"A comparison of the accuracy of various methods of postnatal gestational age estimation; including Ballard Score, Foot Length, Vascularity of the Anterior Lens, Last Menstrual Period and also a clinician's non-structured assessment."

A dissertation submitted in partial fulfilment of the degree of Master of Philosophy in Neonatology , for the University of Cape Town

Alexander Graham Stevenson (STVALE002)

Table of Contents

25 January 2021

Declaration

The research reported in this dissertation is based on independent work performed by the myself, the candidate, and neither the whole nor any part of it has been or is being submitted for another degree to any other university.

This work has not been reported or published prior to registration for this degree of Master in Philosophy, Neonatology.

Signed this 25th day of January 2021

Signed by candidate

Alexander G. Stevenson

Abstract

Rationale

Gestational age is a strong determinant of neonatal mortality and morbidity. Early obstetric ultrasound is the clinical reference standard, but is not widely available in many developing countries. There is a well recognised need to identify reliable and simple methods of postnatal gestational age estimation.

Methods

A prospectively designed methods comparison study in a tertiary referral hospital in a developing country. Early ultrasound (<20 weeks) was the clinical reference standard. Methods evaluated included anthropometric measurements (including foot-length), vascularity of the anterior lens, the New Ballard Score and Last Menstrual Period. Clinicians' non-structured global impression "End of Bed" Assessment was also evaluated.

Results

106 babies were included in the study. Median age at birth was 34 weeks (IQR 29-36). Ballard Score and "End of Bed" Assessment had a mean bias of -0.14 and 0.06 weeks respectively but wide 95% limits of agreement. The physical component of the Ballard score, the total Ballard score and Foot-length's ability to discriminate between term and preterm infants gave an AUROC of 0.97, 0.96 and 0.95 respectively.

Discussion

Although “End of Bed” Assessment and Ballard score had small mean biases, the wide confidence intervals render the methods irrelevant in clinical practice. Foot-length was particularly poor in Small for Gestational Age infants. None of the methods studied were superior to a non-structured clinician’s informal “End of Bed” Assessment.

Conclusion

None of the methods studied met the a priori definition of clinical usefulness. Improving access to early ultrasound remains a priority. Instead of focusing on chronological accuracy, future research should compare the ability of early ultrasound and Ballard score to predict morbidity and mortality.

Acknowledgements

This dissertation was part of a research project with multiple contributors.

Alexander Stevenson, the candidate, conceptualised the research project, wrote the protocol, applied for ethics approval and was one of three researchers who actually performed the gathering of data. He collated the data and liased with the statistician to statistically analyse and represent the data. He wrote the manuscript, but received input from the supervisors and research assistants and the statistician. When the manuscript was submitted for publication, Alexander Stevenson was the first author.

Lloyd Tooke and **Yaseen Joolay** were supervisors and gave input into the protocol design, the statistical analysis and the final version of the manuscript.

Candice Levetan and **Caris Price** assisted in the process of examining the babies and gathering data for the research project. They also had input into the final version of the manuscript.

Stanzi Le Roux joined the team as the statistician and helped organize and analyze and represent the data. She contributed to the final manuscript and guided the writing of the results, tables and figures sections.

List of Tables

List of Figures

SUPPLEMENTAL FIGURE 2. Receiver operating curves for measures to identify neonates ≥ 34 vs < 34 weeks' gestational age
S Figure 2 (a) Ballard scores
S Figure 2 (b) Foot length
S Figure 2 (c) Anthropometric measures

SUPPLEMENTAL FIGURE 3. Receiver operating curves for measures designed to identify neonates ≥ 28 vs <28 weeks' gestational age
S Figure 3 (a) Ballard scores
S Figure 3 (b) Foot length
S Figure 3 (c) Anthropometric measures

SUPPLEMENTAL FIGURE 4. Study Flow Diagram

Abbreviations

AC	Abdominal Circumference
AGA	Appropriate for Gestational Age
ALA	Anterior Lens Assessment
AUROC	Area Under the Receiver Opperator Curve
CI	Confidence Interval
FLC	Foot length, Calipers
FLR	Foot Length, Ruler
HC	Head Circumference
HREC	Human Research Ethics Committee
IQR	Interquartile Range
LBW	Low Birth Weight
LGA	Large for Gestational Age
LMP	Last Menstrual Period
LOA	Limit of Agreement
LR	Likelyhood Ratio
MUAC	Mid Upper Arm Circumference
PTB	Preterm Birth
ROC	Receiver Operator Curve
SGA	Small for Gestational Age

Introduction

Preterm birth is the leading cause of childhood death worldwide (1). The gestational age determines the type and severity of pathologies experienced by the newborn. Sub-Saharan Africa has the highest rates of preterm birth in the world. The highest is in Malawi (18.1%) whilst South Africa is reported to have a rate of 12.4% (2).

Most research on the accuracy of postnatal gestational age was performed in the era before early ultrasound scanning became routine (3)(4). These studies were flawed as the methods of gestational age estimate were evaluated against an imperfect standard: the last menstrual period(5) . Prior to 14 weeks gestation, obstetric ultrasound has an accuracy of 5 to 7 days. In the first half of the second trimester (before 20 weeks) the accuracy is of within 7 to 10days (6,7). In developing countries relatively few mothers have access to early (dating) ultrasounds (8) (9). Improving methods of ascertaining gestational age has been identified as a research priority in the World Health Organisation's Every Newborn Action Plan (10).

A systemic review in 2019 suggested that scoring systems such as Dubowitz or Ballard had relatively little mean bias but wide 95% Confidence Intervals which limited their clinical usefulness(4). The paper highlighted the need to specifically look at gestational age estimates in Small for Gestational Age (SGA) babies.

A Cape Town study suggested foot length may be useful but the study excluded SGA babies (11). Although clinicians can form an opinion about a baby's gestational age through a non-structured examination of the baby from the "End of the Bed", this has not previously been scientifically analysed.

The primary objective of the study was to identify the most accurate post-natal clinical tools to estimate gestational age in a secondary or tertiary level hospital. Secondary objectives included evaluating clinical test performance in subgroups of infants based on size at birth (low birthweight and/or SGA) and prematurity (extremely preterm infants).

Methods

This methods comparison study was designed prospectively, with the clinical reference standard measurement (early antenatal ultrasound) preceding the index tests.

Study participants were recruited at Groote Schuur Maternity Hospital, a tertiary level referral maternity hospital in Cape Town, South Africa. Most mothers delivering at Groote Schuur are from peri-urban, low income areas. In the Western Cape, neonatal mortality was 8.2/1000 in 2010-2015 (12) with preterm birth (PTB) and congenital abnormalities being the two leading causes of death(13). Pregnant women in the West Metropole of Cape Town are offered a dating scan at their community clinic performed by a trained radiographer.

The researchers were only present on the neonatal unit six days a week, so convenience sampling was used. Mothers with newborn infants aged less than 48 hours were invited to participate in the study if they had had a dating ultrasound scan performed before 20 weeks of gestation documented in their antenatal record and the baby could be examined before 48 hours of age. Hospital staff notified study clinicians of potentially eligible neonates during working hours. Exclusion criteria included major congenital or genetic abnormalities, neurological abnormalities, sedation, critical illness (including mechanical ventilation) and structural abnormalities of the lower limbs. Babies were also excluded if the researchers were inadvertently told the gestational age before assessing the baby.

Measurements

The primary index test was the New Ballard Score(14); secondary index tests included (1) last menstrual period (LMP) (overall, and sub analysis of women who were confident of the LMP date); (2) foot length measurements :Foot Length with Callipers (FLC) and Foot Length with Ruler, (FLR) (3), anterior lens assessments (ALA); and (4) anthropometric measures (including birthweight [BW] and birth length ([BL]; and mid-upper arm [MUAC], abdominal [AC] and head circumferences [HC] measured at time of assessment (15). Calculation of scores, lens vascularity and anthropometric measurements was performed according to published data (11,14,16,17). Standard definitions of growth restriction and weight categories were used, summarised in Supplementary table 1 (18,19) (17).

Maternal interviews, data abstraction and clinical assessments were completed by study staff, consisting of three junior medical clinicians who had completed their internship and had been working in the neonatal unit for at least six months. Specialized neonatologists provided structured, standardized training in foot-length measurement, Ballard scoring and anterior lens assessment. Data abstraction from clinical records included estimated date of delivery from early ultrasound, date of birth, and birth measurements including sex and anthropometry.

Clinical examinations for gestational dating were blinded to the ultrasound-based gestational age at birth and any other available history. Each clinical evaluation was repeated by a second study clinician, also blinded to the findings of the first clinician. The researchers began by simply looking at the baby and performing a brief non structured examination to formulate a best guess “End of Bed” Assessment which was recorded. They then measured anthropometry including circumferential measures (using non-stretchable measuring tape following standardized procedures) for the mid-upper arm, head circumference, length and the abdominal circumference. Foot length measurements were done from the middle of the heel to the end of the longest toe, following standard protocol(11). Each measure was repeated by the clinician and analysis used the average of two measures per assessor to allocate an estimated gestation (supplemental table 2). Ballard scoring followed published guidelines (supplemental table

2)(14). Finally, study clinicians assessed the vascularity of anterior lens (anterior lens analysis, ALA) using direct ophthalmoscopy, without dilatation. If the baby's eye lids could be separated, visualisation of the vascularity on the lens was not technically difficult. However if the baby resisted opening their eyes to gentle manual pressure the procedure was abandoned and no lid retractors were used. Grading of lens vascularity was based on previously published guidelines(16) (supplemental table 2).

