OFD

Transthalamic plane (%)



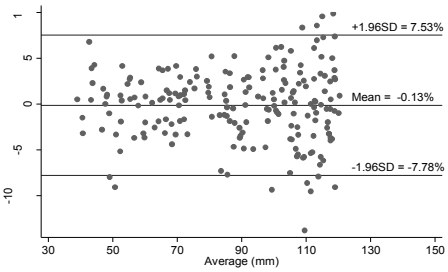
Interobserver reproducibility for OFD

Transventricular plane (mm)



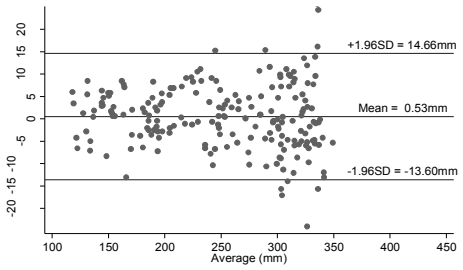
Interobserver reproducibility for OFD

Transventricular plane (%)



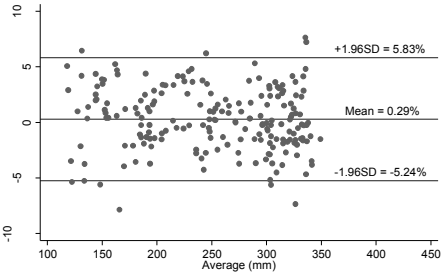
Interobserver reproducibility for $HC_{calculated}$

Transthalamic plane (mm)



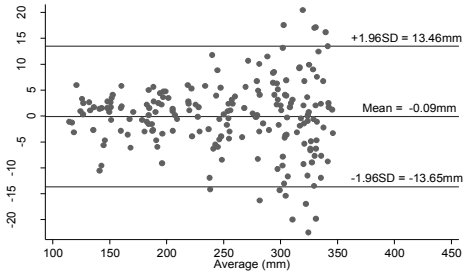
Interobserver reproducibility for $HC_{calculated}$

Transthalamic plane (%)



Interobserver reproducibility for $HC_{calculated}$

Transvetricular plane (mm)



Interobserver reproducibility for $HC_{calculated}$

Transventricular plane (%)

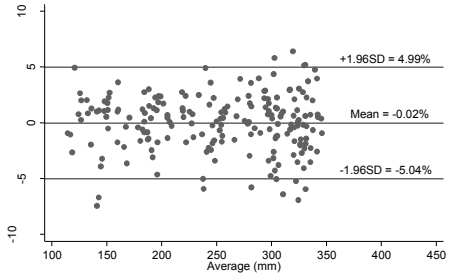
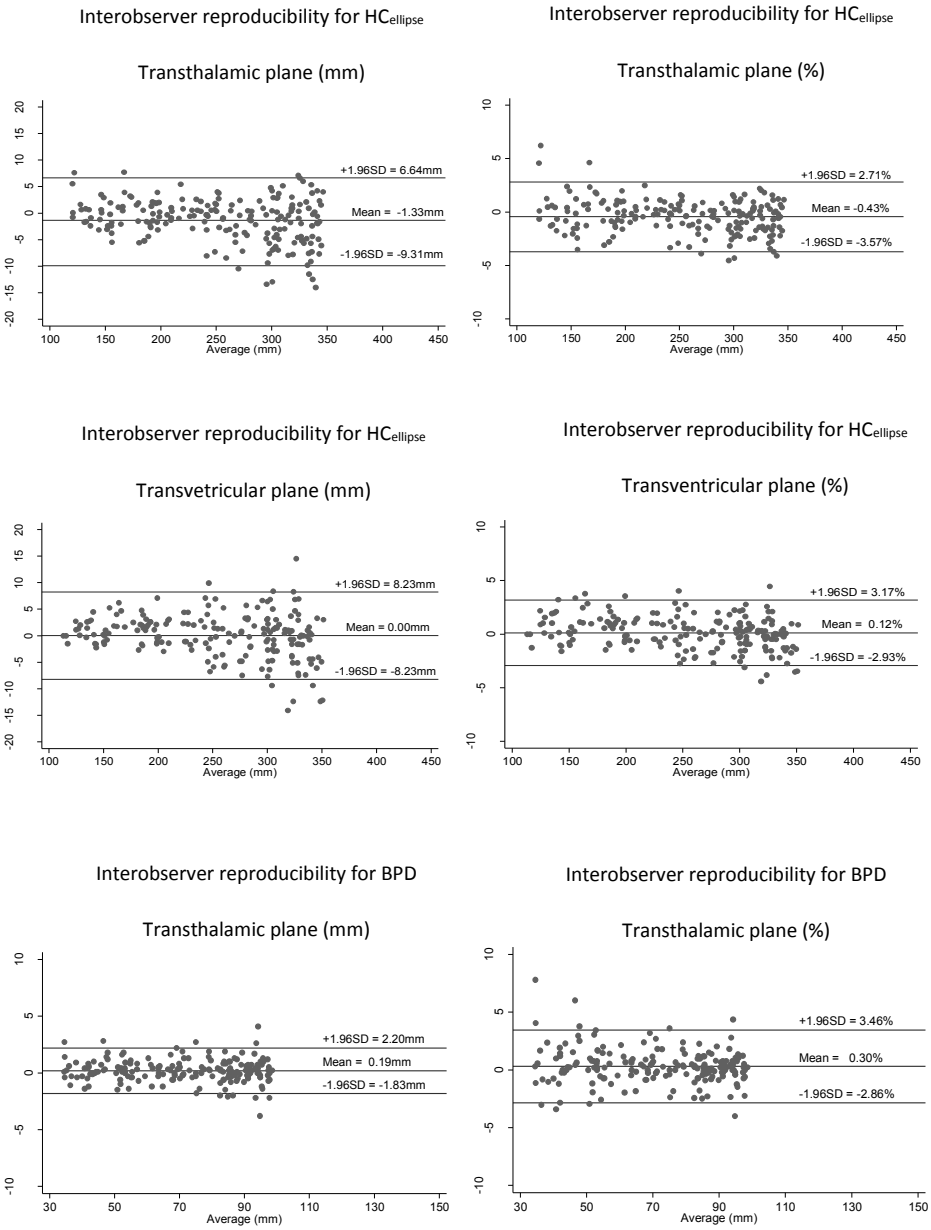
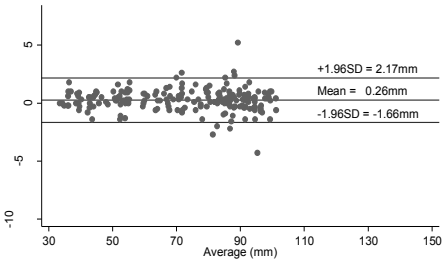


Figure S5: Bland–Altman plots showing interobserver reproducibility of caliper replacement, in transthalamic and transventricular planes, for measuring head circumference using the ellipse facility ($HC_{ellipse}$), biparietal diameter (BPD), occipitofrontal diameter (OFD) and head circumference calculated from the two perpendicular head diameters BPD and OFD ($HC_{calculated}$). Plots on left show absolute difference (in mm) and plots on right show reproducbility as a percentage.



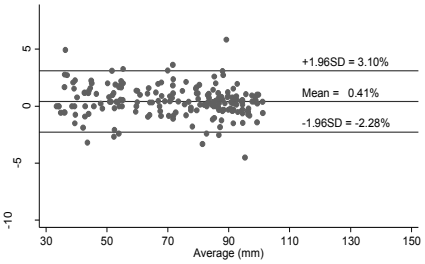
Interobserver reproducibility for BPD

Transventricular plane (mm)



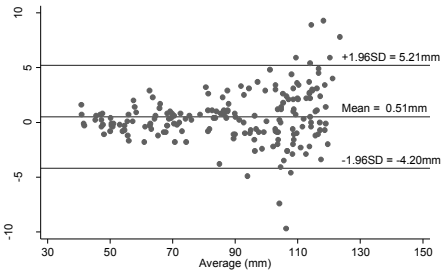
Interobserver reproducibility for BPD

Transventricular plane (%)



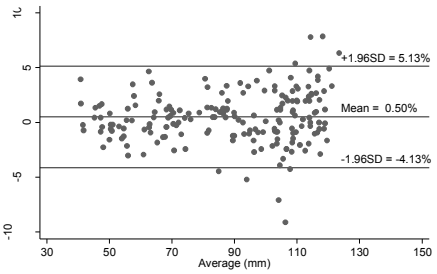
Interobserver reproducibility for OFD

Transthalamic plane (mm)



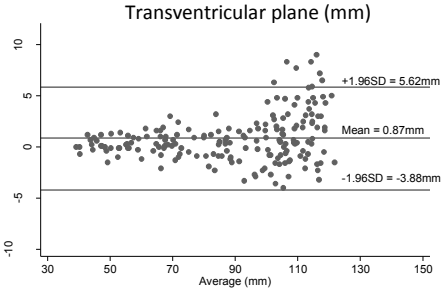
Interobserver reproducibility for OFD

Transthalamic plane (%)



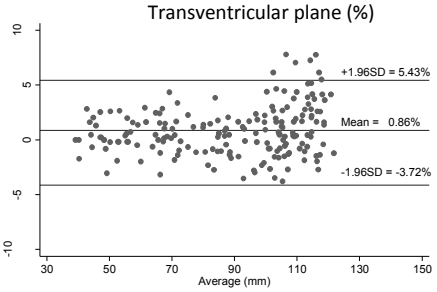
Interobserver reproducibility for OFD

Transventricular plane (mm)



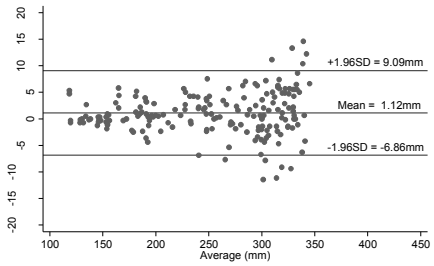
Interobserver reproducibility for OFD

Transventricular plane (%)



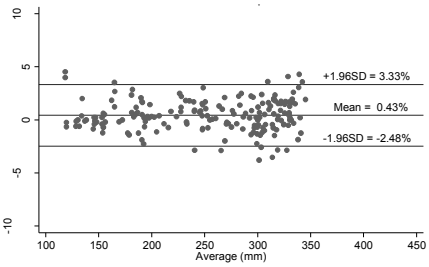
Interobserver reproducibility for $HC_{calculated}$

Transthalamic plane (mm)



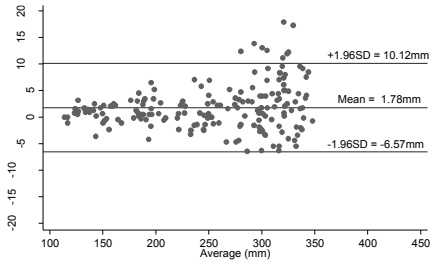
Interobserver reproducibility for $HC_{calculated}$

Transthalamic plane (%)



Interobserver reproducibility for $HC_{calculated}$

Transvetricular plane (mm)



Interobserver reproducibility for $HC_{calculated}$

Transventricular plane (%)

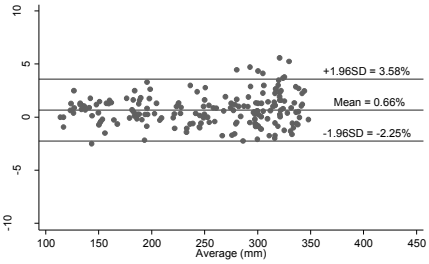
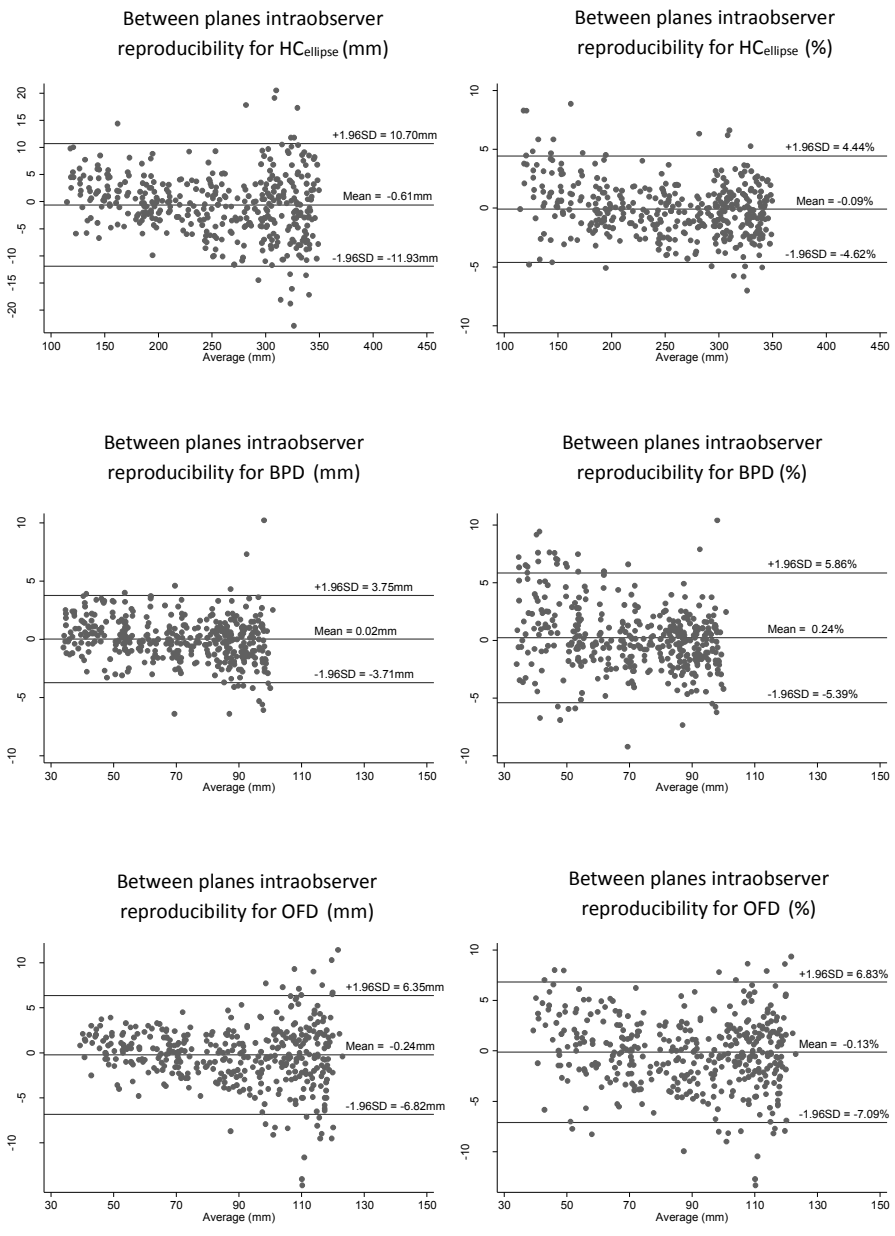


Figure S6: Bland–Altman plots showing between-plane intraobserver reproducibility in the transthalamic and transventricular planes, of acquiring and measuring the head circumference using the ellipse facility ($HC_{ellipse}$), biparietal diameter (BPD), occipitofrontal diameter (OFD), head circumference calculated from the two perpendicular head diameters ($HC_{calculated}$). Plots on left show absolute difference (in mm) and plots on right show reproducbility as a percentage.