Study clinicians recorded the date of last menstrual period from maternal interviews, including an indicator for how sure the mother was about the date (binary: sure vs unsure).

Statistical methods

We defined clinically acceptable mean bias of gestational age *a priori* as ±1 week or less. With $\alpha=0.05$, we estimated a sample size of 110 would provide 84% power to detect ≥1 week (paired mean) difference in gestational age, assuming a known standard deviation of 3.5; or 74% power assuming a known standard deviation of 4. Throughout, estimates are presented with 95% confidence intervals (CI) to demonstrate achieved precision. Exploratory analysis utilized standard biostatistical approaches. Ballard scores, foot length measures and ALA grading were used to calculate continuous measures of gestational age in weeks (supplemental table 2)(14,16,20). For analysis of accuracy for different cut-off values, gestational age by ultrasound was dichotomized into (a) ≥37 vs <37, (b) ≥34 vs < 34 and (c) ≥28 vs < 28 weeks. Interrater variability was compared for postnatal clinical measures using Lin's correlation coefficients for continuous and Krippendorff's alpha for categorical measures (21). Continuous measures of gestation obtained by postnatal methods were compared to the reference standard with Lin's covariate analysis and the Bland-Altman approach, with Bradley-Blackwood F-tests to assess concordance over increasing values of gold standard gestational age estimates (test for

significant differences between means and variances; p≥0.05 indicating acceptable concordance) (22,23). Finally, diagnostic accuracy [to identify preterm birth at different thresholds, (a) <37 vs ≥37, (b) <34 vs ≥ 34 and (c) <28 vs ≥ 28 weeks] was compared for the most robust clinical measures (identified as those with low interrater variability; constant variance and small mean bias on Bland-Altman graphs), using area under the curve (AUROC, receiver operating characteristics curve) approaches to identify boundaries that optimize sensitivity and specificity. Additional analyses that were planned a priori were for subgroups of low birthweight (<2500g) and SGA (birthweight < I0th centile, according to Intergrowth-21st reference standards were used(17). Throughout, p-values are two-sided with α=0.05.

Ethical considerations

We obtained written, informed consent from the mother or legal guardian after discussing the study aims, potential risks and benefits and answering any questions (Appendix A). The University of Cape Town Human Research Ethics Committee approved this study (HREC-233/2018). Participant confidentiality was maintained throughout. Data was entered onto a standardised paper form and kept in a secure location within the neonatal unit of Groote Schuur Hospital. De-identified data were entered into Epi-info 7 and then exported to Stata 14.0 (Statacorp, College Station TX) for further cleaning and analysis.

Results

Study population

Overall, 106 infants were included in the analysis (Table 1). The median (interquartile range, IQR) gestation at time of ultrasound was 13 (11-17) weeks, and at birth, 34 (29-36) weeks. Twenty two percent (23/106) infants were term, 75% (79/106) were low birth weight (LBW) <2500g, 48%

(51/ 106) were boys. Distribution of gestational age is shown in figure 1. Additional information regarding anthropometry and allocation of gestational age by the methods under analysis are available in supplemental tables 3 and 4 respectively.

Interrater variability

Interrater agreement (supplemental table 6) was moderate for the overall Ballard score (ρ_c=0.89, 95% CI 0.85-0.93), the physical (external) Ballard score (ρ_c=0.88, 95% CI 0.83-0.92) and the estimated gestation based on Ballard score transformation (ρ_c=0.88, 95% CI 0.83-0.92), but poor for the neuromuscular score (ρ_c=0.77, 95% CI 0.68-0.84). Foot length measures and "End of bed" estimations had good interrater agreement (>0.90 each), as did head and abdominal circumferences. ALA only achieved fair agreement (α_K=0.47, benchmark interval 0.2-0.4).

Measurement correlation and mean bias (Bland-Altman analysis)

Overall, gestation based on total Ballard scores had a small mean bias of -0.14 weeks compared to the reference, but wide 95% limits of agreement (LOA, -2.93 to 2.65) weeks [table 2, fig 2(a)]. There was lack of concordance for gestation based on LMP (all measures as well as when limited to sure dates) or foot length measures (ruler or calipers), as indicated by Bradley-Blackwood p<0.05 (table 2 and supplemental table 5) and the regression lines shown in figures 2(b-e). "End of bed" estimations had an estimated mean bias of 0.06 (95% LOA -2.86 to 2.98) [figure 2f]. Results were similar when restricted to SGA, LBW and very preterm infants (table 2 and supplemental table 5).

Diagnostic accuracy by cut-off values to identify preterm (<37 weeks)

ROC curves for index tests to identify preterm (<37 weeks) demonstrated moderate to good diagnostic performance in most instances (figure 3). Boundaries selected for optimal sensitivity and specificity are shown in table 3, with accompanying point estimates (95% CI) of sensitivity, specificity and likelihood ratio tests (positive, LR+ and negative, LR-). The positive and negative predictive values of the tests will vary with the

background prevalence of the condition. For each measure boundary, we present three scenarios of pre-test probability (i.e. projected underlying true prevalence of preterm <37 weeks): the prevalence experienced in the study, a hypothetical background prevalence of 10% and a hypothetical background prevalence of 50%. Our data supports using different threshold values for postnatal gestational evaluation depending on whether the aim is to include preterm infants (e.g. for specific therapeutic interventions) or to exclude preterm infants (i.e., to identify term infants for discharge and follow-up purposes)(24). Our data (table 3, supplemental table 7) support using a total Ballard score < 30 (LR+ 18.3) or foot length < 70mm (LR+ 20.2) to identify preterm infants specifically. For exclusion purposes, our data support using a Ballard score ≥ 35 (LR- 0.03) or foot length ≥ 75mm (LR- 0.02).

Diagnostic accuracy by cut-off values to identify early preterm (<34 weeks)

Precision was limited due to small sample size (Supplemental Table 8). For inclusion purposes, a total Ballard score of < 25 had the best diagnostic accuracy with high specificity (98%; 95%CI 89.4-99.9%) and Likelyhood Ratio (LR) LR+ (37.5, 95% CI 5.4-263.0). For exclusion purposes (i.e. to identify infants ≥ 34 weeks, for example to facilitate earlier discharge), a total Ballard score ≥ 30 provided the highest sensitivity (96.4%; 95%CI 87.7%-99.6%) and LR- (0.05; 95% CI 0.01-0.19).

Diagnostic accuracy by cut-off values to identify extremely preterm (<28weeks)

Again the very small sample size (n=8) limits precision (Supplemental Table 8). ROC curves are shown in supplemental figure 3. Although imprecise, the best estimated test for inclusion purposes was a Ballard score < 10 (specificity 99%; 95% CI 94%-100%; and LR+ 24.5, 95% CI 2.48-242), and for exclusion purposes, a Ballard score ≥ 15 (sensitivity 87.5%, 95%CI 47.3-99.7%; LR- 0.14, 95% CI 0.02-0.90)

Discussion

No postnatal method of gestational analysis met the a priori definition of clinical usefulness of accuracy within 1 week. This is similar to previous studies(4,25). The best performing methods were the non-structured "End of Bed "Assessment (mean bias 0.06 weeks) and the Ballard (mean bias 0.14 weeks). Although the mean bias was small, the 95% Limits of Agreement were wide: between 2.5 and 3 weeks. Similarly in the subgroup analysis (e.g. SGA infants) Ballard and "End of Bed" Assessment had the smallest mean bias in every sub group analysed whilst foot-length (calliper) had the largest. Foot length tended to underestimate the gestational age, especially in SGA babies. This raises the possibility that despite the efforts at standardisation, the researchers were employing a different technique to that described by that described by Van Wyk et al (20). Another possible explanation is that the population was different. The previously published cohort had excluded SGA babies.

These findings should be interpreted in the light of several study limitations and strengths. As a single-centre study at a tertiary hospital, the results may not be generalisable to primary care settings. This study considered ultrasounds up to 20 weeks gestation as early ultrasounds, but it is known that first trimester ultrasounds are slightly more accurate. That said, the mean gestation at time of initial ultrasound was 13 weeks, so this probably did not introduce much error. The sample size was too small for a methods comparison study, especially for the sub-group analysis. Due to the small numbers in some subgroups including extremely preterm and SGA, precision was limited for sensitivity analyses. High interrater variability in ALA suggested a high risk of misclassification bias, thereby precluding meaningful evaluation of diagnostic accuracy. Despite these limitations, using blinded and repeated assessments allowed estimation of interrater variability, a key strength of the study. Another strength was the wide range of clinical measures reported.

Conclusion

The results from this paper suggest that none of the methods studied have clinical utility in individual patients. Both New Ballard Score and an informal "End of Bed" assessment are very accurate in population analysis, provided the clinicians assessing the infants have been well-trained and have some experience in administering the assessments. .

Instead of pursuing what might be a fruitless task seeking an acceptable postnatal gestational age estimator it may be more worthwhile to invest in increasing access to early ultrasound scans(26).

It is possible that postnatal assessments of maturity may themselves independently predict outcomes such as mortality and morbidity as well, or perhaps even better, than gestational age calculated by early ultrasound and this is an exciting area for future research.

Funding

This research was supported by a Research Award from the Department of Paediatrics and Child Health, University of Cape Town.

Acknowledgements

The authors gratefully acknowledge the statistical analyses provided by Dr Stanzi M le Roux, from the division of Epidemiology and Biostatistics, School of Public Health and Family Medicine, University of Cape Town.

References

1. Chawanpaiboon S, Vogel JP, Moller A-B, Lumbiganon P, Petzold M, Hogan D, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. Lancet Glob Health. 2019 Jan 1;7(1):e37–46.

2. Althabe F, Howson CP, Kinney M, Lawn J, World Health Organization. Born too soon: the global action report on preterm birth [Internet]. 2012 [cited 2020 Apr 19]. Available from: http://www.who.int/pmnch/media/news/2012/201204%5Fborntoosoon-report.pdf

3. Diagnostic Accuracy of Neonatal Assessment for Gestational Age Determination: A Systematic Review | Review Articles | Pediatrics December 2017, 140 (6) e20171423; DOI: https://doi.org/10.1542/peds.2017-1423

4. Diagnostic Accuracy of Neonatal Assessment for Gestational Age Determination: A Systematic Review | American Academy of Pediatrics [Internet]. [cited 2020 Feb 15]. Available from: https://pediatrics.aappublications.org/content/140/6/e20171423

5. Savitz DA, Terry JW, Dole N, Thorp JM, Siega-Riz AM, Herring AH. Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination. Am J Obstet Gynecol. 2002 Dec 1;187(6):1660–6.

6. Benson CB, Doubilet PM. Sonographic prediction of gestational age: accuracy of second- and third-trimester fetal measurements. Am J Roentgenol. 1991 Dec 1;157(6):1275–7.

7. Committee Opinion No 700: Methods for Estimating the Due Date. Obstet Gynecol. 2017 May;129(5):e150.

8. Ultrasound in Africa: what can really be done? in: Journal of Perinatal Medicine Volume 44 Issue 2 (2016)

9. Goldenberg RL, Nathan RO, Swanson D, Saleem S, Mirza W, Esamai F, et al. Routine antenatal ultrasound in low- and middle-income countries: first look – a cluster randomised trial. BJOG Int J Obstet Gynaecol. 2018;125(12):1591–9.

10. Every Newborn Metrics Technical Consultation (3-4 December, 2014 ¦ Ferney Voltaire, France) [Internet]. Healthy Newborn Network. 2014 . Available from: https://www.healthynewbornnetwork.org/blog/every-newborn-event-summary-enap-metrics-meeting/(cited 2020 Feb15)

11. Wyk LV, Smith J. Postnatal Foot Length to Determine Gestational Age: A Pilot Study. J Trop Pediatr. 2016 Apr;62(2):144.

12. Hendricks MK, ChB M, Hawkridge A, ChB M, Jacobs L, ChB M, et al. Tracking progress on the health status and service delivery outcomes for neonates and children in the Metro West geographic service area of the Cape Metropole, 2010 - 2015. :8.

13. Hendricks M, Hawkridge A, Jacobs L, Evans J, Mahomed H, Linley L, et al. Tracking progress on the health status and service delivery outcomes for neonates and children in the Metro West geographic service area of the Cape Metropole, 2010 - 2015. South Afr J Child Health. 2019 Apr 11;13(1):36-43–43.

14. Ballard JL, Khoury JC, Wedig K, Wang L, Eilers-Walsman BL, Lipp R. New Ballard Score, expanded to include extremely premature infants. J Pediatr. 1991 Sep 1;119(3):417–23.

15. Anthropometry Procedures Manual. :120.

16. Hittner HM, Hirsch NJ, Rudolph AJ. Assessment of gestational age by examination of the anterior vascular capsule of the lens. J Pediatr. 1977 Sep;91(3):455–8.

17. Villar J, Giuliani F, Fenton TR, Ohuma EO, Ismail LC, Kennedy SH. INTERGROWTH-21st very preterm size at birth reference charts. The Lancet. 2016 Feb 27;387(10021):844–5.

18. WHO | Anthropometric reference data for international use [Internet]. Available from: https://www.who.int/childgrowth/publications/deonis_habicht_1996/en/(cited 2020 Aug 31)

19. PMNCH | Born Too Soon: The Global Action Report on Preterm Birth [Internet]. WHO. [cited 2020 Feb 12]. Available from: http://www.who.int/pmnch/knowledge/publications/preterm_birth_report/en/

20. Postnatal Foot Length to Determine Gestational Age: A Pilot Study | Journal of Tropical Pediatrics | Oxford Academic [Internet]. [cited 2020 Mar 2]. Available from: https://academic.oup.com/tropej/article/62/2/144/2375047

21. Krippendorff K. Computing Krippendorff's Alpha-Reliability. Dep Pap ASC [Internet]. 2011 Jan 25; Available from: https://repository.upenn.edu/asc_papers/43

22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet Lond Engl. 1986 Feb 8;1(8476):307–10.

23. Blackwood LG, Bradley EL. An omnibus test for comparing two measuring devices [Internet]. 1991 . Available from: https://www.semanticscholar.org/paper/An-omnibus-test-for-comparing-two-measuring-devices-Blackwood-Bradley/e4ea4d66ff59106d84d106bb0e74bce5c5cc0154(cited 2020 April 19)

24. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004 Jul 15;329(7458):168–9.

25. Lee AC, Mullany LC, Ladhani K, Uddin J, Mitra D, Ahmed P, et al. Validity of Newborn Clinical Assessment to Determine Gestational Age in Bangladesh. Pediatrics [Internet]. 2016 Jul 1 ;138(1). Available from: https://pediatrics.aappublications.org/content/138/1/e20153303(cited 2020 Feb 12)

26. Kim ET, Singh K, Moran A, Armbruster D, Kozuki N. Obstetric ultrasound use in low and middle income countries: a narrative review. Reprod Health. 2018 Jul 20;15(1):129.

Tables

TABLE 1. Characteristics of study participants: comparing term (≥ 37 weeks) and preterm (<37 weeks) newborn infants classified using gestational age based on early ultrasound

	Total N=106	**Preterm[1] N=83**	**Term[1] N=23**	**Missing data**
Gestational age at time of ultrasound (weeks)[1]	13 (11-17)	13 (11-17)	14 (10-17)	0
Gestational age at birth, according to early ultrasound (weeks)[1]	34 (29-36)	33 (29-35)	38 (38-38)	0
Male sex	51 (48%)	40 (48%)	11 (48%)	0
Twins	23 (22%)	20 (24%)	3 (13%)	0
Birth weight (g)	1750 (1155-2500)	1560 (1080 – 2000)	3300 (2565-3800)	0
Categories of birth weight				
< 1000g	17 (16%)	17 (20%)	0	
≥ 1000, < 1500g	20 (19%)	20 (24%)	0	
≥ 1500, < 2500g	42 (40%)	38 (46%)	4 (17%)	
≥ 2500g	27 (25%)	8 (10%)	19 (83%)	
Small for Gestational Age	25 (24%)	24(29%)	1(4%)	

	Total N=106	Preterm[1] N=83	Term[1] N=23	Missing data
Appropriate for Gestational Age	72(68%)	55(66%)	17(74%)	
Large for Gestational Age	9(8%)	4(5%)	5(22%)	

Numbers are median (interquartile range) or n (column percentage). Abbreviations: g, gram; cm, centimeters; postnatal clinical measurement values based on average from two assessors

[1] According to ultrasound done at first antenatal booking visit

TABLE 2. Agreement of methods for clinical postnatal gestational age assessment compared to early ultrasound (continuous measures, expressed in weeks) overall and amongst SGA neonates

	GA (weeks) mean (SD; range)	Mean bias[1] (95% Limits of agreement)	Lin's concordance Rho[2] (95% CI)	Correlation: mean and difference[3]	*p*-value[4] (Bradley-Blackwood F-test)
All neonates (N=106)					
Early ultrasound (reference), n=106	33.3 (3.9; 24.1-41.3)	Ref	Ref	-	-
Last menstrual period (all), n=81	34.2 (4.8; 23.0-46.6)	-0.70 (-7.85 to 6.44)	0.63 (0.49-0.74)	-0.29	0.01
Ballard total score, n=106	33.4 (3.9; 24.0-42.0)	-0.14 (-2.93 to 2.65)	0.93 (0.90-0.95)	0.02	0.59