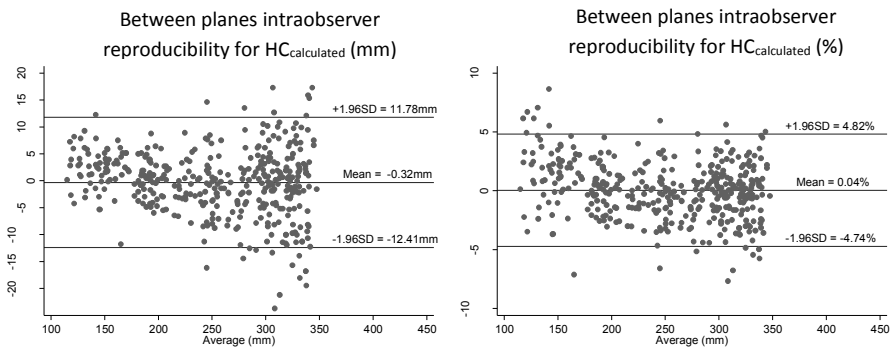
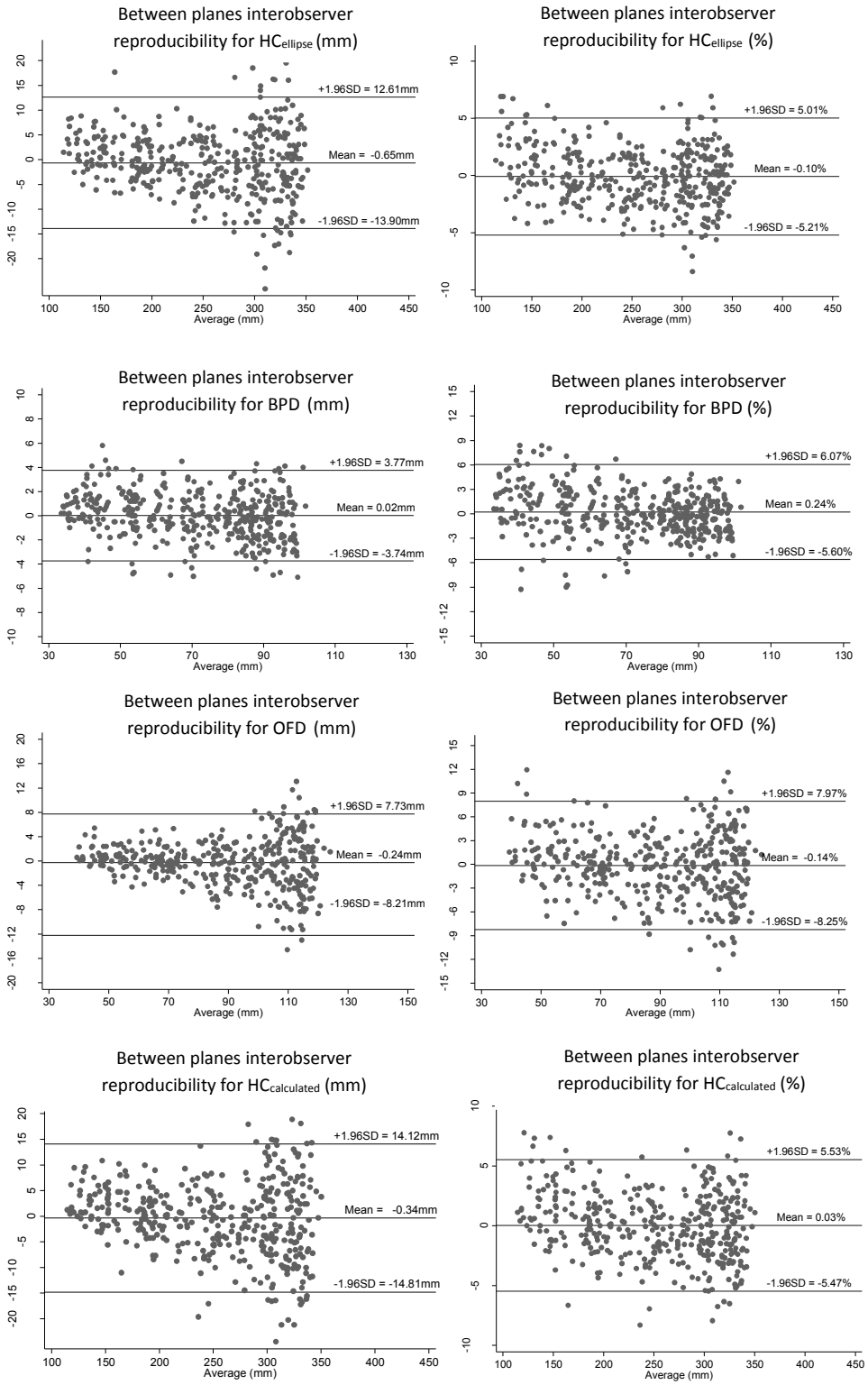

Between planes intraobserver reproducibility for HC calculated (mm)
+1.96SD = 11.78mm
Mean = -0.32mm
-1.96SD = -12.41mm
Average (mm)
Between planes intraobserver reproducibility for HC calculated (%)
+1.96SD = 4.82%
Mean = 0.04%
-1.96SD = -4.74%
Average (mm)

Figure S7: Bland–Altman plots showing between-plane interobserver reproducibility in transthalamic and transventricular planes, of acquiring and measuring head circumference using the ellipse facility ($HC_{ellipse}$), biparietal diameter (BPD), occipitofrontal diameter (OFD), head circumference calculated from the two perpendicular head diameters BPD and OFD ($HC_{calculated}$). Plots on left show absolute difference (in mm) and plots on right show reproducbility as a percentage.



FETAL BRAIN STRUCTURES SIZE CHARTS

Central nervous system of the fetus is routinely assessed by ultrasound antenatally mainly to diagnose fetal anomalies and provide a useful tool to estimate the gestational age late in pregnancy. Central nervous system anomalies are a major component of fetal abnormalities detected antenatally.[2]

Main assessment of fetal brain structures is performed using a subjective analysis of the morphology and a quantitative assessment using structures biometry. The latter provided to be a more objective method, with higher reproducibility,[41, 86] and allowing quantitative calculation to assess the relative growth of brain structures with advancing gestation.[5]

The use of the appropriate chart is therefore essential in research and clinical practice, whereas the use of different reference charts can affect the diagnostic ability of ultrasound in detecting fetal abnormality, can affect clinical decision and impair generalisability of results from research studies using different cut-offs.

A systematic review of the literature has been performed to identify all the studies aimed to create brain structures charts. Only studies reporting on six specific fetal brain structures of relevant clinical interest obtained on axial planes were included in the final analysis (the parieto-occipital fissure (POF) and the sylvian fissure (SF) in the TT plane; the anterior ventricle (AV) and the posterior ventricle (PV) in the TV plane; the transcerebellar diameter (TCD) and cisterna magna (CM) in the TC plane).

Four major electronic databases (MEDLINE, EMBASE, Cochrane Library and Science Citation Index & Conference Proceedings Citation Index) were systematically searched from 1946 to June 2016. Only articles written in English were considered. Article reporting on animal studies, case reports, food, comments, letters, editorials were excluded. A search strategy was formulated in collaboration with a professional information specialist. The following keywords were entered: fetal or foetal or fetus or foetus AND ultrasound or ultrasonogra* or ultra-sonogra* or sonic* or scan* AND brain or cerebral ventricles or lateral ventricles or cisterna magna or cranial fossa, posterior or exp cerebellum OR brain or cerebell* or transcerebell* or cerebral OR cerebellar or transcerebellar or cerebellum or cerebral cortex OR posterior fossa or cisterna magna OR sylvian fissure or lateral sulcus or lateral fissure or perisylvian cortex or cereb* fissure or brain fissure or parietooccipital fissure or parieto-occipital fissure or parietooccipital sulcus or parieto-occipital sulcus OR lateral ventric* OR brain or cereb* or lateral anterior or posterior AND embryonic and fetal Development or fetal development or gestational age AND reference standards or reference values OR reproducibility of results OR predictive value of tests OR observer variation OR reference or normal OR reference or growth OR correlat* or reproducib* or variation or validat* OR nomogram or nomograph OR biometry or biometric OR percentile or centile.

More than 570 articles were identified after removal of duplicates. 95 articles underwent abstract and full paper review and 36 studies were finally identified.[43-52, 87-112] There is substantial heterogeneity in the methodology used. High risk of bias in several domains have been identified including the selection of the population, the ultrasound protocol and the analysis of the

data. Less than 10% of the identified studies reported on maternal and fetal inclusion criteria, pregnancy outcome, ultrasound quality control and statistical description. Most importantly, no studies reported on long term infant outcome, most probably due to the retrospective descriptive design of data collection which was non-specific for the purpose of the study. Not surprisingly, these are common finding in creating fetal biometry charts as found in previous systematic reviews.[36, 38]

To overcome such limitations in previous studies, the main aim of this project is to create international standards for six fetal brain structures by antenatal ultrasound.

The study was conducted in women taking part in the INTERGROWTH-21st Project whose babies have a low risk of FGR and consequently low risk of abnormal neurological outcome. This is confirmed in the study findings as more than 99% of the babies with known motor development were normal at 1 year of age.

Appendix 9: Normal fetal brain structures size: standards based on ultrasound measurements from the Fetal Growth Longitudinal Study of the INTERGROWTH-21st PROJECT.

This is study is under the review of the Scientific Steering Committee of the INTERGROWTH-21st Project and it might undergo substantial review of the data before the publication in the journal peer reviewed process.

ABSTRACT

Objective: To create prescriptive growth charts of six fetal brain structures measured by ultrasound from the optimally grown fetal population of the INTERGROWTH-21st Project.

Methods: This was a prospective multiethinic multicentre cross-sectional study aimed to assess the size of parieto-occipital fissure (POF), sylvian fissure (SF), anterior ventricle (AV), posterior ventricle (PV), transcerebellar diameter (TCD) and cisterna magna (CM) in planes reconstructed from head volumes acquired from women at low risk of abnormal fetal growth and perinatal complications. Fetuses were randomly recruited ensuring an equal distribution between the 8 countries of origin and week of gestation (range: 15 - 36 weeks). Children long term follow up was assessed by motor assessment at 1 year of age. The best fitting powers were provided by second-degree

fractional polynomials and further modelled in a multilevel framework to account for the cross-sectional design of the study.

Results: 451 fetuses were recruited and after exclusions a total of 442 volumes from live singletons without congenital malformations were used to create the charts. Motor assessment was available in 297 cases and it was normal in 98% of them. Structures were measurable in 90% of cases. Mean and standard deviations observed were 5.47 (1.91), 9.45 (4.22), 7.61 (1.54), 6 (1.59), 28.97 (9.32), 5.27 (1.66) mm for the POF, SF, AV, PV, TC and CM respectively, showing increasing size (all) and variability (POF, SF, PV, TCD, CM) with advancing gestation. 5th, 50th, 95th smoothed centile were calculated.

Conclusions: Prescriptive brain structures size charts were created from fetuses at low risk of long term abnormal development. The proposed charts should be recommended as international standards for fetal brain structures measurements by ultrasound.

INTRODUCTION

During pregnancy, the anatomy of the development of the fetal brain can be assessed using ultrasound. On most settings this is undertaken as part of a routine assessment of the fetal anatomy at around 20 weeks of gestation, and the main aims are to demonstrate anatomical integrity; and to diagnose abnormalities of the central nervous system (CNS). Such anomalies can be visualised directly (for example absence of a structure, such as the corpus callosum); or indirectly (such as a banana shaped cerebellum in open spina bifida). Measurement of intracranial structures forms part of this, and often includes assessment of the head size; width of the atrium of the posterior ventricle; cerebellar diameter and cisterna magna.[1] In more advanced neurosonography, undertaken due to indications such as previous or suspected abnormality, measurements of other structures or at different gestations is also practiced – either earlier in gestation such as in cases of previous history, or late in gestation, assessing for instance gyration and sulcation patterns.[2-6] Measurement of structures can also be assessed antenatally to estimate gestational age.[7]

Because subjective evaluation of fetal brain structures is associated with high variability,[8] quantitative estimation using biometric measurements are generally used; however, there are several limitations of existing charts.[9-16] This may contribute to variability of interpretation of ultrasound diagnosis of CNS abnormalities.[17]

In some sense, these aspects are generic to ultrasound measurement. A similar lack of a standard approach due to use of different reference charts has also been observed in fetal biometry and pregnancy dating.[18, 19] In addition, the recommendation for evaluation using a prescriptive approach (using standards) rather than descriptive approach (using references) has led us to produce international standards for pregnancy dating, fetal growth and other aspects of pregnancy care.[20-23]

To complement these products we present here standards for size estimation of six fetal brain structures in a multiethnic population of healthy women taking part in the INTERGROWTH-21st Project whose babies have a low risk of abnormal developmental outcome.

METHODS

Study population

The study was performed in women recruited as part of INTERGROWTH-21st Project (www.intergrowth21.org.uk), a multicentre, multiethnic, population-based project, conducted between 2008 and 2013 in eight countries. The Fetal Growth Longitudinal Study (FGLS) involves both two-dimensional and three-dimensional serial fetal scans performed every 5 weeks from 14+0 to 41+6 weeks.[21] Women participating in this study have low-risk pregnancies that fulfil well defined and strict inclusion criteria at recruitment.[24] Briefly,

inclusion criteria were maternal age between 18 and 35 years, body mass index (BMI) ≥18.5 and <30 kg/m2, a singleton pregnancy, normal pregnancy history without relevant past medical history, no evidence of socioeconomic constraints likely to impede fetal growth, no use of tobacco or recreational drugs and no heavy alcohol consumption. Women also had to have a known date of last menstrual period (LMP) with regular cycles without hormonal contraceptive use or breastfeeding during the 2 months before pregnancy and natural conception; gestational age was based on LMP if standardized ultrasound measurement of crown–rump length between 9+0 and 14+0 weeks was in agreement within 7 days.[25]

Detailed pregnancy outcomes, and where available, motor assessment at age 1 year are reported. One-year follow up of infants was collected by interview of parents or assessment by a certified examiner. Achievement of milestones (sitting without support, standing with assistance, hand-and-knees-crawling, walking with assistance, standing alone and walking alone) were considered normal if the proportion of babies achieving milestones was similar to expected windows of achievement (less than the 99th centile child age for each of the expected windows).[26]

All ultrasound scans were performed by sonographers trained, standardised and regularly audited according to the FGLS standards.[27, 28] The same commercially available ultrasound machine (Philips HD-9, Philips Ultrasound, Bothell, WA, USA) with curvilinear abdominal two-dimensional transducers (C5-2, C6-3) and one curvilinear abdominal three-dimensional transducer (V7-3) was used for all growth scans. For the purposes of the INTERGROWTH-21st Project, the manufacturer reprogrammed the machine's software to

ensure that the measurement values do not appear on screen during the scan, in order to reduce operator “expected value” bias. The INTERGROWTH-21st Project was approved by the Oxfordshire Research Ethics Committee “C” (ref: 08/H0606/ 139); all the pregnant women involved gave informed written consent.

Three-dimensional ultrasound volumes of the fetal head were selected using computer randomisation from pregnancies recruited into the FGLS and ensuring an equal distribution between country of origin and gestational age week between 16 and 36 weeks.

Structures measured, volume manipulation and measurement methodology

Based on an extensive scoping exercise; review of the literature; and a pilot study involving 90 volumes, we aimed to create standards for three commonly used brain structures, namely the PV, TCD and CM[1] and three other, clinically relevant structures that may be relevant in an extended examination (POF, SF, AV).[9, 10, 13, 14, 29] These fetal brain structures were measured on still images retrieved from three-dimensional head volumes acquired in all eight recruiting units participating in the main study (Brazil, Italy, Oman, UK, USA, China, India, Kenya).

Detailed definitions of the methodology for volume acquisition are provided elsewhere.[27, 30] Briefly, head volumes were acquired at the level of the axial transthalamic plane. Six predefined quality control criteria for the transthalamic plane had to be satisfied to acquire the volume (Table 1) (Figure

1).[28] Acquisition was undertaken with the volume data box and angle of sweep (usually 70 degrees) adjusted to include the entire skull; during fetal quiescence; with the mother asked to hold her breath; and with the transducer held steady. The real time image was observed during acquisition to confirm that the sweep included the entire skull with no maternal or fetal movement during the sweep, otherwise the process was repeated. All data were then sent to the coordinating unit in Oxford.

Offline analysis was undertaken by four experienced sonographers at the coordinating unit, who were trained and standardised in volume manipulation and fetal neurosonography. Volume manipulation for plane reconstruction and measurements were performed using the manufacturer software of the ultrasound machine or using the open-source image analysis software program MITK (Medical Imaging Interaction Toolkit MITK, version 0.12.2, German Cancer Research Center, Division of Medical and Biological Informatics, www.mitk.org).[31] First, stored volumes of the fetal head were upload onto the multiplanar mode facility. Second, three standard two-dimensional fetal brain measurement planes were extracted from each volume, namely the transventricular, transthalamic and transcerebellar planes. As the transthalamic plane was the plane of volume acquisition, it required minimum manipulation for the relevant structures to be visualised and was chosen for the measurements of the fissures. Starting from this plane, the operator rotated or scrolled the volume in orthogonal planes with the fulcrum or rotation primarily in the middle of the cavum of the septi pellucidi.[11, 32] A movement to a more cranial level resulted in the transventricular plane, with the lateral ventricles located symmetrically on each side of the midline, the

anterior and posterior horns both visible, and the posterior ventricle cavity visualised as a hypoechoic structure (Figure 1).[33] By rotating the volume onto the Y axis the transcerebellar plane was visualised including the cerebellum at its largest diameter (Figure 1).[29]

Image quality criteria were used to ensure the maximum possible standard for each extracted plane (Table 1) before measurement of the following six structures: the parieto-occipital fissure (POF) and the sylvian fissure (SF) in the transthalamic plane; the anterior ventricle (AV) and the posterior ventricle (PV) in the transventricular plane; the transcerebellar diameter (TCD) and cisterna magna (CM) in the transcerebellar plane.