	GA (weeks) mean (SD; range)	Mean bias[1] (95% Limits of agreement)	Lin's concordance Rho[2] (95% CI)	Correlation: mean and difference[3]	*p*-value[4] (Bradley-Blackwood F-test)
Foot length, caliper, n=105	31.5 (4.2; 23.0-41.0)	1.85 (-2.24 to 5.95)	0.79 (0.71-0.84)	-0.16	<0.0001
End of bed assessment, n=87	33.1 (3.8; 26.0-40.0)	0.06 (-2.86 to 2.98)	0.92 (0.89-0.95)	-0.04	0.87
SGA neonates (N=25)					
Early ultrasound (reference), n=25	32.0 (4.0; 24.1-41.3)	Ref	Ref	-	-
Last menstrual period (all), n=19	33.9 (5.2; 23.0-42.1)	-1.78 (-8.45 to 4.89)	0.69 (0.40-0.86)	-0.31	0.04
Ballard total score, n=25	31.7 (3.8; 24.0-40.0)	0.35 (-2.02 to 2.71)	0.95 (0.89-0.98)	0.21	0.22
Foot length, caliper, n=24	28.3 (3.3; 23.0-36.0)	3.78 (0.29 to 7.27)	0.57 (0.38-0.72)	0.46	<0.0001
End of bed assessment, n=21	31.1 (3.9; 26.0-39.0)	1.15 (-1.37 to 3.67)	0.90 (0.79-0.96)	0.02	0.003

Abbreviations: GA, gestational age at birth; SD, standard deviation; IQR, inter-quartile range; CI, confidence interval
[1]"Mean bias": average difference in estimated gestational age between gold standard (early ultrasound) and each comparator (US – comparator); therefore, a positive mean bias indicates underestimation by the comparator, a negative mean bias indicates overestimation; [2] Correlation coefficient for gold standard and comparator; [3] Closer to zero is required to meet model assumptions; greater correlation implies that the relationship between the gold standard and comparator method varies over different values of gestational age; [4] F-test simultaneously tests for significant differences between means and variances (non-significant p-value indicates concordance)

TABLE 3. Diagnostic accuracy of postnatal clinical methods to identify preterm infants (< 37 weeks) using different clinical cut-off values for different pre-test probabilities (underlying prevalence estimates, sample-specific and hypothetical)

Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
Ballard total overall score	Score < 35	97.6 (91.6-99.7)	73.9 (51.6-89.8)	3.74 (1.88-7.45)	0.03 (0.01-0.13)	78% (69-86%)	93.1 (85.6-97.4)	89.5 (66.9-98.7)
						10% *(hypothetical)*	29.4 (17.3-45.3)	99.6 (98.6-99
						50% *(hypothetical)*	78.9 (65.3-88.2)	96.8 (88.4-99.2)
	Score < 30	79.5 (69.2-87.6)	95.7 (78.1-99.9)	18.3 (2.68-125.0)	0.21 (0.14-0.33)	78% (69-86%)	98.5 (92.0-100.0)	56.4 (39.6-72.2)
						10% *(hypothetical)*	67.0 (23.0-93.3)	97.7 (96.5-98.5)
						50% *(hypothetical)*	94.8 (72.8-99.2)	82.4 (75.2-87.8)
Ballard total neuro-	Score <16	84.3 (74.7-91.4)	87.0 (66.4-97.2)	6.47 (2.24-18.6)	0.18 (0.11-0.30)	78% (69-86%)	95.9 (88.5-99.1)	60.6 (42.1-77.1)
						10% *(hypothetical)*	41.8 (19.9-67.4)	98 (96.7-98.8)

Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
muscular score						50% *(hypothetical)*	86.6 (69.2-94.9)	84.7 (76.7-90.4)
Ballard total external (physical) score	Score <17	91.6 (83.4-96.5)	87.0 (66.4-97.2)	7.02 (2.44-20.2)	0.10 (0.05-0.20)	78% (69-86%)	96.2 (89.3-99.2)	74.1 (53.7-88.9)
						10% *(hypothetical)*	43.8 (21.3-69.2)	98.9 (97.8-99.5)
						50% *(hypothetical)*	87.5 (70.9-95.3)	91.2 (83.3-95.5)
Birthweight	< 2500g	90.4 (81.9-95.7)	82.6 (61.2-95.0)	5.2 (2.1-12.7)	0.12 (0.06-0.23)	78% (69-86%)	94.9 (87.5-98.6)	70.4 (49.8-86.2)
						10% *(hypothetical)*	36.6 (19.1-58.5)	98.7 (97.5-99.4)
						50% *(hypothetical)*	83.9 (68.0-92.7)	89.6 (81.2-94.4)
	< 2000g	73.5 (62.7-82.6)	100.0 (85.2-100.0)	-	0.26 (0.19-0.38)	78% (69-86%)	100.0 (94.1-100.0)	51.1 (35.8-66.3)
						10% *(hypothetical)*	100.00 (-)	97.1 (96.0-98.0)

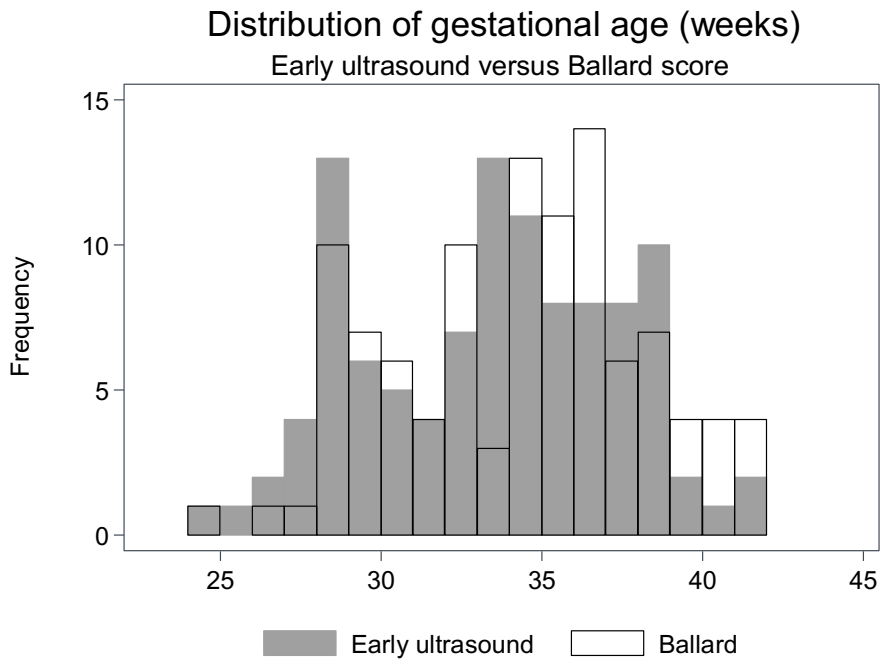
Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
						50% *(hypothetical)*	100.00 (-)	79.0 (72.5-84.4)
Head circumference	< 33cm	94.9 (87.4-98.6)	71.4 (47.8-88.7)	3.32 (1.69-6.54)	0.07 (0.03-0.19)	78% (69-86%)	92.5 (84.4-97.2)	78.9 (54.4-93.9)
						10% *(hypothetical)*	27.0 (15.8-42.1)	99.2 (97.9-99.7)
						50% *(hypothetical)*	76.9 (62.8-86.7)	93.3 (83.8-97.4)
	< 32cm	83.3 (73.2-90.8)	85.7 (63.7-97.0)	5.83 (2.04-16.7)	0.19 (0.11-0.33)	78% (69-86%)	95.6 (87.6-99.1)	58.1 (39.1-75.5)
						10% *(hypothetical)*	39.3 (18.5-65.0)	97.9 (96.5-98.7)
						50% *(hypothetical)*	85.4 (67.1-94.4)	83.7 (75.2-89.7)
Foot length, caliper	< 75mm	98.8 (93.4-100.0)	60.9 (38.5-80.3)	2.52 (1.52-4.21)	0.02 (0.003-0.14)	78% (69-86%)	90.0 (81.9-95.3)	93.3 (68.1-99.8)
						10% *(hypothetical)*	21.9 (14.4-31.8)	99.8 (98.4-100.0)
						50% *(hypothetical)*	71.6 (60.2-80.8)	98.0 (87.4-99.7)