Caliper placement for measurement acquisition

The POF, the SF, the AV and the PV were measured in the distal hemisphere of the respective plane (due to the lower resolution in the proximal hemisphere). The POF was measured by placing the caliper form the inner edge of the falx to the inner edge of the fissure ('inner to inner') at its widest point, parallel to the biparietal diameter (modified from Alves et al.)[9] The SF was measured from the lateral edge of the roof of the fissure to the medial edge of the skull at its widest point, parallel to the biparietal diameter ('inner to inner').[13] Calipers for the AV were positioned between the internal margin of the midline falx and the lateral wall of anterior horn ('inner to inner').[14] Calipers for the PV were positioned between the internal margin of the medial and lateral wall of the ventricle cavity ('inner to inner'), at the level of the glomus of the choroid plexus, on an axis perpendicular to the long axis of the lateral

ventricle (Figure 1).[29] The TCD was measured in the transcerebellar plane, perpendicular to the falx, with the calipers placed “outer to outer” between the distal margins of the hemispheres at the largest transverse diameter of the cerebellum.[34] The CM was measured in the transcerebellar plane by placing the calipers from the posterior wall of the cerebellum at a level middle to the vermis to the inner wall of the skull (‘inner to inner’) (Figure 1).[29]

Reproducibility

This was assessed in a subset of 90 volumes. The first sonographer uploaded the volume, extracted the three planes and measured the six structures twice (intraobserver reproducibility for plane reconstruction and measurement acquisition). A second sonographer, re-upload the same volume and repeated this process (interobserver reproducibility for plane reconstruction and measurement acquisition). To assess the contribution of caliper replacement, the second sonographer replaced the calipers on still images and repositioned them to measure all structures in each plane stored by the first sonographer (interobserver reproducibility for calipers replacement on stored images). All sonographers were blinded to their own and the others measurements during the reproducibility study but also the main study.

Statistical analysis

The sample size was based on pragmatic and statistical considerations; the former was based on time frame necessary to obtain all the measurements

from the volumes (20 minutes); the latter focused on the precision at the 5th or the 95th centile, and regression-based reference limits. A sample of 300 scans would obtain precision of 0.1 SD at the 5th or the 95th centile.[35] Assuming a rate of exclusion of 5% and working on a conservative estimate that in 40% of the volumes at least one structure would not be measureable (the upper limit of the confidence interval estimated from the pilot study, primarily due to missing data and movement artefact), it was estimated that 441 volumes would lead to a minimum of 300 volumes analysed. In the event, the actual number measurable was higher than this.

After comparing results from the various approaches, there was no evidence to support a non-normal distribution for a specific gestational age. The study was cross-sectional as volumes were analysed once. Goodness of fitness was assessed by Q-Q plots and a scatter plot of Z-scores by GA. Mean differences between the observed and fitted centiles were also calculated.

For the reproducibility study, Bland-Altman plots were used to estimate mean systematic differences and 95% limits of agreement. Differences between and within observers were expressed in absolute values (mm) for the POF, SF, AV, PV and CM; while they were expressed as a percentage of fetal dimensions for the TCD, to take account of the increase in cerebellar size with gestational age. Percentages were calculated as the difference between two measurements divided by the average of the two measurements multiplied by 100. All analyses were performed using STATA 11 (StataCorp, College Station, Texas, USA).

RESULTS

A total of 451 volumes were selected and after exclusions a total of 442 volumes used to reconstruct planes and create the charts (Figure 2). No congenital malformations were detected either antenatally or postnatally. Maternal demographics and pregnancy outcomes were similar to the overall FGLS population, confirming a low risk of perinatal complications (Table 2).

Follow up of infants by interview of parents was available in 297 out of 442 cases (67%), and 289 infants were assessed by a certified examiner (65%) at 1 year (mean 12.3 months, range 10.9 - 19.4). Motor assessment reported by parents was normal in 99% of the infants, with milestone not achieved (>99th centile of the window of achievement) in 3 (1%), 3 (1%), 0, 0, 0, 0 infants for sitting without support, standing with assistance, hand-and-knees-crawling, walking with assistance, standing alone and walking alone respectively. There was overall good agreement between the achievements of milestones reported by parents compared to examination (average agreement 96%, range 92 to 100%). Reassuringly, in almost all cases where disagreement was present, the examiner reported more precocious milestone achievement than that reported by the parents, confirming the low risk for abnormal long term outcome in our cohort.

In total, 2439 measurements of fetal brain structures were acquired. On average structures were optimally measurable in a high quality extracted

plane in 90% of cases, with the CM being the structure least frequently measurable. After removal of outliers measurements were available to create centiles for POF, SF, AV, PV, TC and CM in 420 (95%), 404 (91.4%), 370 (83.7%), 422 (95%), 390 (88.2%), 352 (79.6%) cases respectively.

The time required for analysis and structures measurement of a single volume was 9 ± 0.8 SD minutes (pilot study). Mean and SD of each measurement in mm were 5.47 (1.91), 9.45 (4.22), 7.61 (1.54), 6 (1.59), 28.97 (9.32), 5.27 (1.66) for the POF, SF, AV, PV, TC and CM respectively.

The gestational age-specific 5th, 50th, and 95th smoothed centiles for POF, SF, AV, PV, TCD and CM are presented in Figure 3. 5th, 50th, and 95th centiles according to gestational age for these ultrasound measures were calculated and reported in Supplementary Table 1.

Goodness of fit by gestational age-specific comparisons of empirical centiles to smoothed centile curves (3rd, 50th, and 97th centiles) and comparing Z-scores showed good agreement. Mean differences between the observed and smoothed centiles for the 3rd, 50th, and 97th centiles, respectively, were small: 0.22 mm (0.5), 0 mm (0.4), 0.17 mm (0.6) for the POF, 0.02 mm (1.1), 0.03 mm (0.7), 0.09 mm (1.1) for the SF, 0.19 mm (0.8), 0.01 mm (0.4), 0.12 mm (0.7) for the AV, 0.22 mm (0.8), 0.07 mm (0.5), 0.04 mm (0.8) for the PV, 0.52 mm (1.6), 0.09 mm (1.1), 0.51 mm (2.6) for the TCD and 0.1 mm (0.36), 0.05 mm (0.4), 0.01 mm (0.9) for the CM .

The equations for the mean and standard deviation from the multilevel regression models for each structure measure are presented in Table 3, allowing for calculations by readers of any desired centiles according to

gestational age in exact weeks. The best fitting powers were provided by second-degree fractional polynomials and further modelled in a multilevel framework to account for the cross-sectional design of the study.

The actual values for these centiles according to gestational age are presented in Supplementary Table 1.

As regards the reproducibility study, the mean difference and 95% limits of agreement are shown in Table 4. All measurements were reproducible within less than 3mm or 12% (all mean differences were less than 0.1mm and 0.5%). The greatest proportion of variability was due to caliper replacement accounting for more than 50% of the intra- and interobserver variability for all structures.

DISCUSSION

In this study we have produced international standards for ultrasound measurements of brain structures, derived from a multi-ethnic population from the FGLS of the INTERGROWTH-21st Project. The design was prescriptive and selected a population of healthy, well nourished pregnant women and their fetuses and newborn babies.[21] The populations were at low risk of obstetric complications and motor assessment at 1 year in keeping with expected norms (Table 2).

We used international guidelines to obtain measurements of the TCD, the CM and the PV;[1, 29] as we were unable to find accepted guidelines on measurement of the depth of the SF or the POF, we developed methods for this.

Previous studies on the subjective assessment of brain fissures report variable results in terms of reproducibility (Kappa coefficients varying from 0.56 to 0.95).[8, 36] One aim of our international standards is to reduce the variability from such subjective non-quantitative assessment of fetal brain size and development.[8, 37, 38]

One of the pitfall in neurosonographic subjective assessment is in the absence of plane standardisation. Using different planes in fetal head biometry can lead to significant measurement difference.[33] In some studies landmarks for plane acquisition are not specified,[10] in other studies various oblique planes with numerous landmarks are proposed.[8, 37] One of the strengths of our study is the use of standardised axial planes recommended in routine clinical practice for biometry assessment (Table 1), and reconstruction from volumes allowed optimal plane finding. The approach of using standardised planes improve reproducibility[36, 39] leading, in our case, to a high percentage of structures measured (90% on average) with high reproducibility (95% limits of agreement were within <0.3mm or <6%) (Table 3). Studies involving experts in neurosonography report similar results in structures visualisation from volume analysis.[40]

We searched for previous studies aimed to create fetal brain structures charts and we identified substantial heterogeneity in the methodology used.[9-16, 34, 41-67] There is high risk of bias in several domains including the selection of the population, the ultrasound protocol and the analysis of the data. Less than 10% of the identified studies reported on maternal and fetal inclusion criteria, pregnancy outcome, ultrasound quality control and statistical method description. Most importantly, no studies reported on long term infant outcome, most probably due to the retrospective descriptive design and the method of collection of the data which was non-specific for the purpose of the study. Not surprisingly, these are common findings in creating fetal biometry charts as found in previous systematic reviews.[18, 19] We identified only three studies reporting charts on the SF[9, 10, 13] and only two on the POF.[9, 10] Increasing variability with advancing gestation is evident from the plotted values of the above studies but this was not computed in the analysis. Reassuringly, our observed measurements range did not differ substantially from previous studies with the lowest risk of methodological bias for each of the six structure.[10, 13-15, 41, 42] Despite all brain structures increase in size with advancing gestation, currently used cut-offs for normality can still be considered safe. For example <1% of PV and CM measurements were above 10mm in our study.

Limitations and strengths

There are some limitations to our study. It can be argued that the use of a large number of sonographers obtaining data might have an impact on the

results; however, we felt that this more accurately reflects clinical practice.68 In addition, the quality of the images obtained in the study was of a high standard and in accordance to a predefined protocol.[27] The setting of near-optimal conditions for scanning was done to minimise potential contribution of confounding factors and this could also be seen as a strength. It is possible that measurements acquired on planes extracted from three-dimensional volumes do not represent of fetal two-dimensional measurements. Although volumetry is associated with high degree or variability if not standardized,[38] once rigorous methodology is adopted, two-dimensional measurements from reconstructed planes can be at least as reproducible as real time measurements and concord to them.[11, 30]

The main strength of our study is the prescriptive design, rather than the descriptive (how structures should grow rather than how they have grown at a specific point in time) which aimed to avoid limitations in previous studies reporting on reference charts. It was truly prospective where women were healthy, well nourished, educated, and at low risk of pregnancy complications. The strategy for population selection was population-based that initially selected geographical regions where women were at low risk of fetal growth restriction, from which, in a second step, pregnant women for FGLS were identified. The ultrasound measurements were taken specifically for the purpose of constructing international standards using a rigorous method implemented across all study sites; standardisation was performed using centrally trained staff; each study site and the coordinating unit used the same specially adapted ultrasound equipment to allow blinding of measurements; we developed a novel quality control strategy for all ultrasound

measurements, including assessment of intraobserver and interobserver variability at all sites and continual independent image review and scoring at the coordinating unit. Finally, the appropriate statistical methods were used to analyse the dataset.

The inevitable and recurrent question related to the implementation of international, prescriptive growth standards is whether or not they can be generalised to all populations. Some authors report on differences in fetal brain structures size across populations. However, these studies are difficult to compare as populations have different demographics between each other, women included have high risk of fetal growth abnormality and outcome is scarcely reported.[44, 49, 57, 63, 69] The generalisability of anthropometric standards based on a prescriptive approach and international sampling frames of geographically and ethnically diverse populations is supported by the uncertainty surrounding the identification of functionally significant, common genetic variants that are unique to ethnic groups in quantitative, complex traits. This is confirmed in neonatal studies analysing fetal brain size.[70]

Our aim was to create international standards, using recommended methods for the analysis and the creation of charts,[71, 72] that can be used in clinical practice. However, we did not propose to produce criteria and cut-offs for detection of abnormality.

Conclusion

International standards for six fetal brain structures growth are reported. Objective and quantitative measurements can help to improve the screening and diagnostic performance of prenatal ultrasound.[73, 74]

The above should represent the standards for protocols of ultrasound measurements and allow comparison between studies on fetal brain structures size and development.

Acknowledgement: This project was supported by a generous grant from the Bill & Melinda Gates Foundation to the University of Oxford (Oxford, UK), for which we are very grateful.

Table 1: Quality criteria for acquisition of the three planes.

TRANSTHALAMIC PLANE	TRANSVENTRICULAR PLANE	TRANSCEREBELLAR PLANE
Symmetrical hemispheres	Symmetrical hemispheres	Symmetrical hemispheres
Cavum of the septum pellucidum present	Cavum of the septum pellucidum present	Cavum of the septum pellucidum present
Thalami visible	Lateral ventricles visible	Thalami visible
No cerebellum visible	No cerebellum visible	Cerebellum present at the maximum diameter
Magnification of 30% image	Magnification of 30% image	Magnification of 30% image

Table 2: Demographic details of the two populations of women recruited in the Fetal Brain Charts Study and the FGLS.

Characteristics	Fetal Brain Charts Study	FGLS
	N= 442	N = 4321
Maternal age, years	28.2 (3.9)	28.4 (3.9)
BMI, kg/m^2	23.4 (3.0)	23.3 (3.0)
Nulliparous (%)	283 (64%)	2955 (68%)
Gestational age at first visit, weeks	11.8 (1.3)	11.8 (1.4)
Years of formal education, years	14.0 (2.9)	15.0 (2.8)
Preterm (<37 weeks)	22 (4.9%)	195 (5%)
Term LBW (<2500 g; ≥37 weeks)	10 (2.2%)	128 (3%)
Birthweight (≥37 weeks), kg	3.2 (0.4)	3.3 (0.4)
Birth length (≥37 weeks), cm	49.2 (1.9)	49.4 (1.9)
Birth head circumference (≥37 weeks), cm	33.9 (1.3)	33.9 (1.3)
Pre-eclampsia	4 (<1%)	31 (<1%)
PPROM (<37 weeks)	6 (1.3%)	80 (2%)
Caesarean section	171 (38%)	1541 (36%)
NICU admission >1 day	33 (7.4%)	240 (6%)
Neonatal mortality	1 (<1%)	7 (<1%)
Mother admitted to intensive care unit	1 (<1%)	17 (<1%)

Maternal baseline characteristics were measured at less than 14 weeks of gestation. Data are mean (SD) or number (%). FGLS = fetal growth longitudinal study. BMI = Body Mass Index. LBW=low birthweight. PPROM=preterm prelabour rupture of membranes. NICU=neonatal intensive care unit.

Table 3: Equations for the estimation of the mean and SD (in mm) of each fetal brain structure measurement according to exact gestational age (in weeks)

Parieto-occipital fissure	
Mean	10.29428 + (-12.28447*(GA/10)^-1) + (0.0103835*(GA/10)^3)
SD	1.596042 + (-2.572297*(GA/10)^-2)
Sylvian fissure	
Mean	80.27012 + (-83.29849*(GA/10)^-0.5) + (-31.67315*((GA/10)^-0.5*LN(GA/10)))
SD	2.304501 + (-3.53814*(GA/10)^-2)
Anterior ventricle	
Mean	6.396214 + (0.0620535*(GA/10)^3)
SD	1.204454
Posterior ventricle	
Mean	4.389214 + (3.810015*(GA/10)^-1) + (0.0020063*(GA/10)^3)
SD	0.6707227 + (0.034258*(GA))
Transcerebellar diameter	
Mean	6.856038+(2.913928*(GA/10)^3)+(-1.66686*(GA/10)^3*LN(GA/10))
SD	0.21404 + (0.1119059*(GA/10)^3)
Cisterna Magna	
Mean	EXP(2.098095 + (-2.390659*(GA/10)^-2) + (-0.0001547*(GA/10)^3))
SD	0.2297936 + (0.081872*(GA/10)^-2))

LN: natural logarithm, GA=exact gestational age.

Table 4: Reproducibility study

	Intraobserver Reproducibility Mean (95% LOA)	Interobserver Reproducibility Mean (95% LOA)	Caliper replacement Reproducibility Mean (95% LOA)
Parieto-occipital fissure (mm)	-0.02 (1.6)	0 (0.19)	-0.01 (0.19)
Sylvian fissure (mm)	-0.01 (2.1)	0 (0.22)	0 (0.28)
Anterior ventricle (mm)	-0.01 (0.18)	-0.02 (0.2)	0 (0.1)
Posterior ventricle (mm)	0 (0.11)	0 (0.18)	0.01 (1.7)
Transcerebellar diameter (%)	-0.08 (8.6)	-0.47 (11.9)	-0.32 (10.54)
Cisterna magna (mm)	0 (1.6)	-0.02 (0.19)	0.01 (1.85)

M: mean, LOA; limits of agreement; Ultrasound: ultrasound machine Philips HD9 using multiplanar 3D measurement modality.

Supplementary Table 1A: Smoothed centiles for parieto-occipital fissure (in mm) according to exact gestational age (in weeks).

GA	Sample	C5	C50	C95
15	18	1.39	2.14	2.88
16	18	1.69	2.66	3.63
17	18	1.96	3.12	4.28
18	19	2.21	3.53	4.85
19	19	2.45	3.90	5.35
20	21	2.67	4.24	5.80
21	16	2.87	4.54	6.21
22	18	3.07	4.82	6.57
23	21	3.25	5.08	6.90
24	18	3.43	5.32	7.21
25	20	3.59	5.54	7.49
26	19	3.75	5.75	7.75
27	19	3.90	5.95	7.99
28	19	4.05	6.13	8.22
29	22	4.19	6.31	8.43
30	21	4.32	6.48	8.63
31	20	4.46	6.64	8.83
32	17	4.58	6.80	9.01
33	22	4.71	6.94	9.18
34	21	4.83	7.09	9.35
35	19	4.95	7.23	9.51
36	15	5.07	7.37	9.67
Total Measurements	420			

Supplementary Table 1B: Smoothed centiles for sylvian fissure (in mm) according to exact gestational age (in weeks).