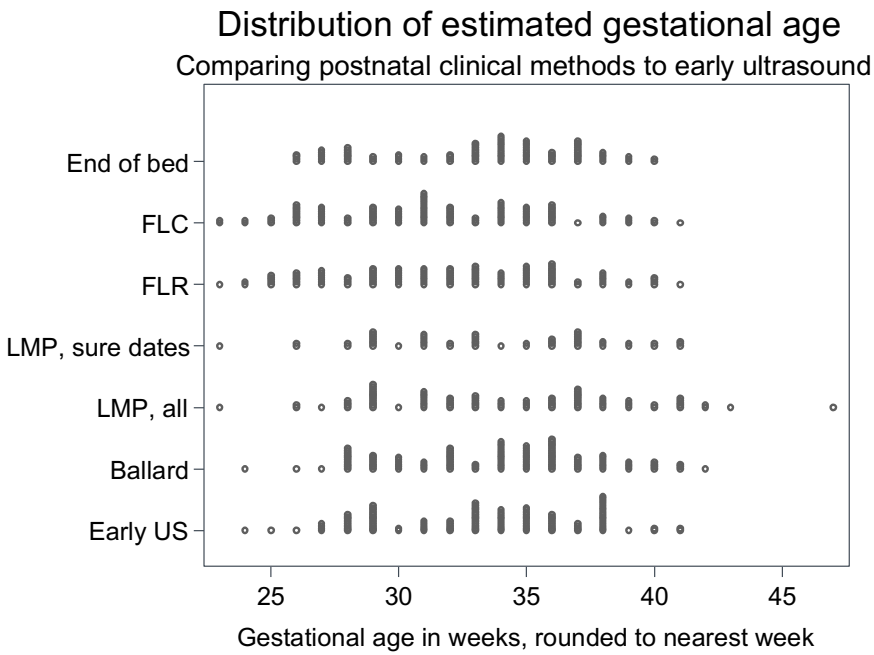
Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
	< 70mm	87.8 (78.7-94.0)	95.7 (78.1-99.9)	20.2 (2.96-138.0)	0.13 (0.07-0.23)	78% (69-86%)	98.6 (92.6-100.0)	68.8 (50.0-83.9)
						10% *(hypothetical)*	69.2 (24.8-93.9)	98.6 (97.5-99.2)
						50% *(hypothetical)*	95.3 (74.8-99.3)	88.7 (81.3-93.4)

Abbreviations: CI, confidence interval; sensitivity = Pr(Test+|Disease+); specificity = Pr(Test-|Disease-); LRT, likelihood ratio (interpretation: +LR, positive likelihood ratio= true positives/false positives; -LRT, negative likelihood ratio=false negatives/true negatives); PPV, positive predictive value = Pr(Disease+|Test+), using given prevalence as pretest probability; NPV, negative predictive value = Pr(Disease-|Test-), using given prevalence of pretest probability
Prevalence estimates: (a) actual prevalence in sample, with 95% CI; (b) hypothetical prevalence of 10%; (c) hypothetical prevalence of 50%

Figures

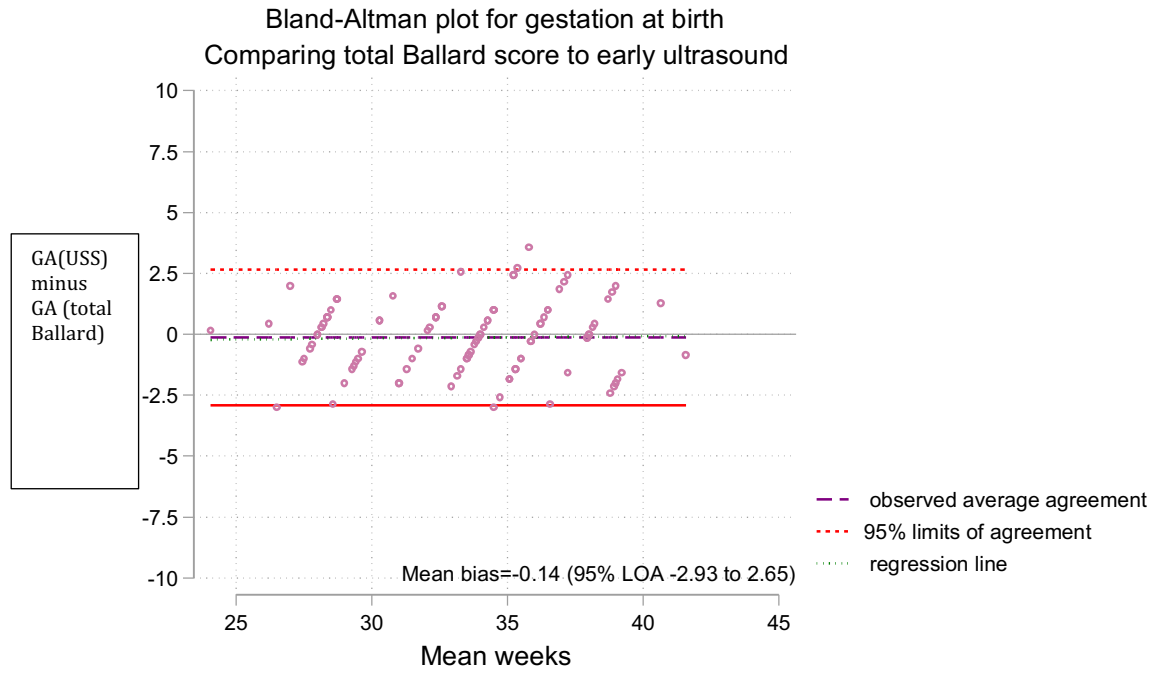
FIGURE 1. Distribution of gestational age comparing early ultrasound to postnatal clinical measures: (a) Histogram comparing early ultrasound and Ballard; (b) One-way graph comparing distributions of early ultrasound and all postnatal clinical measures





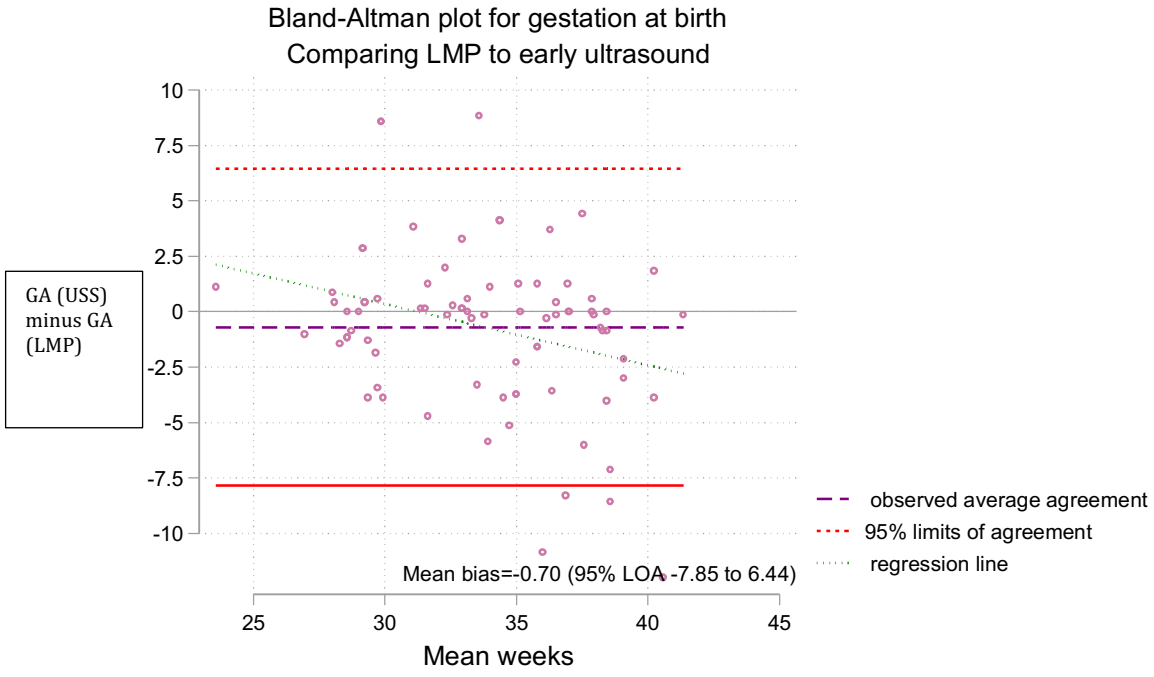
FLC, foot length by calipers; FLR, foot length by ruler; LMP, last menstrual period; US, ultrasound

FIGURE 2 (a) Bland-Altman plot of gestational age in weeks, comparing total Ballard score to early ultrasound



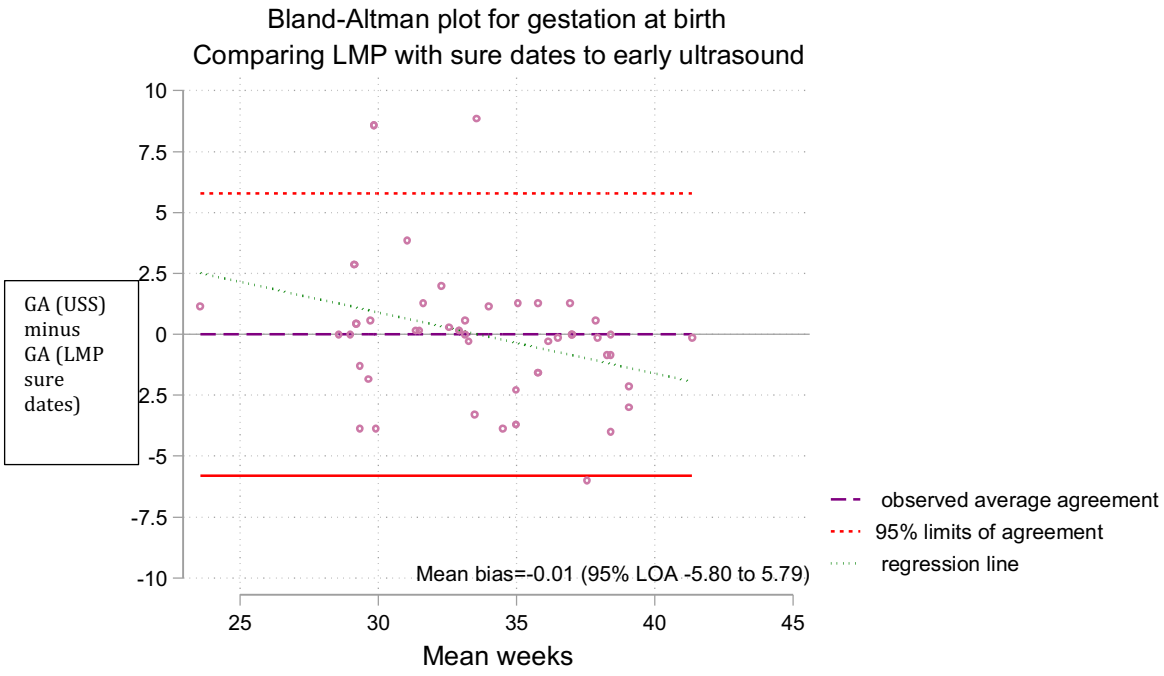
n=106 newborn infants

FIGURE 2 (b) Bland-Altman plot of gestational age in weeks, comparing last menstrual period to early ultrasound



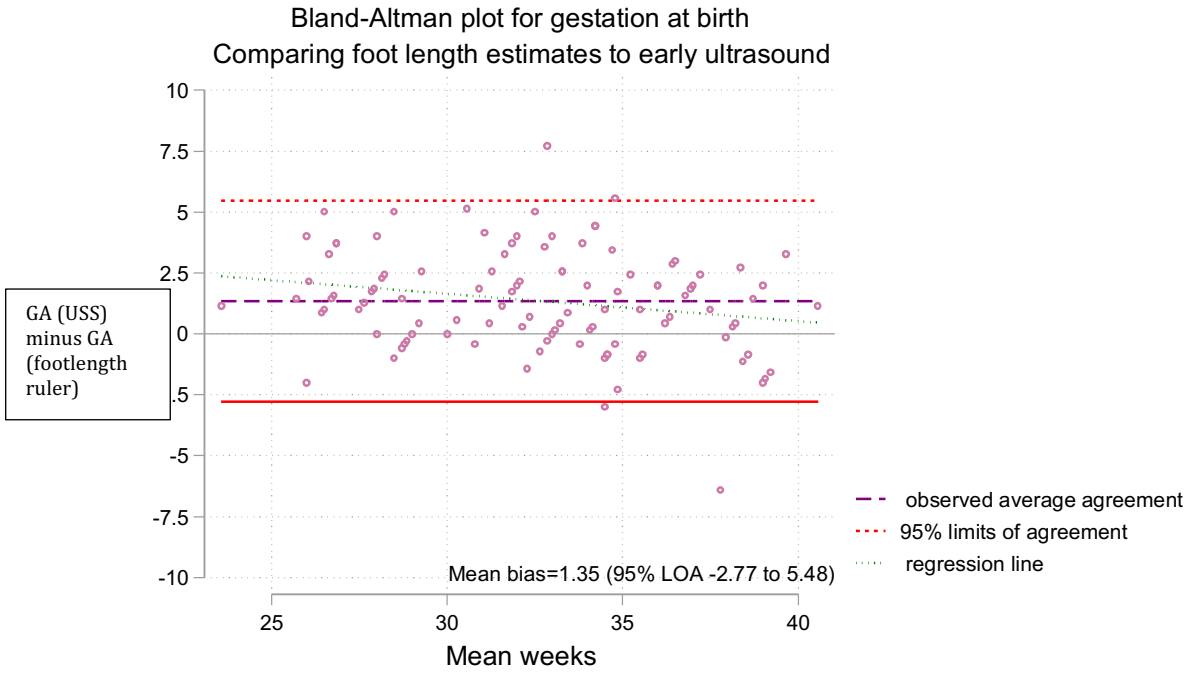
LMP, last menstrual period (all); n=81 newborn infants

FIGURE 2 (c) Bland-Altman plot of gestational age in weeks, comparing last menstrual period (sure dates) to early ultrasound



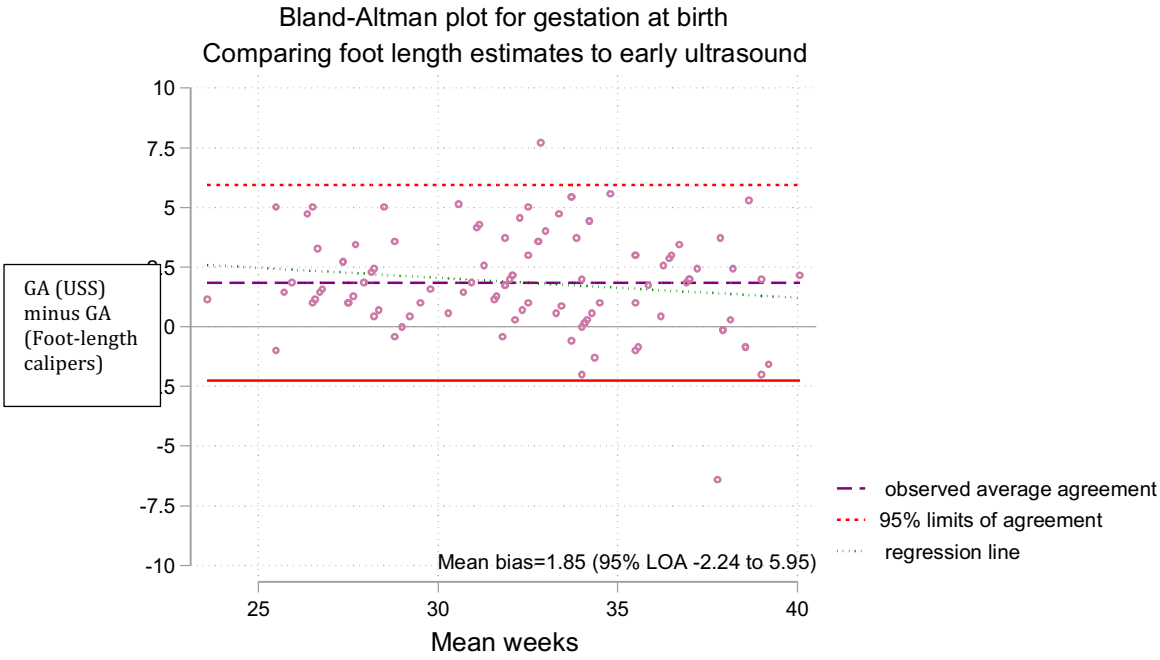
LMP, last menstrual period (sure dates only); n=51 newborn infants

FIGURE 2 (d) Bland-Altman plot of gestational age in weeks, comparing gestational age based on foot length measurements (ruler) to early ultrasound



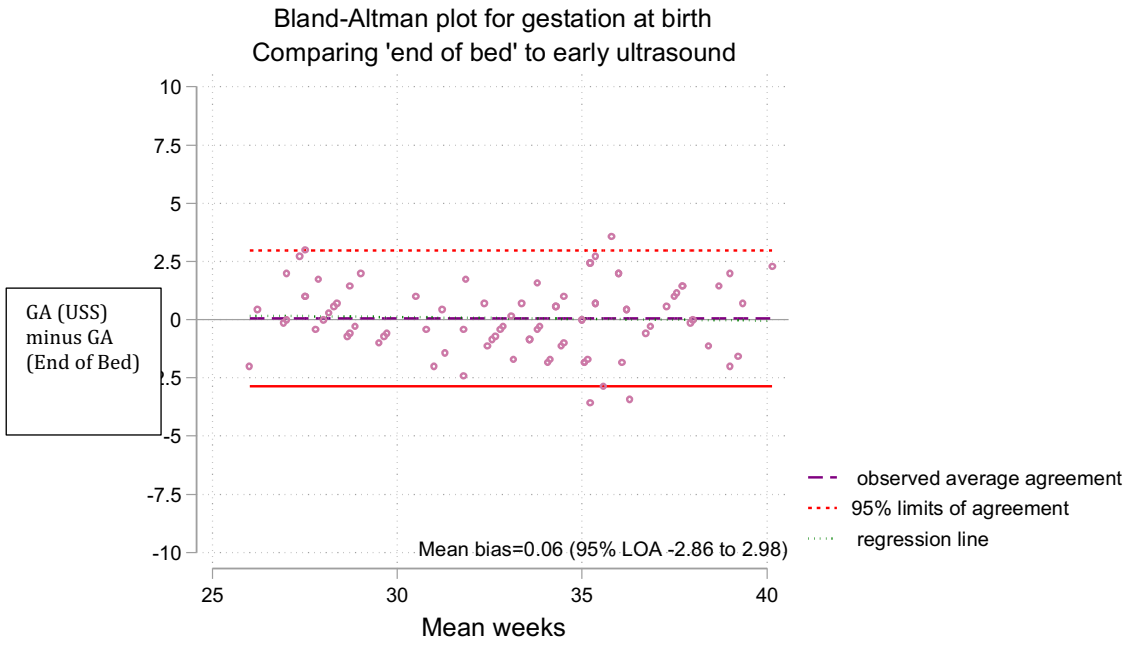
FLR = foot length by ruler converted into GA; n=106 newborn infants