GA	Sample	C5	C50	C95
15	18	0.57	1.77	2.98
16	15	1.13	2.65	4.17
17	18	1.72	3.49	5.27
18	18	2.31	4.31	6.30
19	17	2.91	5.09	7.27
20	20	3.51	5.85	8.18
21	15	4.10	6.57	9.04
22	18	4.69	7.27	9.86
23	20	5.26	7.95	10.64
24	17	5.82	8.60	11.38
25	20	6.37	9.23	12.09
26	18	6.91	9.84	12.77
27	16	7.44	10.43	13.42
28	19	7.95	11.00	14.05
29	22	8.45	11.55	14.65
30	20	8.94	12.09	15.23
31	20	9.42	12.61	15.79
32	17	9.89	13.11	16.33
33	22	10.34	13.60	16.86
34	22	10.79	14.07	17.36
35	18	11.22	14.54	17.85
36	14	11.64	14.99	18.33
Total Measurements	404			

Supplementary Table 1C: Smoothed centiles for anterior ventricle (in mm) according to exact gestational age (in weeks).

GA	Sample	C5	C50	C95
15	17	4.62	6.61	8.59
16	15	4.67	6.65	8.63
17	17	4.72	6.70	8.68
18	18	4.78	6.76	8.74
19	19	4.84	6.82	8.80
20	20	4.91	6.89	8.87
21	15	4.99	6.97	8.95
22	18	5.08	7.06	9.04
23	21	5.17	7.15	9.13
24	15	5.27	7.25	9.24
25	19	5.38	7.37	9.35
26	18	5.51	7.49	9.47
27	17	5.64	7.62	9.60
28	19	5.78	7.76	9.74
29	22	5.93	7.91	9.89
30	19	6.09	8.07	10.05
31	17	6.26	8.24	10.23
32	13	6.45	8.43	10.41
33	18	6.65	8.63	10.61
34	17	6.85	8.84	10.82
35	15	7.08	9.06	11.04
36	9	7.31	9.29	11.27
Total Measurements	378			

Supplementary Table 1D: Smoothed centiles for posterior ventricle (in mm) according to exact gestational age (in weeks).

GA	Sample	C5	C50	C95
15	18	4.99	6.94	8.88
16	19	4.77	6.78	8.78
17	18	4.58	6.64	8.70
18	19	4.40	6.52	8.64
19	19	4.23	6.41	8.58
20	22	4.08	6.31	8.54
21	16	3.94	6.22	8.51
22	18	3.80	6.14	8.49
23	21	3.67	6.07	8.47
24	18	3.55	6.00	8.46
25	20	3.43	5.94	8.46
26	19	3.32	5.89	8.46
27	19	3.22	5.84	8.46
28	19	3.11	5.79	8.48
29	22	3.01	5.75	8.49
30	21	2.92	5.71	8.51
31	20	2.83	5.68	8.53
32	17	2.74	5.65	8.55
33	22	2.65	5.62	8.58
34	22	2.57	5.59	8.61
35	19	2.49	5.56	8.64
36	14	2.41	5.54	8.67
Total Measurements	422			

Supplementary Table 1E: Smoothed centiles for transcerebellar diameter (in mm) according to exact gestational age (in weeks).

GA	Sample	C5	C50	C95
15	19	13.44	14.41	15.38
16	18	14.48	15.58	16.69
17	18	15.57	16.83	18.08
18	19	16.71	18.14	19.56
19	19	17.89	19.50	21.12
20	21	19.10	20.92	22.75
21	16	20.33	22.39	24.45
22	18	21.58	23.89	26.20
23	21	22.83	25.42	28.01
24	18	24.07	26.97	29.86
25	17	25.29	28.52	31.75
26	19	26.49	30.08	33.67
27	17	27.65	31.62	35.60
28	18	28.76	33.15	37.54
29	21	29.80	34.64	39.48
30	18	30.77	36.09	41.41
31	17	31.65	37.48	43.32
32	16	32.43	38.81	45.19
33	19	33.09	40.06	47.02
34	17	33.62	41.21	48.80
35	14	34.02	42.26	50.50
36	10	34.25	43.19	52.13
Total Measurements	390			

Supplementary Table 1F: Smoothed centiles for cisterna magna (in mm) according to exact gestational age (in weeks).

GA	Sample	C5	C50	C95
15	19	1.82	2.82	4.36
16	17	2.08	3.20	4.92
17	17	2.33	3.56	5.44
18	18	2.56	3.89	5.92
19	19	2.77	4.20	6.36
20	21	2.97	4.48	6.76
21	15	3.15	4.73	7.12
22	18	3.31	4.97	7.45
23	21	3.46	5.18	7.75
24	16	3.60	5.37	8.02
25	17	3.72	5.55	8.27
26	19	3.83	5.71	8.50
27	15	3.94	5.85	8.70
28	16	4.03	5.99	8.89
29	20	4.12	6.11	9.06
30	16	4.20	6.22	9.22
31	13	4.27	6.33	9.36
32	14	4.34	6.42	9.49
33	12	4.40	6.51	9.62
34	13	4.46	6.59	9.73
35	11	4.51	6.66	9.83
36	5	4.56	6.73	9.92
Total Measurements	352			

Figure 1: Planes reconstructed and caliper placement for brain structures acquisition at different weeks of gestation. W: completed weeks of gestation, TT: transthalamic plane, TV: transventricular plane, TC: transcerebellar plane, POF: parieto-occipital fissure, SF: sylvian fissure, AV: anterior ventricle, PV: posterior ventricle, TCD: transcerebellar diameter, CM: cisterna magna.

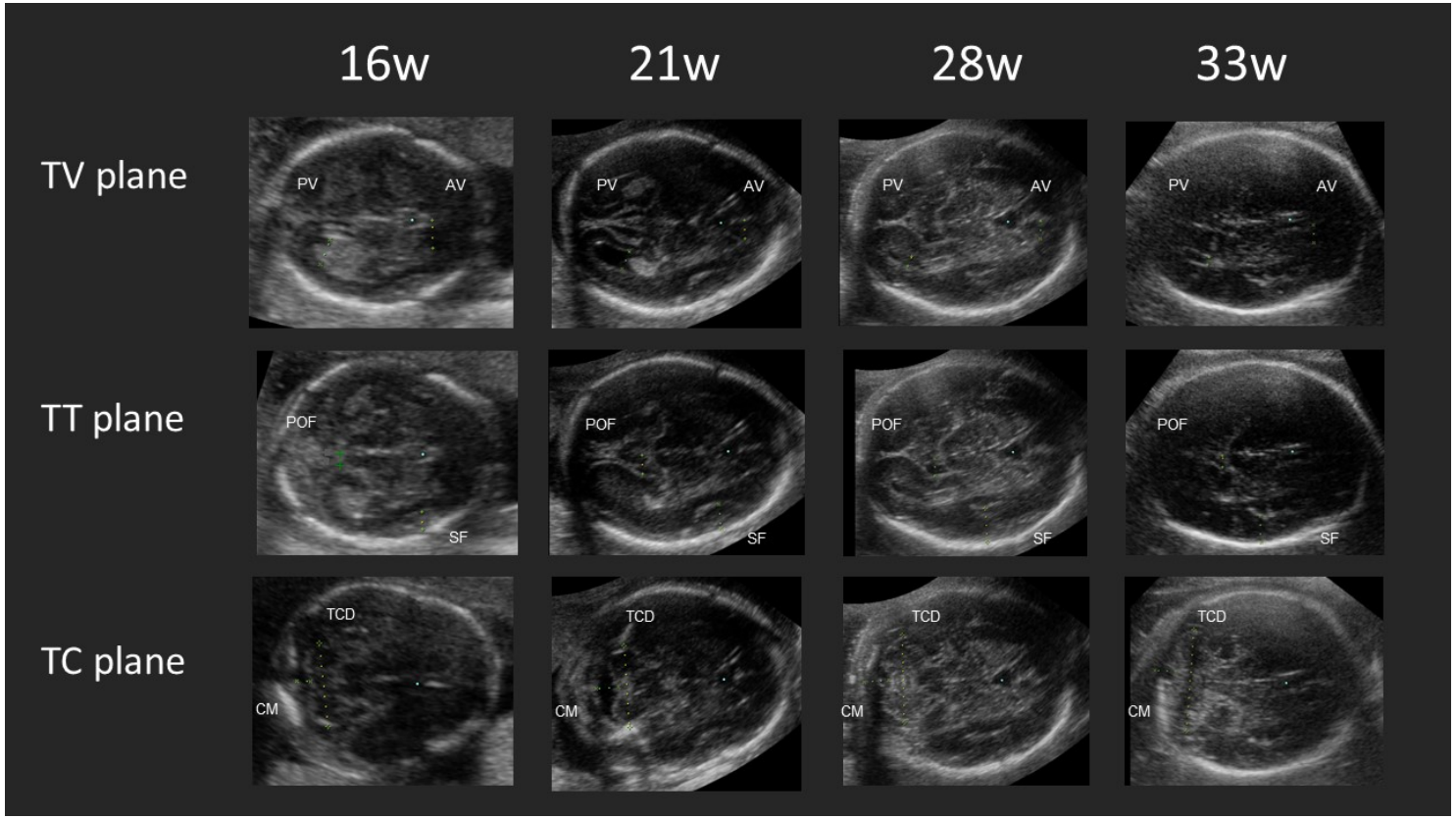


Figure 2: Fetal brain charts study flow chart

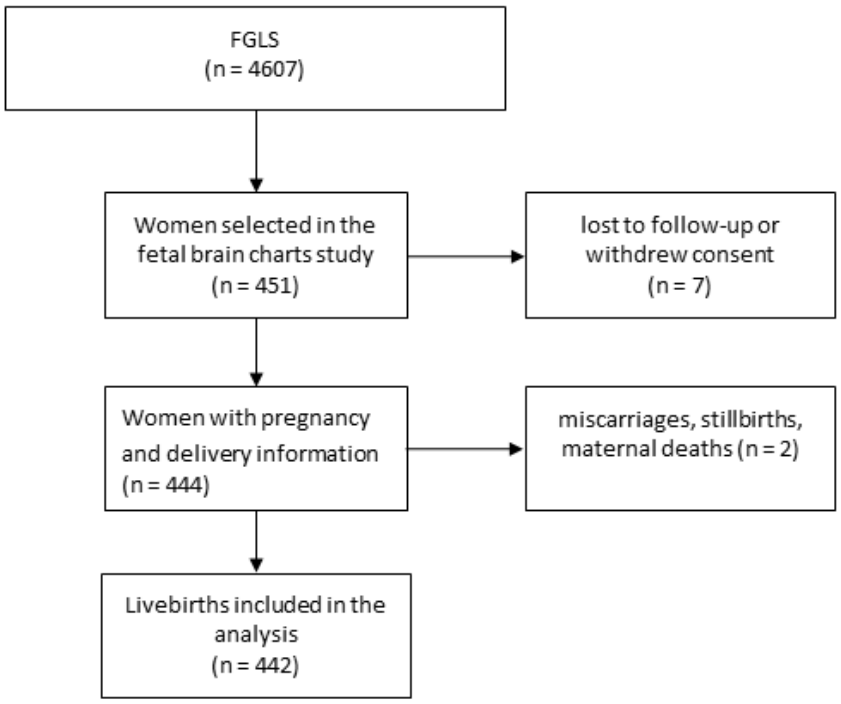


Figure 3A: Fitted 5th, 50th, and 95th smoothed centile curves of parieto-occipital fissure.

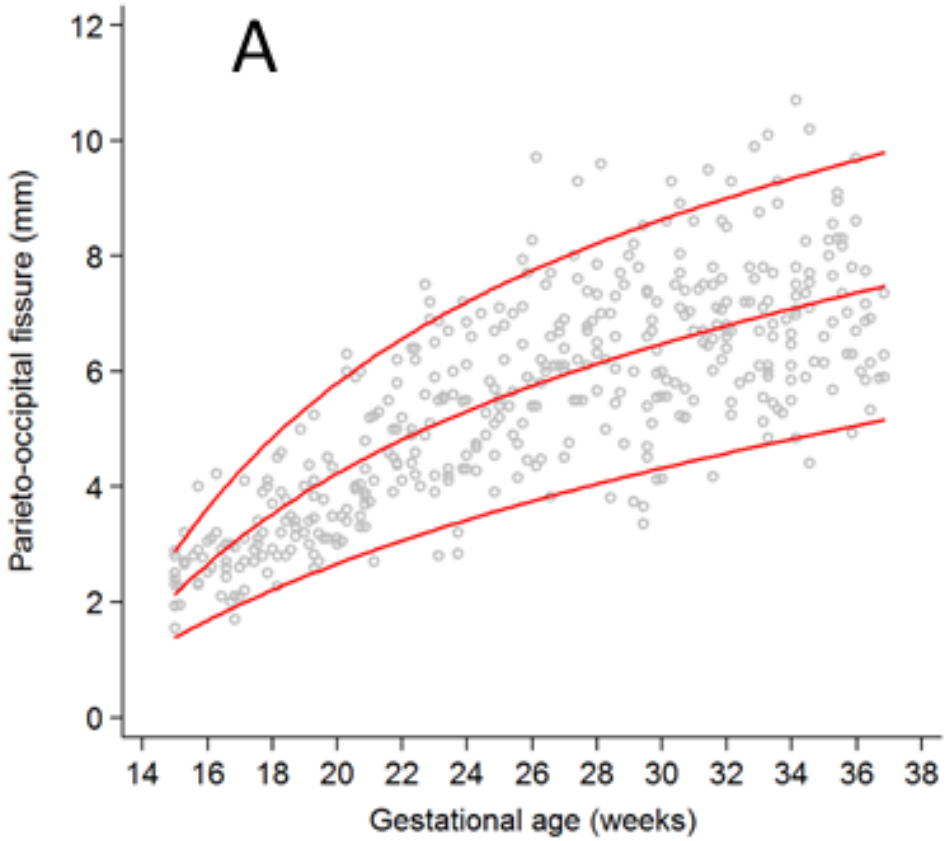


Figure 3B: Fitted 5th, 50th, and 95th smoothed centile curves of sylvian fissure.

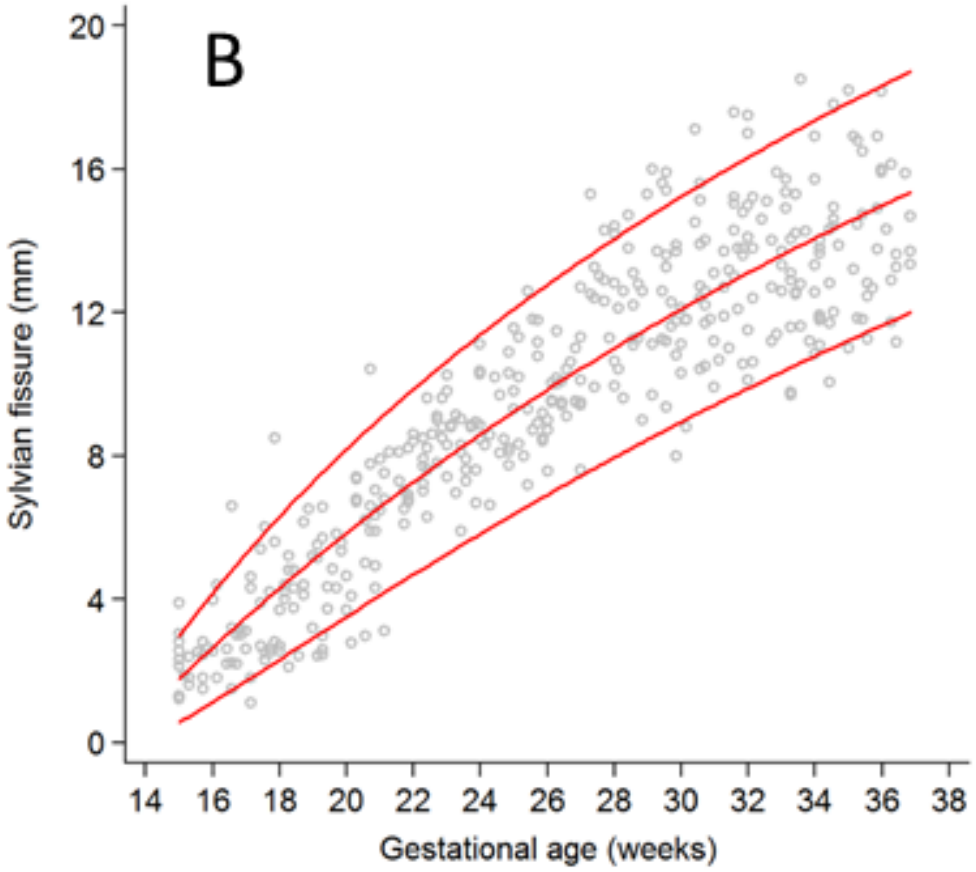


Figure 3C: Fitted 5th, 50th, and 95th smoothed centile curves of anterior ventricle.

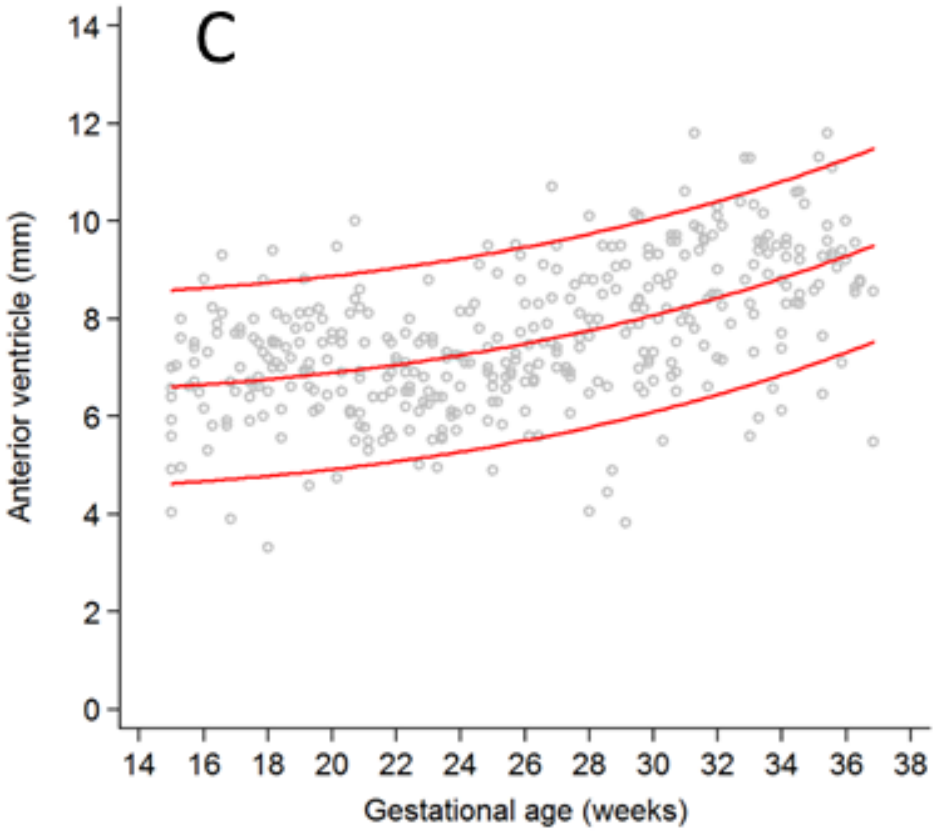


Figure 3D: Fitted 5th, 50th, and 95th smoothed centile curves of posterior ventricle.

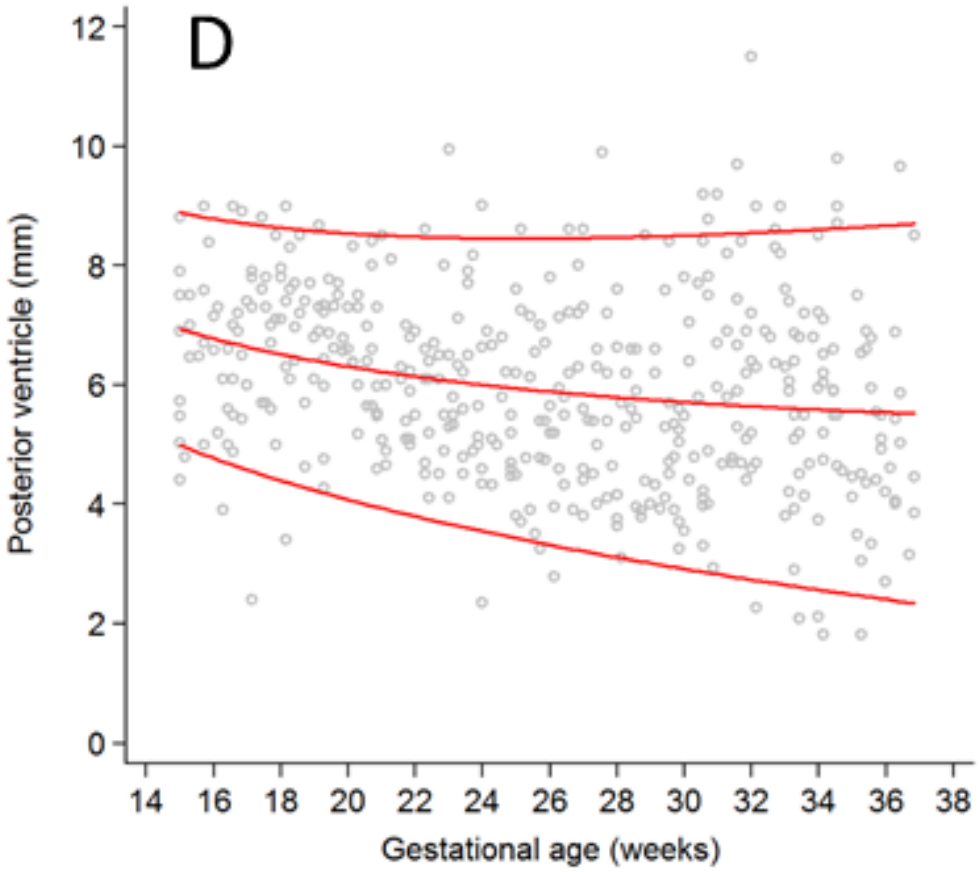


Figure 3E: Fitted 5th, 50th, and 95th smoothed centile curves of transcerebellar diameter.

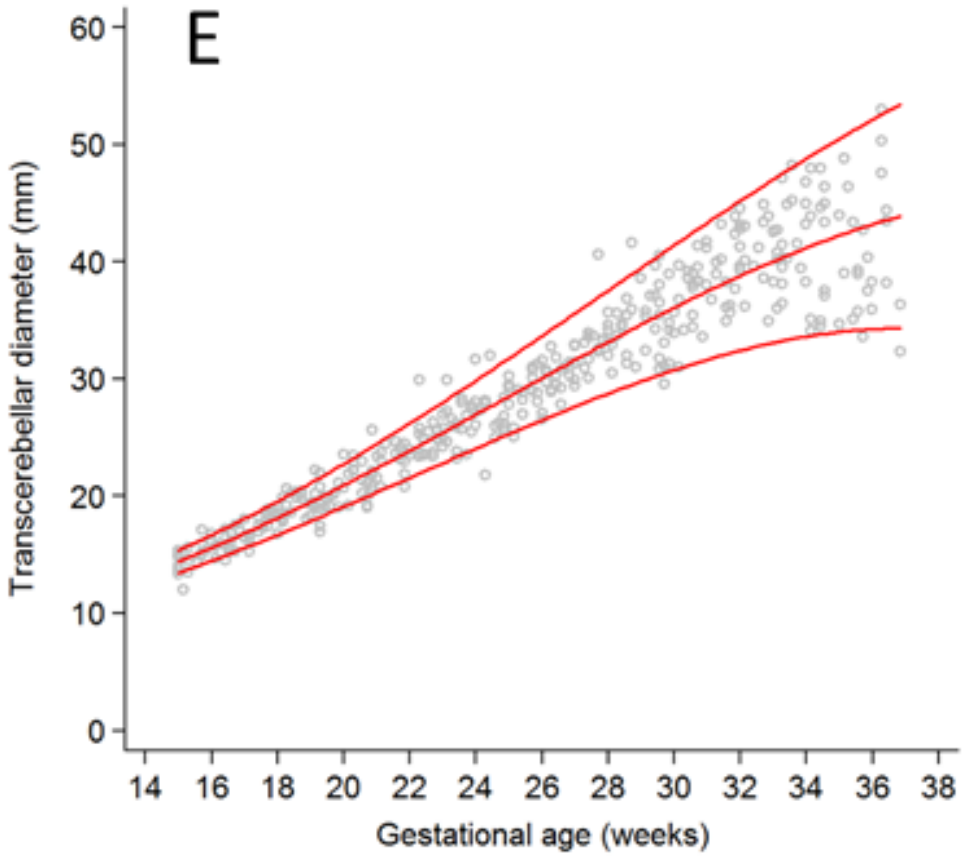
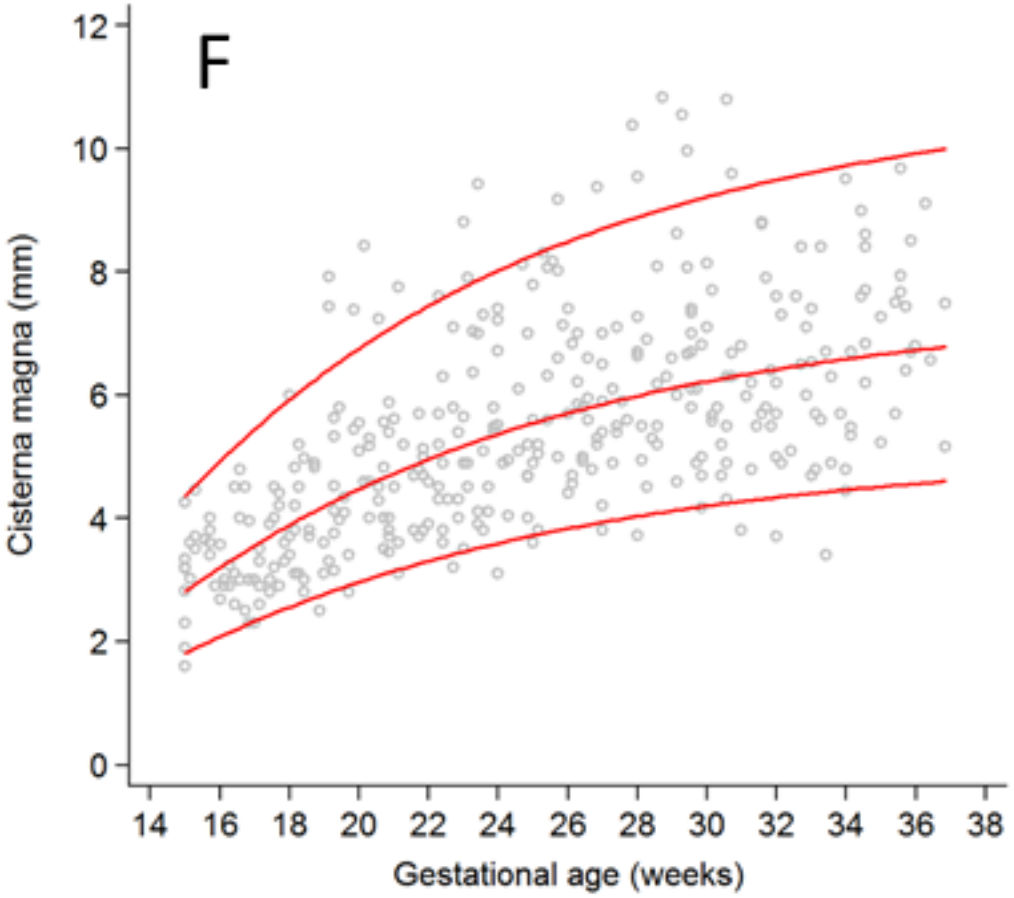
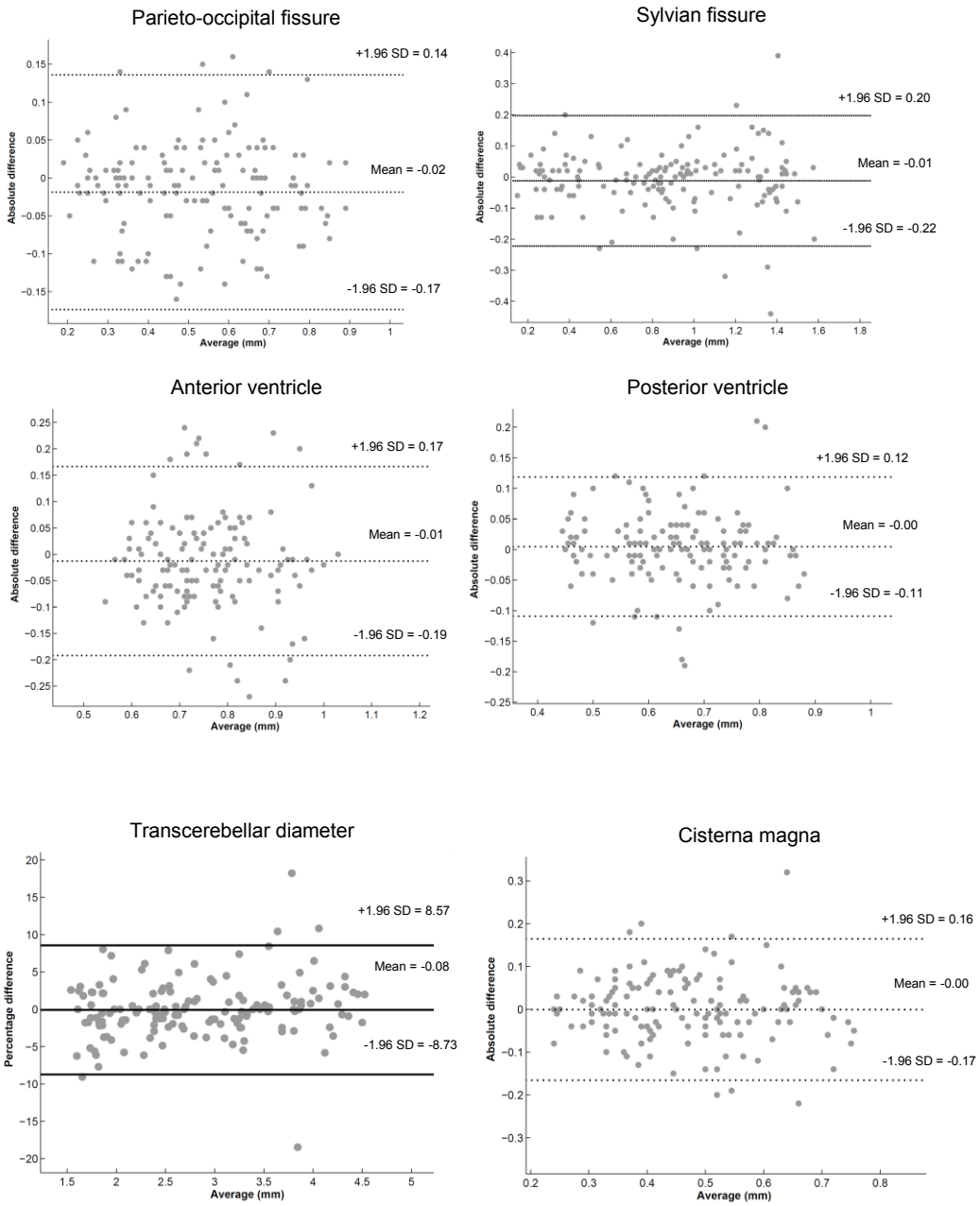


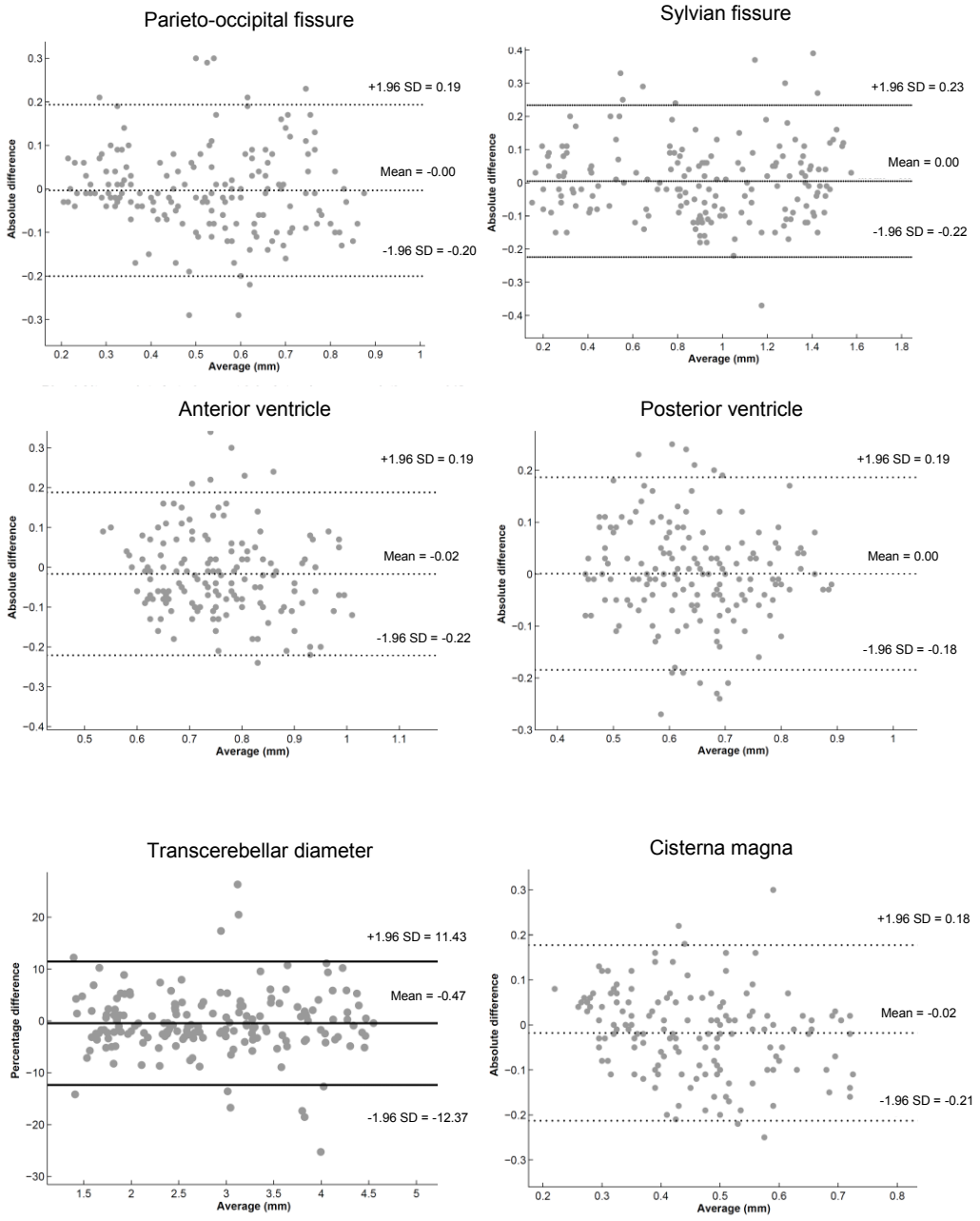
Figure 3F: Fitted 5th, 50th, and 95th smoothed centile curves of cisterna magna.