FIGURE 2 (e) Bland-Altman plot of gestational age in weeks, comparing gestational age based on foot length measurements (calipers) to early ultrasound



FLC = foot length by caliper converted into GA; n=105 newborn infants

FIGURE 2 (f) Bland-Altman plot of gestational age in weeks, comparing informal, clinical "end of bed" assessment to early ultrasound



'End of bed' = informal clinical assessment; n=87 newborn infants

FIGURE 3. Receiver operating curves for measures designed to identify infants ≥ 37 weeks' gestational age

Figure 3 (a) Ballard scores

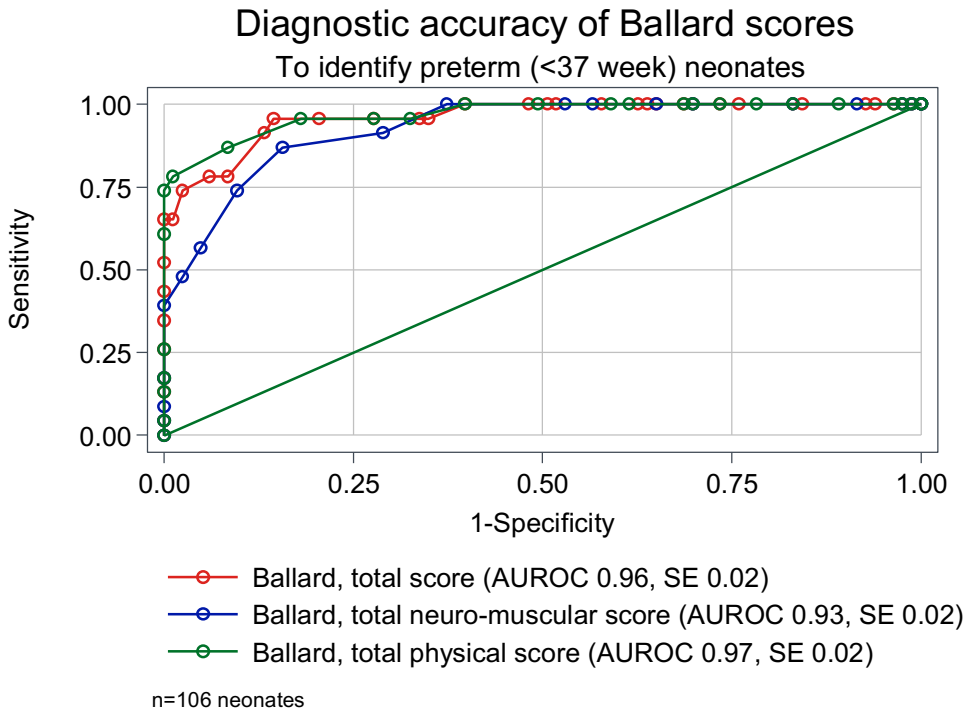


Figure 3 (b) Foot length

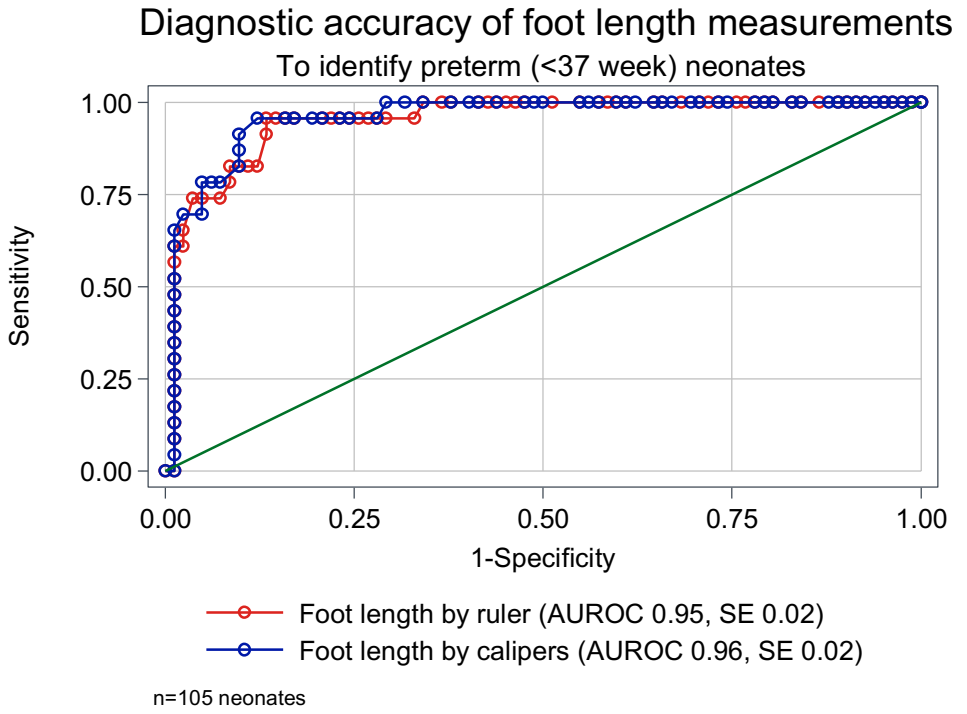
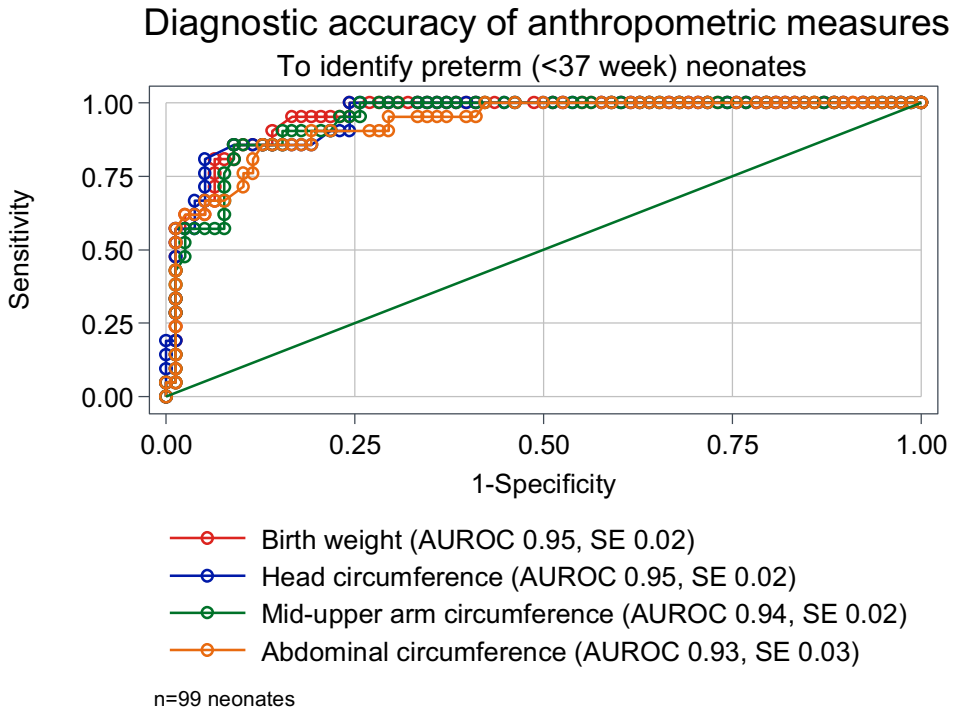


Figure 3 (c) Anthropometric measures



Supplimental Tables

SUPPLIMENTAL TABLE 1: Standard definitions used:

Small for Gestational Age (SGA)	Weight <10th centile for gestational age
Appropriate for Gestational Age (AGA)	Weight between 10th and 90th centile for gestational age
Large for Gestational Age (LGA)	Weight > 90th centile for gestational age
Low Birth Weight (LBW)	Birth weight <2500g
Very Low Birth Weight (VLBW)	Birth weight <1500g
Extremely Low Birth Weight (ELBW)	Birth weight <1000g

SUPPLEMENTAL TABLE 2. Transformation of clinical postnatal measurements into estimated gestational age, by published guidelines and operationalized for analysis