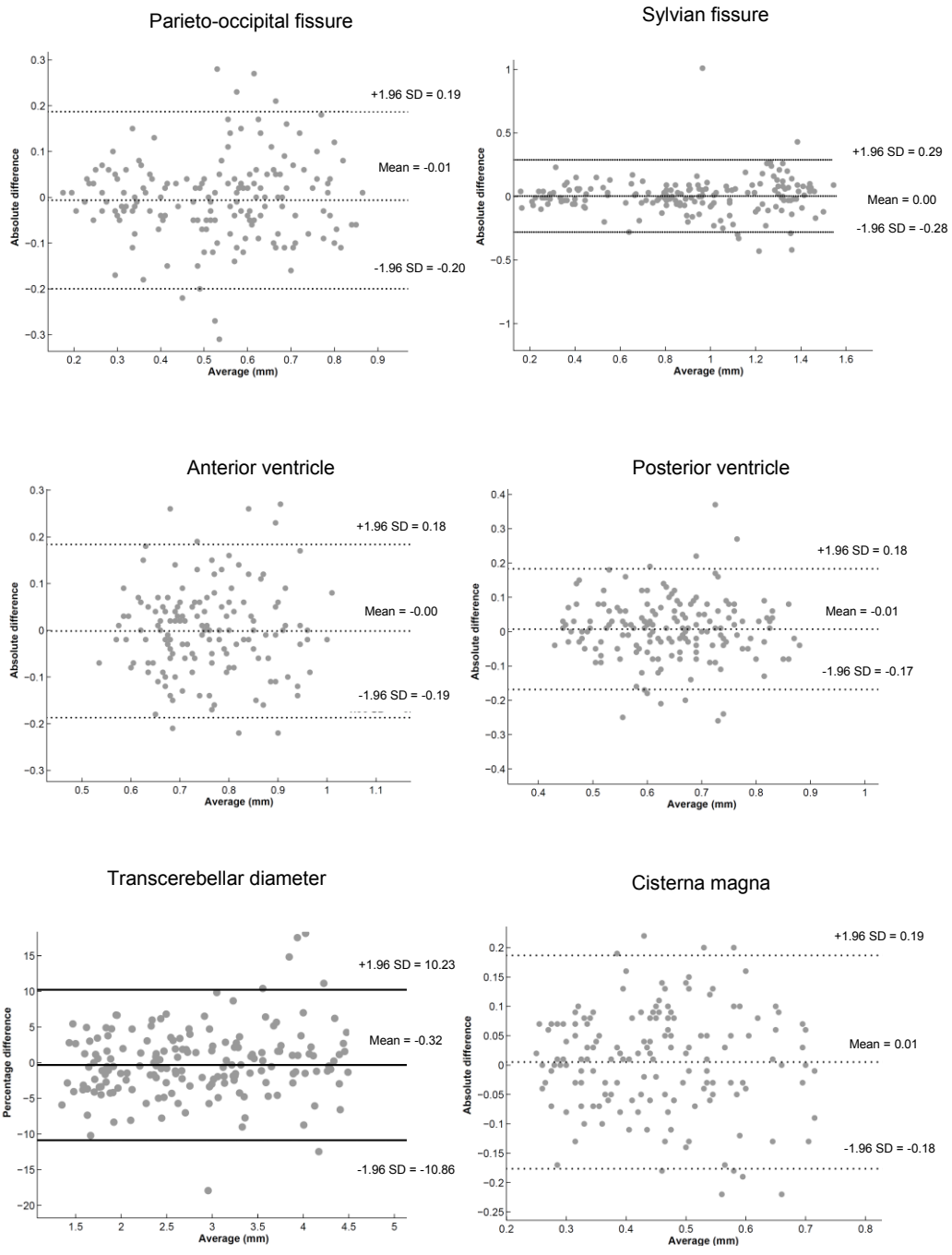
Supplementary Figure 1A: Bland–Altman plots showing intraobserver reproducibility for volume manipulation and caliper placement for measurement acquisition.



Supplementary Figure 1B: Bland–Altman plots showing interobserver reproducibility for volume manipulation and caliper placement for measurement acquisition.



Supplementary Figure 1C: Bland–Altman plots showing interobserver reproducibility for caliper replacement on stored planes.



REFERENCES

1. Salomon LJ, Alfirevic Z, Berghella V, Bilardo C, Hernandez-Andrade E, Johnsen SL, et al. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2011 Jan;37(1):116-26.

2. Malinger G, Zakut H. The corpus callosum: normal fetal development as shown by transvaginal sonography. AJR Am J Roentgenol. 1993 Nov;161(5):1041-3.

3. Paladini D, Volpe P. Posterior fossa and vermian morphometry in the characterization of fetal cerebellar abnormalities: a prospective three-dimensional ultrasound study. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2006 May;27(5):482-9.

4. Napolitano R, Thilaganathan B. Late termination of pregnancy and foetal reduction for foetal anomaly. Best practice & research Clinical obstetrics & gynaecology. 2010 Aug;24(4):529-37.

5. Kuklisova-Murgasova M, Cifor A, Napolitano R, Papageorghiou A, Quaghebeur G, Rutherford MA, et al. Registration of 3D fetal neurosonography and MRI. Medical image analysis. 2013 Dec;17(8):1137-50.

6. Malinger G, Lerman-Sagie T, Watemberg N, Rotmensch S, Lev D, Glezerman M. A normal second-trimester ultrasound does not exclude intracranial structural pathology. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2002 Jul;20(1):51-6.

7. Namburete AI, Stebbing RV, Kemp B, Yaqub M, Papageorghiou AT, Alison Noble J. Learning-based prediction of gestational age from ultrasound images of the fetal brain. Medical image analysis. 2015 Apr;21(1):72-86.

8. Pistorius LR, Stoutenbeek P, Groenendaal F, de Vries L, Manten G, Mulder E, et al. Grade and symmetry of normal fetal cortical development: a longitudinal two- and three-dimensional ultrasound study. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2010 Dec;36(6):700-8.

9. Alves CM, Araujo Junior E, Nardozza LM, Goldman SM, Martinez LH, Martins WP, et al. Reference ranges for fetal brain fissure development on 3-dimensional sonography in the multiplanar mode. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine. 2013 Feb;32(2):269-77.

10. Alonso I, Borenstein M, Grant G, Narbona I, Azumendi G. Depth of brain fissures in normal fetuses by prenatal ultrasound between 19 and 30 weeks of gestation. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2010 Dec;36(6):693-9.

11. Chang CH, Chang FM, Yu CH, Ko HC, Chen HY. Three-dimensional ultrasound in the assessment of fetal cerebellar transverse and antero-posterior diameters. Ultrasound in medicine & biology. 2000 Feb;26(2):175-82.

12. Goldstein I, Reece EA, Pilu G, Bovicelli L, Hobbins JC. Cerebellar measurements with ultrasonography in the evaluation of fetal growth and development. American journal of obstetrics and gynecology. 1987 May;156(5):1065-9.

13. Mittal P, Goncalves LF, Kusanovic JP, Espinoza J, Lee W, Nien JK, et al. Objective evaluation of sylvian fissure development by multiplanar 3-dimensional ultrasonography. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine. 2007 Mar;26(3):347-53.

14. Snijders RJ, Nicolaides KH. Fetal biometry at 14-40 weeks' gestation. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 1994 Jan 01;4(1):34-48.

15. Salomon LJ, Bernard JP, Ville Y. Reference ranges for fetal ventricular width: a non-normal approach. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2007 Jul;30(1):61-6.

16. Almog B, Gamzu R, Achiron R, Fainaru O, Zalel Y. Fetal lateral ventricular width: what should be its upper limit? A prospective cohort study and reanalysis of the current and previous data. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine. 2003 Jan;22(1):39-43.

17. Rossi AC, Prefumo F. Additional value of fetal magnetic resonance imaging in the prenatal diagnosis of central nervous system anomalies: a systematic review of the literature. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2014 Oct;44(4):388-93.

18. Napolitano R, Dhami J, Ohuma EO, Ioannou C, Conde-Agudelo A, Kennedy SH, et al. Pregnancy dating by fetal crown-rump length: a systematic review of charts. BJOG. 2014 Apr;121(5):556-65.

19. Ioannou C, Talbot K, Ohuma E, Sarris I, Villar J, Conde-Agudelo A, et al. Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size. Bjog. 2012;119(12):1425-39.

20. Papageorghiou AT, Kennedy SH, Salomon LJ, Ohuma EO, Cheikh Ismail L, Barros FC, et al. International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown-rump length in the first trimester of pregnancy. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2014 Dec;44(6):641-8.

21. Papageorghiou AT, Ohuma EO, Altman DG, Todros T, Cheikh Ismail L, Lambert A, et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. Lancet. 2014 Sep 6;384(9946):869-79.

22. Villar J, Cheikh Ismail L, Victora CG, Ohuma EO, Bertino E, Altman DG, et al. International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. Lancet. 2014 Sep 6;384(9946):857-68.

23. Papageorghiou AT, Ohuma EO, Gravett MG, Hirst J, da Silveira MF, Lambert A, et al. International standards for symphysis-fundal height based on serial measurements from the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project: prospective cohort study in eight countries. BMJ. 2016 Nov 07;355:i5662.

24. Villar J, Altman DG, Purwar M, Noble JA, Knight HE, Ruyan P, et al. The objectives, design and implementation of the INTERGROWTH-21st Project. BJOG. 2013 Sep;120 Suppl 2:9-26, v.

25. Wanyonyi SZ, Napolitano R, Ohuma EO, Salomon LJ, Papageorghiou AT. Image-scoring system for crown-rump length measurement. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2014 Dec;44(6):649-54.

26. Group WHOMGRS. WHO Motor Development Study: windows of achievement for six gross motor development milestones. Acta Paediatr Suppl. 2006 Apr;450:86-95.

27. Papageorghiou AT, Sarris I, Ioannou C, Todros T, Carvalho M, Pilu G, et al. Ultrasound methodology used to construct the fetal growth standards in the INTERGROWTH-21st Project. BJOG. 2013 Sep;120 Suppl 2:27-32, v.

28. Sarris I, Ioannou C, Ohuma EO, Altman DG, Hoch L, Cosgrove C, et al. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. BJOG. 2013 Sep;120 Suppl 2:33-7, v.

29. Sonographic examination of the fetal central nervous system: guidelines for performing the 'basic examination' and the 'fetal neurosonogram'. Ultrasound in obstetrics

& gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2007 Jan;29(1):109-16.

30. Sarris I, Ohuma E, Ioannou C, Sande J, Altman DG, Papageorghiou AT, et al. Fetal biometry: how well can offline measurements from three-dimensional volumes substitute real-time two-dimensional measurements? Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2013 Nov;42(5):560-70.

31. Ioannou C, Sarris I, Napolitano R, Ohuma E, Javaid MK, Papageorghiou AT. A longitudinal study of normal fetal femur volume. Prenat Diagn. 2013 Nov;33(11):1088-94.

32. Pilu G, Ghi T, Carletti A, Segata M, Perolo A, Rizzo N. Three-dimensional ultrasound examination of the fetal central nervous system. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2007 Aug;30(2):233-45.

33. Napolitano R, Donadono V, Ohuma EO, Knight CL, Wanyonyi SZ, Kemp B, et al. Scientific basis for standardization of fetal head measurements by ultrasound: a reproducibility study. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2016 Jul;48(1):80-5.

34. Chavez MR, Ananth CV, Smulian JC, Lashley S, Kontopoulos EV, Vintzileos AM. Fetal transcerebellar diameter nomogram in singleton gestations with special emphasis in the third trimester: a comparison with previously published nomograms. American journal of obstetrics and gynecology. 2003 Oct;189(4):1021-5.

35. Altman DG, Ohuma EO, International F, Newborn Growth Consortium for the 21st C. Statistical considerations for the development of prescriptive fetal and newborn growth standards in the INTERGROWTH-21st Project. BJOG. 2013 Sep;120 Suppl 2:71-6, v.

36. Quarello E, Stirnemann J, Ville Y, Guibaud L. Assessment of fetal Sylvian fissure operculization between 22 and 32 weeks: a subjective approach. Ultrasound in obstetrics

& gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2008 Jul;32(1):44-9.

37. Timor-Tritsch IE, Monteagudo A. Transvaginal fetal neurosonography: standardization of the planes and sections by anatomic landmarks. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 1996 Jul;8(1):42-7.

38. Ioannou C, Sarris I, Salomon LJ, Papageorghiou AT. A review of fetal volumetry: the need for standardization and definitions in measurement methodology. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2011 Dec;38(6):613-9.

39. Sarris I, Ioannou C, Dighe M, Mitidieri A, Oberto M, Qingqing W, et al. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2011 Dec;38(6):681-7.

40. Bornstein E, Monteagudo A, Santos R, Strock I, Tsymbal T, Lenchner E, et al. Basic as well as detailed neurosonograms can be performed by offline analysis of three-dimensional fetal brain volumes. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2010 Jul;36(1):20-5.

41. Verburg BO, Steegers EA, De Ridder M, Snijders RJ, Smith E, Hofman A, et al. New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2008 Apr;31(4):388-96.

42. Passos AP, Araujo Junior E, Bruns RF, Nardozza LM, Moron AF. Reference ranges of fetal cisterna magna length and area measurements by 3-dimensional ultrasonography using the multiplanar mode. J Child Neurol. 2015 Feb;30(2):209-15.

43. Vinkesteijn AS, Mulder PG, Wladimiroff JW. Fetal transverse cerebellar diameter measurements in normal and reduced fetal growth. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2000 Jan;15(1):47-51.

44. Lei H, Wen SW. Ultrasonographic examination of intrauterine growth for multiple fetal dimensions in a Chinese population. Central-South China Fetal Growth Study Group. American journal of obstetrics and gynecology. 1998 May;178(5):916-21.

45. Serhatlioglu S, Kocakoc E, Kiris A, Sapmaz E, Boztosun Y, Bozgeyik Z. Sonographic measurement of the fetal cerebellum, cisterna magna, and cavum septum pellucidum in normal fetuses in the second and third trimesters of pregnancy. J Clin Ultrasound. 2003 May;31(4):194-200.

46. Goldstein I, Reece EA, Pilu GL, Hobbins JC, Bovicelli L. Sonographic evaluation of the normal developmental anatomy of fetal cerebral ventricles: I. The frontal horn. Obstet Gynecol. 1988 Oct;72(4):588-92.

47. Alagappan R, Browning PD, Laorr A, McGahan JP. Distal lateral ventricular atrium: reevaluation of normal range. Radiology. 1994 Nov;193(2):405-8.

48. Cardoza JD, Goldstein RB, Filly RA. Exclusion of fetal ventriculomegaly with a single measurement: the width of the lateral ventricular atrium. Radiology. 1988 Dec;169(3):711-4.

49. Araujo Junior E, Martins WP, Nardozza LM, Pires CR, Filho SM. Reference range of fetal transverse cerebellar diameter between 18 and 24 weeks of pregnancy in a Brazilian population. J Child Neurol. 2015 Feb;30(2):250-3.

50. Araujo Junior E, Martins WP, Rolo LC, Pires CR, Zanforlin Filho SM. Normative data for fetal cisterna magna length measurement between 18 and 24 weeks of pregnancy. Childs Nerv Syst. 2014 Jan;30(1):9-12.

51. Brown RN. Reassessment of the normal fetal cisterna magna during gestation and an alternative approach to the definition of cisterna magna dilatation. Fetal Diagn Ther. 2013;34(1):44-9.

52. Farrell TA, Hertzberg BS, Kliewer MA, Harris L, Paine SS. Fetal lateral ventricles: reassessment of normal values for atrial diameter at US. Radiology. 1994 Nov;193(2):409-11.

53. Goel P, Singla M, ., Ghai R, Jain S, Budhiraja V, Badu R. Transverse cerebellar diameter – a marker for estimation of gestational age. J Anat Soc India. 2010;59:158-61.

54. Goldstein I, Reece EA, Pilu GL, Hobbins JC. Sonographic evaluation of the normal developmental anatomy of fetal cerebral ventricles. IV.: The posterior horn. Am J Perinatol. 1990 Jan;7(1):79-83.

55. Hilpert PL, Hall BE, Kurtz AB. The atria of the fetal lateral ventricles: a sonographic study of normal atrial size and choroid plexus volume. AJR Am J Roentgenol. 1995 Mar;164(3):731-4.

56. Ishola A, Asaleye CM, Ayoola OO, Loto OM, Idowu BM. Reference Ranges of Fetal Cerebral Lateral Ventricle Parameters by Ultrasonography. Rev Bras Ginecol Obstet. 2016 Sep;38(9):428-35.

57. Joshi BR. Fetal transcerebellar diameter nomogram in Nepalese population. J Inst Med. 2010;32:19-23.

58. Koktener A, Dilmen G, Kurt A. The cisterna magna size in normal second-trimester fetuses. J Perinat Med. 2007;35(3):217-9.

59. Koktener A, Dilmen G, Yildirim M, Kosehan D, Akin K, Cakir B. Growth of the lateral ventricle in normal second-trimester fetuses: Is a nomogram practical? J Pediatr Neuroradiol. 2012;1:37-41.

60. Mahony BS, Callen PW, Filly RA, Hoddick WK. The fetal cisterna magna. Radiology. 1984 Dec;153(3):773-6.

61. McGahan JP, Phillips HE. Ultrasonic evaluation of the size of the trigone of the fetal ventricle. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine. 1983 Jul;2(7):315-9.

62. Peixoto AB, Caldas TM, Barbosa MF, Romao Ldos A, Martins WP, Araujo Junior E. Reference values for the fetal lateral ventricle atrium measurement in the second and third trimesters of pregnancy in a Brazilian population. J Matern Fetal Neonatal Med. 2016;29(14):2337-40.

63. Uerpairojkit B, Charoenvidhya D, Manotaya S, Tanawattanachareon S, Wacharaprechanont T, Tannirandorn Y. Fetal transverse cerebellar diameter in Thai population. J Med Assoc Thai. 2001 Jun;84 Suppl 1:S346-51.

64. Hata K, Hata T, Senoh D, Makihara K, Aoki S, Takamiya O, et al. Ultrasonographic measurement of the fetal transverse cerebellum in utero. Gynecol Obstet Invest. 1989;28(2):111-2.

65. Hayata K, Hiramatsu Y, Masuyama H, Etou E, Nobumoto E, Mitsui T. Creation of a cerebellar diameter reference standard and its clinical application to the detection of cerebellar hypoplasia unique to trisomy 18. J Obstet Gynaecol Res. 2015 Dec;41(12):1899-904.

66. McLeary RD, Kuhns LR, Barr M, Jr. Ultrasonography of the fetal cerebellum. Radiology. 1984 May;151(2):439-42.

67. Smith PA, Johansson D, Tzannatos C, Campbell S. Prenatal measurement of the fetal cerebellum and cisterna cerebellomedullaris by ultrasound. Prenat Diagn. 1986 Mar-Apr;6(2):133-41.

68. Sarris I, Ioannou C, Chamberlain P, Ohuma E, Roseman F, Hoch L, et al. Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2012 Mar;39(3):266-73.

69. Jacquemyn Y, Sys SU, Verdonk P. Fetal transverse cerebellar diameter in different ethnic groups. J Perinat Med. 2000;28(1):14-9.

70. Bai J, Abdul-Rahman MF, Rifkin-Graboi A, Chong YS, Kwek K, Saw SM, et al. Population differences in brain morphology and microstructure among Chinese, Malay, and Indian neonates. PLoS One. 2012;7(10):e47816.

71. Altman DG, Chitty LS. Charts of fetal size: 1. Methodology. British journal of obstetrics and gynaecology. 1994 Jan;101(1):29-34.

72. Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Committee. World Health Organ Tech Rep Ser. 1995;854:1-452.

73. D'Antonio F, Khalil A, Garel C, Pilu G, Rizzo G, Lerman-Sagie T, et al. Systematic review and meta-analysis of isolated posterior fossa malformations on prenatal ultrasound imaging (part 1): nomenclature, diagnostic accuracy and associated anomalies. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2016 Jun;47(6):690-7.

74. Melchiorre K, Bhide A, Gika AD, Pilu G, Papageorghiou AT. Counseling in isolated mild fetal ventriculomegaly. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2009 Aug;34(2):212-24.

CONCLUSION

The main aim of this project was to create international standards for fetal brain structures size using ultrasound measurements.

Fetal brain growth and development is routinely studied using prenatal ultrasound. Ultrasound is used mainly in antenatal care to diagnose fetal abnormalities[6-14] but it is also essential in evaluating fetal growth and central nervous system development. It has been demonstrated how FGR can affect neurodevelopment and therefore a population at low risk of FGR should be selected to create such standards (Appendix 1).[31, 32]

Women recruited in the FGLS of the INTERGROWTH-21st Project represent the ideal candidate. This study uses a 'prescriptive' other than a 'descriptive' design to study the fetal growth, i.e. only children from populations with minimal environmental constraints on growth were included. Previous studies on fetal growth are associated with high risk of bias when a descriptive approach was used (Appendix 2). The population recruited in the fetal structures brain study was representative of the FGLS in view of low incidence of adverse pregnancy outcome and normal motor development.

Ultrasound technology requires the input of several software analysis functions to increase the diagnostic performance, the clinical use and assist in the measurements evaluation when creating standards. A second

line of research associated with this aim has been reported in this manuscript showing promising results of automatic software analysis in terms of accuracy (80-90%) and reproducibility when compared with experts (Appendix 3, 4, 5).

One of the source of high variability between different charts reporting of fetal growth is the absence of a comprehensive quality control strategy in fetal ultrasound.[36-38] The above has been a novel component of the FGLS study[39-42] and several strategies to implement quality control have been studied and demonstrated to be highly reproducible (Appendix 6, 7, 8).

A systematic review of the literature has been performed to identify all the studies aimed to create brain structures charts. There is substantial heterogeneity in the methodology used in previous studies aimed to create brain structures charts.[43-52] and high risk of bias in several domains. Most importantly, no studies reported on long term infant outcome. As a conclusion international standards are required.

To conclude, international standards for six fetal brain structures size measured by antenatal ultrasound have been created with the highest quality methodology and good results of the model fitted. Those standards provide guidelines for ultrasound evaluation of the fetal brain and further understanding into the fetal brain development process in babies at low risk of abnormal neurological outcome (Appendix 9).

REFERENCES

1. Royal College of Obstetricans and Gynaecologists (2013). Green-top guidelines No. 31. The investigation and management of the Small-for-Gestational-Age Fetus. London: Royal College of Obstetricans and Gynaecologists. Available at https://www.rcog.org.uk/en/guidelines-research-services/guidelines/gtg31 (accessed 11/10/2014). .
2. Sonographic examination of the fetal central nervous system: guidelines for performing the 'basic examination' and the 'fetal neurosonogram'. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2007 Jan;29(1):109-16.
3. NICE. Antenatal Care: Routine care for the healthy pregnant woman. London; 2008.
4. American College of O, Gynecologists. ACOG Practice bulletin no. 134: fetal growth restriction. Obstetrics and gynecology. 2013 May;121(5):1122-33.
5. Salomon LJ, Alfirevic Z, Berghella V, Bilardo C, Hernandez-Andrade E, Johnsen SL, et al. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2011 Jan;37(1):116-26.
6. Di Fraja D, Sarno L, Migliucci A, Acampora E, Napolitano R, Maruotti GM, et al. Prenatal diagnosis of beta-thalassemia: nuchal translucency in affected fetuses. Minerva Ginecol. 2011 Dec;63(6):491-4.
7. Mallia Milanes G, Napolitano R, Quaglia F, Mazzarelli LL, Agangi A, Tessitore G, et al. Prenatal diagnosis of arthrogryposis. Minerva Ginecol. 2007 Apr;59(2):203-4.
8. Maruotti GM, Agangi A, Napolitano R, Mazzarelli LL, Quaglia F, Carbone IF, et al. Prenatal diagnosis of Ulbright-Hodes syndrome. J Ultrasound Med. 2009 Mar;28(3):385-8.
9. Maruotti GM, Fabbrini F, Napolitano R, Genesio R, Conti A, Mallia Milanes G, et al. Trisomy 18 caused by isochromosome 18p and 18q formation: Is there a milder phenotype? Am J Med Genet A. 2011 Jan;155A(1):225-7.
10. Maruotti GM, Paladini D, Napolitano R, Mazzarelli LL, Russo T, Quarantelli M, et al. Prenatal 2D and 3D ultrasound diagnosis of diprosopus: case report with post-mortem magnetic resonance images (MRI) and review of the literature. Prenat Diagn. 2009 Oct;29(10):992-4.
11. Migliucci A, Di Fraja D, Sarno L, Acampora E, Mazzarelli LL, Quaglia F, et al. Prenatal diagnosis of congenital rubella infection and ultrasonography: a preliminary study. Minerva Ginecol. 2011 Dec;63(6):485-9.
12. Napolitano R, Maruotti GM, Quarantelli M, Martinelli P, Paladini D. Prenatal diagnosis of Seckel Syndrome on 3-dimensional sonography and magnetic resonance imaging. J Ultrasound Med. 2009 Mar;28(3):369-74.
13. Napolitano R, Thilaganathan B. Late termination of pregnancy and foetal reduction for foetal anomaly. Best practice & research Clinical obstetrics & gynaecology. 2010 Aug;24(4):529-37.

14. Simioli S, Napolitano R, Quaglia F, Mazzarelli LL, Agangi A, Milanes GM, et al. Fetal borderline cerebral ventriculomegaly: clinical significance and management. Minerva Ginecol. 2009 Apr;61(2):109-12.
15. Dias T, Arcangeli T, Bhide A, Napolitano R, Mahsud-Dornan S, Thilaganathan B. First-trimester ultrasound determination of chorionicity in twin pregnancy. Ultrasound Obstet Gynecol. 2011 Nov;38(5):530-2.
16. Martinelli P, Agangi A, Sansone M, Napolitano R, Maruotti GM. An important clinical lesson from a patient infected with HIV with diabetic nephropathy and retinopathy. BMJ Case Rep. 2009.
17. Maruotti GM, Anfora R, Scanni E, Rispoli M, Mazzarelli LL, Napolitano R, et al. Anesthetic management of a parturient with spinal muscular atrophy type II. J Clin Anesth. 2012 Nov;24(7):573-7.
18. Maruotti GM, Sarno L, Napolitano R, Mazzarelli LL, Quaglia F, Capone A, et al. Preeclampsia in women with chronic kidney disease. J Matern Fetal Neonatal Med. 2012 Aug;25(8):1367-9.
19. Morlando M, Sarno L, Napolitano R, Capone A, Tessitore G, Maruotti GM, et al. Placenta accreta: incidence and risk factors in an area with a particularly high rate of cesarean section. Acta Obstet Gynecol Scand. 2013 Apr;92(4):457-60.
20. Napolitano R, Campanile A, Sarno L, Anastasio A, Maruotti GM, Morlando M, et al. GRK2 levels in umbilical arteries of pregnancies complicated by gestational hypertension and preeclampsia. Am J Hypertens. 2012 Mar;25(3):366-71.
21. Napolitano R, Maruotti GM, Mazzarelli LL, Quaglia F, Tessitore G, Pecoraro M, et al. Prenatal diagnosis of placental chorioangioma: our experience. Minerva Ginecol. 2005 Dec;57(6):649-54.
22. Napolitano R, Melchiorre K, Arcangeli T, Dias T, Bhide A, Thilaganathan B. Screening for pre-eclampsia by using changes in uterine artery Doppler indices with advancing gestation. Prenat Diagn. 2012 Feb;32(2):180-4.
23. Napolitano R, Rajakulasingam R, Memmo A, Bhide A, Thilaganathan B. Uterine artery Doppler screening for pre-eclampsia: comparison of the lower, mean and higher first-trimester pulsatility indices. Ultrasound Obstet Gynecol. 2011 May;37(5):534-7.
24. Napolitano R, Santo S, D'Souza R, Bhide A, Thilaganathan B. Sensitivity of higher, lower and mean second-trimester uterine artery Doppler resistance indices in screening for pre-eclampsia. Ultrasound Obstet Gynecol. 2010 Nov;36(5):573-6.
25. Napolitano R, Thilaganathan B. Mean, lowest, and highest pulsatility index of the uterine artery and adverse pregnancy outcome in twin pregnancies. Am J Obstet Gynecol. 2012 Jun;206(6):e8-9.
26. Napolitano R, Thilaganathan B. Comment on "first trimester uterine artery Doppler velocimetry in the prediction of birth weight in a low-risk population". Prenat Diagn. 2013 Dec;33(13):1317.
27. Villar J, Altman DG, Purwar M, Noble JA, Knight HE, Ruyan P, et al. The objectives, design and implementation of the INTERGROWTH-21st Project. BJOG : an international journal of obstetrics and gynaecology. 2013 Sep;120 Suppl 2:9-26.
28. Meher S, Hernandez-Andrade E, Basheer SN, Lees C. Impact of cerebral redistribution on neurodevelopmental outcome in small-for-gestational-age or growth-restricted babies: a systematic review. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2015 Oct;46(4):398-404.
29. Guellec I, Lapillonne A, Marret S, Picaud JC, Mitanchez D, Charkaluk ML, et al. Effect of Intra- and Extrauterine Growth on Long-Term Neurologic Outcomes of Very Preterm Infants. J Pediatr. 2016 Aug;175:93-9 e1.

30. Takeuchi A, Yorifuji T, Takahashi K, Nakamura M, Kageyama M, Kubo T, et al. Behavioral outcomes of school-aged full-term small-for-gestational-age infants: A nationwide Japanese population-based study. Brain Dev. 2017 Feb;39(2):101-6.
31. Lees C, Marlow N, Arabin B, Bilardo CM, Brezinka C, Derks JB, et al. Perinatal morbidity and mortality in early-onset fetal growth restriction: cohort outcomes of the trial of randomized umbilical and fetal flow in Europe (TRUFFLE). Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2013 Oct;42(4):400-8.
32. Lees CC, Marlow N, van Wassenaer-Leemhuis A, Arabin B, Bilardo CM, Brezinka C, et al. 2 year neurodevelopmental and intermediate perinatal outcomes in infants with very preterm fetal growth restriction (TRUFFLE): a randomised trial. Lancet. 2015 May 30;385(9983):2162-72.
33. Boers KE, Vijgen SM, Bijlenga D, van der Post JA, Bekedam DJ, Kwee A, et al. Induction versus expectant monitoring for intrauterine growth restriction at term: randomised equivalence trial (DIGITAT). Bmj. 2010;341:c7087.
34. Abe S, Takagi K, Yamamoto T, Kato T. Assessment of cortical gyrus and sulcus formation using magnetic resonance images in small-for-gestational-age fetuses. Prenat Diagn. 2004 May;24(5):333-8.
35. Papageorghiou AT, Ohuma EO, Altman DG, Todros T, Cheikh Ismail L, Lambert A, et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. Lancet. 2014 Sep 6;384(9946):869-79.
36. Ioannou C, Talbot K, Ohuma E, Sarris I, Villar J, Conde-Agudelo A, et al. Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size. BJOG : an international journal of obstetrics and gynaecology. 2012;119(12):1425-39.
37. Ioannou C, Sarris I, Salomon LJ, Papageorghiou AT. A review of fetal volumetry: the need for standardization and definitions in measurement methodology. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2011 Dec;38(6):613-9.
38. Napolitano R, Dhami J, Ohuma EO, Ioannou C, Conde-Agudelo A, Kennedy SH, et al. Pregnancy dating by fetal crown-rump length: a systematic review of charts. BJOG : an international journal of obstetrics and gynaecology. 2014 Apr;121(5):556-65.
39. Papageorghiou AT, Sarris I, Ioannou C, Todros T, Carvalho M, Pilu G, et al. Ultrasound methodology used to construct the fetal growth standards in the INTERGROWTH-21st Project. BJOG : an international journal of obstetrics and gynaecology. 2013 Sep;120 Suppl 2:27-32.
40. Sarris I, Ioannou C, Chamberlain P, Ohuma E, Roseman F, Hoch L, et al. Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2012 Mar;39(3):266-73.
41. Sarris I, Ioannou C, Dighe M, Mitidieri A, Oberto M, Qingqing W, et al. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2011 Dec;38(6):681-7.
42. Sarris I, Ioannou C, Ohuma E, Altman D, Hoch L, Cosgrove C, et al. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21 Project. BJOG : an international journal of obstetrics and gynaecology. 2013 Sep;120 Suppl 2:33-7.

43. Almog B, Gamzu R, Achiron R, Fainaru O, Zalel Y. Fetal lateral ventricular width: what should be its upper limit? A prospective cohort study and reanalysis of the current and previous data. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine. 2003 Jan;22(1):39-43.
44. Alonso I, Borenstein M, Grant G, Narbona I, Azumendi G. Depth of brain fissures in normal fetuses by prenatal ultrasound between 19 and 30 weeks of gestation. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2010 Dec;36(6):693-9.
45. Alves CM, Araujo Junior E, Nardozza LM, Goldman SM, Martinez LH, Martins WP, et al. Reference ranges for fetal brain fissure development on 3-dimensional sonography in the multiplanar mode. J Ultrasound Med. 2013 Feb;32(2):269-77.
46. Chang CH, Chang FM, Yu CH, Ko HC, Chen HY. Three-dimensional ultrasound in the assessment of fetal cerebellar transverse and antero-posterior diameters. Ultrasound in medicine & biology. 2000 Feb;26(2):175-82.
47. Chavez MR, Ananth CV, Smulian JC, Lashley S, Kontopoulos EV, Vintzileos AM. Fetal transcerebellar diameter nomogram in singleton gestations with special emphasis in the third trimester: a comparison with previously published nomograms. American journal of obstetrics and gynecology. 2003 Oct;189(4):1021-5.
48. Goldstein I, Reece EA, Pilu G, Bovicelli L, Hobbins JC. Cerebellar measurements with ultrasonography in the evaluation of fetal growth and development. Am J Obstet Gynecol. 1987 May;156(5):1065-9.
49. Mittal P, Goncalves LF, Kusanovic JP, Espinoza J, Lee W, Nien JK, et al. Objective evaluation of sylvian fissure development by multiplanar 3-dimensional ultrasonography. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine. 2007 Mar;26(3):347-53.
50. Salomon LJ, Bernard JP, Ville Y. Reference ranges for fetal ventricular width: a non-normal approach. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2007 Jul;30(1):61-6.
51. Snijders RJ, Nicolaides KH. Fetal biometry at 14-40 weeks' gestation. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 1994 Jan 01;4(1):34-48.
52. Verburg BO, Steegers EA, De Ridder M, Snijders RJ, Smith E, Hofman A, et al. New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2008 Apr;31(4):388-96.
53. Black RE, Allen LH, Bhutta ZA, Caulfield LE, de Onis M, Ezzati M, et al. Maternal and child undernutrition: global and regional exposures and health consequences. Lancet. 2008 Jan 19;371(9608):243-60.
54. Wixey JA, Chand KK, Colditz PB, Bjorkman ST. Neuroinflammation in intrauterine growth restriction. Placenta. 2016 Nov 25.
55. Bricker L, Neilson JP, Dowswell T. Routine ultrasound in late pregnancy (after 24 weeks' gestation). The Cochrane database of systematic reviews. 2008(4):CD001451.
56. Arcangeli T, Thilaganathan B, Hooper R, Khan KS, Bhide A. Neurodevelopmental delay in small babies at term: a systematic review. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2012 Sep;40(3):267-75.
57. Murray E, Fernandes M, Fazel M, Kennedy SH, Villar J, Stein A. Differential effect of intrauterine growth restriction on childhood neurodevelopment: a systematic review. BJOG : an international journal of obstetrics and gynaecology. 2015 Jul;122(8):1062-72.

58. Faa G, Manchia M, Pintus R, Gerosa C, Marcialis MA, Fanos V. Fetal programming of neuropsychiatric disorders. Birth Defects Res C Embryo Today. 2016 Sep;108(3):207-23.
59. Barker DJ. The developmental origins of well-being. Philos Trans R Soc Lond B Biol Sci. 2004 Sep 29;359(1449):1359-66.
60. Visser GH, Bilardo CM, Derks JB, Ferrazzi E, Fratelli N, Frusca T, et al. The TRUFFLE study; fetal monitoring indications for delivery in 310 IUGR infants with 2 year's outcome delivered before 32 weeks of gestation. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2016 Nov 11.
61. Uauy R, Casanello P, Krause B, Kuzanovic JP, Corvalan C, (INTERGROWTH-21st) ftIFaNGCftsC. Conceptual basis for prescriptive growth standards from conception to early childhood: present and future. BJOG : an international journal of obstetrics and gynaecology. 2013;120 Suppl 2:3-8.
62. Villar J, Papageorghiou AT, Pang R, Salomon LJ, Langer A, Victora C, et al. Monitoring human growth and development: a continuum from the womb to the classroom. Am J Obstet Gynecol. 2015 Oct;213(4):494-9.
63. 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 Feb 01;151(3):333-7.
64. Giuliani F, Ohuma E, Spada E, Bertino E, Al Dhaheri AS, Altman DG, et al. Systematic review of the methodological quality of studies designed to create neonatal anthropometric charts. Acta Paediatr. 2015 Oct;104(10):987-96.
65. Giuliani F, Cheikh Ismail L, Bertino E, Bhutta ZA, Ohuma EO, Rovelli I, et al. Monitoring postnatal growth of preterm infants: present and future. Am J Clin Nutr. 2016 Feb;103(2):635S-47S.
66. De Onis M, Garza C, Onyango AW, Martorell R. WHO child growth standards. Acta Paediatr Suppl. 2006;450:1-101.
67. Papageorghiou AT, Kennedy SH, Salomon LJ, Ohuma EO, Cheikh Ismail L, Barros FC, et al. International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown-rump length in the first trimester of pregnancy. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2014 Dec;44(6):641-8.
68. Papageorghiou AT, Kemp B, Stones W, Ohuma EO, Kennedy SH, Purwar M, et al. Ultrasound-based gestational-age estimation in late pregnancy. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2016 Dec;48(6):719-26.
69. Stirnemann J, Villar J, Salomon LJ, Ohuma E, Ruyan P, Altman DG, et al. International Estimated Fetal Weight Standards of the INTERGROWTH-21st Project. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2016 Nov 02.
70. Papageorghiou AT, Ohuma EO, Gravett MG, Hirst J, da Silveira MF, Lambert A, et al. International standards for symphysis-fundal height based on serial measurements from the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project: prospective cohort study in eight countries. Bmj. 2016 Nov 07;355:i5662.
71. Cheikh Ismail L, Bishop DC, Pang R, Ohuma EO, Kac G, Abrams B, et al. Gestational weight gain standards based on women enrolled in the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project: a prospective longitudinal cohort study. Bmj. 2016 Feb 29;352:i555.

72. Villar J, Cheikh Ismail L, Victora CG, Ohuma EO, Bertino E, Altman DG, et al. International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. Lancet. 2014 Sep 6;384(9946):857-68.
73. Villar J, Giuliani F, Fenton TR, Ohuma EO, Ismail LC, Kennedy SH, et al. INTERGROWTH-21st very preterm size at birth reference charts. Lancet. 2016 Feb 27;387(10021):844-5.
74. Fernandes M, Stein A, Newton CR, Cheikh-Ismail L, Kihara M, Wulff K, et al. The INTERGROWTH-21st Project Neurodevelopment Package: a novel method for the multi-dimensional assessment of neurodevelopment in pre-school age children. PLoS One. 2014;9(11):e113360.
75. Napolitano R, Donadono V, Ohuma EO, Knight CL, Wanyonyi SZ, Kemp B, et al. Scientific basis for standardization of fetal head measurements by ultrasound: a reproducibility study. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2016 Jul;48(1):80-5.
76. Napolitano R, Ghosh M, Gillott DJ, Ojha K. Three-dimensional Doppler sonography in asymptomatic and symptomatic women after medical termination of pregnancy. J Ultrasound Med. 2014 May;33(5):847-52.
77. Namburete AI, Stebbing RV, Kemp B, Yaqub M, Papageorghiou AT, Alison Noble J. Learning-based prediction of gestational age from ultrasound images of the fetal brain. Medical image analysis. 2015 Apr;21(1):72-86.
78. Maraci MA, Napolitano R, Papageorghiou A, Noble JA. Fisher vector encoding for detecting objects of interest in ultrasound videos. . IEEE 12th International Symposium on Biomedical Imaging (ISBI). 2015:651-4.
79. Waechter-Stehle I, Klinder T, Rouet JM, Roundhill D, Andrews G, Cavallaro A, et al. Learning from redundant but inconsistent reference data: Anatomical views and measurements for fetal brain screening. SPIE Medical Imaging. 2016;9784.
80. Altman DG, Chitty LS. Charts of fetal size: 1. Methodology. British journal of obstetrics and gynaecology. 1994 Jan;101(1):29-34.
81. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb 8;1(8476):307-10.
82. Salomon LJ, Bernard JP, Duyme M, Doris B, Mas N, Ville Y. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2006 Jan;27(1):34-40.
83. Wanyonyi SZ, Napolitano R, Ohuma EO, Salomon LJ, Papageorghiou AT. Image-scoring system for crown-rump length measurement. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2014 Dec;44(6):649-54.
84. Ioannou C, Sarris I, Hoch L, Salomon LJ, Papageorghiou AT. Standardisation of crown-rump length measurement. BJOG : an international journal of obstetrics and gynaecology. 2013 Sep;120 Suppl 2:38-41.
85. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. Journal of clinical epidemiology. 2009 Aug;62(8):797-806.
86. Quarello E, Stirnemann J, Ville Y, Guibaud L. Assessment of fetal Sylvian fissure operculization between 22 and 32 weeks: a subjective approach. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2008 Jul;32(1):44-9.

87. Passos AP, Araujo Junior E, Bruns RF, Nardozza LM, Moron AF. Reference ranges of fetal cisterna magna length and area measurements by 3-dimensional ultrasonography using the multiplanar mode. J Child Neurol. 2015 Feb;30(2):209-15.
88. Vinkesteijn AS, Mulder PG, Wladimiroff JW. Fetal transverse cerebellar diameter measurements in normal and reduced fetal growth. Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology. 2000 Jan;15(1):47-51.
89. Lei H, Wen SW. Ultrasonographic examination of intrauterine growth for multiple fetal dimensions in a Chinese population. Central-South China Fetal Growth Study Group. Am J Obstet Gynecol. 1998 May;178(5):916-21.
90. Serhatlioglu S, Kocakoc E, Kiris A, Sapmaz E, Boztosun Y, Bozgeyik Z. Sonographic measurement of the fetal cerebellum, cisterna magna, and cavum septum pellucidum in normal fetuses in the second and third trimesters of pregnancy. J Clin Ultrasound. 2003 May;31(4):194-200.
91. Goldstein I, Reece EA, Pilu GL, Hobbins JC, Bovicelli L. Sonographic evaluation of the normal developmental anatomy of fetal cerebral ventricles: I. The frontal horn. Obstetrics and gynecology. 1988 Oct;72(4):588-92.
92. Alagappan R, Browning PD, Laorr A, McGahan JP. Distal lateral ventricular atrium: reevaluation of normal range. Radiology. 1994 Nov;193(2):405-8.
93. Cardoza JD, Goldstein RB, Filly RA. Exclusion of fetal ventriculomegaly with a single measurement: the width of the lateral ventricular atrium. Radiology. 1988 Dec;169(3):711-4.
94. Hata K, Hata T, Senoh D, Makihara K, Aoki S, Takamiya O, et al. Ultrasonographic measurement of the fetal transverse cerebellum in utero. Gynecol Obstet Invest. 1989;28(2):111-2.
95. Hayata K, Hiramatsu Y, Masuyama H, Etou E, Nobumoto E, Mitsui T. Creation of a cerebellar diameter reference standard and its clinical application to the detection of cerebellar hypoplasia unique to trisomy 18. J Obstet Gynaecol Res. 2015 Dec;41(12):1899-904.
96. Joshi BR. Fetal transcerebellar diameter nomogram in Nepalese population. J Inst Med. 2010;32:19-23.
97. Koktener A, Dilmen G, Kurt A. The cisterna magna size in normal second-trimester fetuses. Journal of perinatal medicine. 2007;35(3):217-9.
98. Koktener A, Dilmen G, Yildirim M, Kosehan D, Akin K, Cakir B. Growth of the lateral ventricle in normal second-trimester fetuses: Is a nomogram practical? J Pediatr Neuroradiol. 2012;1:37-41.
99. Mahony BS, Callen PW, Filly RA, Hoddick WK. The fetal cisterna magna. Radiology. 1984 Dec;153(3):773-6.
100. McGahan JP, Phillips HE. Ultrasonic evaluation of the size of the trigone of the fetal ventricle. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine. 1983 Jul;2(7):315-9.
101. McLeary RD, Kuhns LR, Barr M, Jr. Ultrasonography of the fetal cerebellum. Radiology. 1984 May;151(2):439-42.
102. Ishola A, Asaleye CM, Ayoola OO, Loto OM, Idowu BM. Reference Ranges of Fetal Cerebral Lateral Ventricle Parameters by Ultrasonography. Rev Bras Ginecol Obstet. 2016 Sep;38(9):428-35.
103. Araujo Junior E, Martins WP, Nardozza LM, Pires CR, Filho SM. Reference range of fetal transverse cerebellar diameter between 18 and 24 weeks of pregnancy in a Brazilian population. J Child Neurol. 2015 Feb;30(2):250-3.

104. Araujo Junior E, Martins WP, Rolo LC, Pires CR, Zanforlin Filho SM. Normative data for fetal cisterna magna length measurement between 18 and 24 weeks of pregnancy. Childs Nerv Syst. 2014 Jan;30(1):9-12.
105. Brown RN. Reassessment of the normal fetal cisterna magna during gestation and an alternative approach to the definition of cisterna magna dilatation. Fetal diagnosis and therapy. 2013;34(1):44-9.
106. Smith PA, Johansson D, Tzannatos C, Campbell S. Prenatal measurement of the fetal cerebellum and cisterna cerebellomedullaris by ultrasound. Prenat Diagn. 1986 Mar-Apr;6(2):133-41.
107. Goel P, Singla M, ., Ghai R, Jain S, Budhiraja V, Badu R. Transverse cerebellar diameter – a marker for estimation of gestational age. J Anat Soc India. 2010;59:158-61.
108. Goldstein I, Reece EA, Pilu GL, Hobbins JC. Sonographic evaluation of the normal developmental anatomy of fetal cerebral ventricles. IV.: The posterior horn. Am J Perinatol. 1990 Jan;7(1):79-83.
109. Farrell TA, Hertzberg BS, Kliewer MA, Harris L, Paine SS. Fetal lateral ventricles: reassessment of normal values for atrial diameter at US. Radiology. 1994 Nov;193(2):409-11.
110. Uerpairojkit B, Charoenvidhya D, Manotaya S, Tanawattanachareon S, Wacharaprechanont T, Tannirandorn Y. Fetal transverse cerebellar diameter in Thai population. Journal of the Medical Association of Thailand. 2001 Jun;84 Suppl 1:S346-51.
111. Hilpert PL, Hall BE, Kurtz AB. The atria of the fetal lateral ventricles: a sonographic study of normal atrial size and choroid plexus volume. AJR Am J Roentgenol. 1995 Mar;164(3):731-4.
112. Peixoto AB, Caldas TM, Barbosa MF, Romao Ldos A, Martins WP, Araujo Junior E. Reference values for the fetal lateral ventricle atrium measurement in the second and third trimesters of pregnancy in a Brazilian population. The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstet. 2016;29(14):2337-40.

ACKNOWLEDGEMENT

Research projects involved in the this manuscript have been conducted at the Nuffield Department of Obstetrics and Gynaecology, University of Oxford, under the supervision of Professor Aris T. Papageorghiou.

I would like to thank Prof. Papageorghiou for his support. He contributed a great deal to my knowledge and understanding of fetal medicine, human growth and research methodology. I would like also to express my gratitude to all the members of the OMPHI office and of the INTERGROWTH-21st Project who I had the privilege to work with. Particularly I would like to mention Prof. S.K. Kennedy and Prof. J. Villar who gave me the opportunity to join the department and do my research. Most of the projects were supported by a grant from the Bill & Melinda Gates Foundation to the University of Oxford, for which I am very grateful.

Dr E.O. Ohuma, Dr V. Donadono, Dr A. Cavallaro, Dr S.K. Wanyonyi, Dr M. Molloholli have been precious co-authors in the analysis of the data, leading and co-authoring projects.

I want to express my gratitude also to Prof. Martinelli for his guidance and the Department of Neuroscience in allowing my residency in the UK to conduct my research. I cannot not mention the team in Naples of the High Risk Pregnancy Unit, University Federico II for the collaboration. I would like also to thank all of my friends, God, my mum, my brother and the enlarged family.

My special thought is for my love Linda, who supported me and walked with me during these transition years in the UK toward a bright future together.