	Published guidelines		Application in analysis (using average of measures by two assessors)	
Measurement	**Published categories**	**Associated gestational age (weeks)**	**Categorization of score used in analysis**	**Allocated gestational age (weeks) in analysis**
Ballard Total Score[1]	≥ -11, <-5	20	≥ -11, < -7.5	20
			≥ -7.5, <-5	21
	≥ -5, < 0	22	≥ -5, <-2.5	22
			≥ -2.5, < 0	23
	≥ 0, < 5	24	≥ 0, < 2.5	24
			≥ 2.5, < 5	25
	≥ 5, < 10	26	≥ 5, < 7.5	26
			≥ 7.5, < 10	27
	≥ 10, < 15	28	≥ 10, < 12.5	28

Measurement	Published guidelines		Application in analysis (using average of measures by two assessors)	
	Published categories	Associated gestational age (weeks)	Categorization of score used in analysis	Allocated gestational age (weeks) in analysis
			≥ 12.5, < 15	29
	≥ 15, < 20	30	≥ 15, < 17.5	30
			≥ 17.5, < 20	31
	≥ 20, < 25	32	≥ 20, < 22.5	32
			≥ 22.5, < 25	33
	≥ 25, < 30	34	≥ 25, < 27.5	34
			≥ 27.5, < 30	35
	≥ 30, < 35	36	≥ 30, < 32.5	36
			≥ 32.5, < 35	37
	≥ 35, < 40	38	≥ 35, < 37.5	38
			≥ 37.5, < 40	39
	≥ 40, < 45	40	≥ 40, < 42.5	40
			≥ 42.5, < 45	41
	≥ 45, < 50	42	≥ 45, < 47.5	42
			≥ 47.5, < 50	43
	≥ 50	44	≥ 50	44
Foot length (mm)[2]	≤44 mm	<24	<44 mm	23
	44.1-45.9 mm	24	≥44, <46 mm	24
	46.0-48.9 mm	25	≥46, <49 mm	25
	49.0-51.9 mm	26	≥49, <52 mm	26
	52.0 – 53.9 mm	27	≥52, <54 mm	27
	54.0-55.9 mm	28	≥54, <56 mm	28
	56.0-58.9 mm	29	≥56, <59 mm	29
	59.0-60.9 mm	30	≥59, 61< mm	30
	61.0-63.9 mm	31	≥61, 64< mm	31
	64.0-65.9 mm	32	≥64, <66 mm	32

	Published guidelines		Application in analysis (using average of measures by two assessors)	
Measurement	**Published categories**	**Associated gestational age (weeks)**	**Categorization of score used in analysis**	**Allocated gestational age (weeks) in analysis**
	66.0-68.9 mm	33	≥66, <69 mm	33
	69.0-70.9 mm	34	≥69, <71 mm	34
	71.0-72.9 mm	35	≥71, <73 mm	35
	73.0-75.9 mm	36	≥73, <76 mm	36
	76.0-77.9 mm	37	≥76, <78 mm	37
	78.0-80.9 mm	38	≥78, <81 mm	38
	81.0-82.9 mm	39	≥81, <83 mm	39
	83.0-84.9 mm	40	≥83, <85 mm	40
	≥ 85.0 mm	>40	≥85	41
Vascularity of anterior lens[3]	Grade 0	≥ 35	Grade 0	≥ 35
	Grade 1	33-34	Grade 1	33-34
	Grade 2	31-32	Grade 2	31-32
	Grade 3	29-30	Grade 3	29-30
	Grade 4	27-28	Grade 4	27-28

1 From Ballard et al (1991) 2 from Van Wyk, Johan Smith (2016); 3 from Hittner et al (1977) average grading rounded up

SUPPLIMENTAL TABLE 3. Characteristics and anthropometry of study participants: comparing term (≥ 37 weeks) and preterm (<37 weeks) newborn infants classified using gestational age based on early ultrasound

	Total N=106	Preterm[1] N=83	Term[1] N=23	Missing data
Gestational age at time of ultrasound (weeks)[1]	13 (11-17)	13 (11-17)	14 (10-17)	0
Gestational age at birth, according to early ultrasound (weeks)[1]	34 (29-36)	33 (29-35)	38 (38-38)	0
Category of prematurity (ultrasound)				0
Term (≥37 weeks)	23 (22%)	0	23 (100%)	
Moderate/late preterm (≥32, <37 weeks)	47 (44%)	47 (57%)	0	
Very preterm (≥28, <32 weeks)	28 (26%)	28 (34%)	0	
Extremely preterm (<28 weeks)	8 (8%)	8 (9%)	0	
Male sex	51 (48%)	40 (48%)	11 (48%)	0
Twins	23 (22%)	20 (24%)	3 (13%)	0
Birth weight (g)	1750 (1155-2500)	1560 (1080 – 2000)	3300 (2565-3800)	0
Categories of birth weight				0
< 1000g	17 (16%)	17 (20%)	0	

	Total N=106	Preterm[1] N=83	Term[1] N=23	Missing data
≥ 1000, < 1500g	20 (19%)	20 (24%)	0	
≥ 1500, < 2500g	42 (40%)	38 (46%)	4 (17%)	
≥ 2500g	27 (25%)	8 (10%)	19 (83%)	
Birthweight percentile for gestational age, categories[1]				0
<3rd percentile	17 (16%)	16 (19%)	1 (4%)	
≥ 3rd, < 10th percentile	8 (7%)	8 (10%)	0	
≥ 10th, < 50th percentile	44 (42%)	36 (43%)	8 (35%)	
≥ 50th, < 90th percentile	28 (26%)	19 (23%)	9 (39%)	
≥ 90th, 97th percentile	5 (5%)	3 (4%)	2 (9%)	
≥ 97th percentile	4 (4%)	1 (1%)	3 (13%)	
Birth length (cm)	42.3 (37.8-46.0)	40.5 (35.0-43.8)	50.0 (46.0–52.5)	16
Birth length percentile for gestational age[1]				9
<3rd percentile	7 (7%)	6 (8%)	1 (5%)	
≥ 3rd, < 10th percentile	9 (9%)	9 (12%)	0	
≥ 10th, < 50th percentile	42 (43%)	33 (43%)	9 (43%)	
≥ 50th, < 90th percentile	29 (30%)	24 (31%)	5 (24%)	
≥ 90th, 97th percentile	5 (5%)	2 (3%)	3 (14%)	
≥ 97th percentile	5 (5%)	2 (3%)	3 (14%)	

	Total N=106	Preterm[1] N=83	Term[1] N=23	Missing data
Head circumference (cm)	30.8 (26.7-32.5)	29.4 (26.5–31.1)	35.0 (32.8–35.8)	7
Mid-upper arm circumference (cm)	7.6 (6.5–9.6)	7.2 (6.0-9.0)	10.7 (9.9-12.1)	0
Abdominal circumference (cm)	25.4 (22.2-28.5)	24.1 (21.5-26.5)	31.3 (28.7-34.2)	0
Foot length from ruler (cm)	65.1 (56.0-72.5)	61.5 (53.7-67.7)	78.5 (73.5–81.2)	0
Foot length from calipers (cm)	63.5 (55.8-71.7)	60.3 (53.2-65.5)	75.2 (72.5 – 80.0)	1

SUPPLIMENTAL TABLE 4. Comparison of Gestational Age allocation by various postnatal methods of gestational age estimation: comparing term (≥ 37 weeks) and preterm (<37 weeks) newborn infants classified using gestational age based on early ultrasound

	Total N=106	Preterm[1] N=83	Term[1] N=23	Missing data
Gestational age at time of ultrasound (weeks)[1]	13 (11-17)	13 (11-17)	14 (10-17)	0

	Total N=106	Preterm[1] N=83	Term[1] N=23	Missing data
Gestational age at birth, according to early ultrasound (weeks)[1]	34 (29-36)	33 (29-35)	38 (38-38)	0
Gestational age at birth, according to postnatal methods (weeks)				
By Ballard score	34 (30-36)	32 (30-34)	38 (36-40)	0
By last menstrual period (all)	34 (31-37)	32 (29-36)	38 (37-39)	25
By last menstrual period (sure dates)	33 (30-37)	32 (29-36)	38 (37-39)	55
By foot length (caliper)	31 (28-35)	30 (27-32)	36 (35-38)	1
By informal "end of bed" assessment	34 (30-36)	33 (29-35)	38 (37-39)	19
Total Ballard score (raw score)	26 (18-32)	24 (15-29)	38 (34-41)	0
Anterior lens assessment				21
Grade 0 (~ ≥35 weeks)	30 (35%)	14 (22%)	16 (76%)	
Grade 1 (~ 33-34 weeks)	22 (26%)	17 (27%)	5 (24%)	
Grade 2 (~ 31-32 weeks)	9 (11%)	9 (14%)	0	
Grade 3 (~ 29-30 weeks)	5 (6%)	5 (8%)	0	
Grade 4 (~ 27-28 weeks)	19 (22%)	19 (30%)	0	

Numbers are median (interquartile range) or n (column percentage). Abbreviations: g, gram; cm, centimeters; postnatal clinical measurement values based on average from two assessors

[1] According to ultrasound done at first antenatal booking visit

SUPPLIMENTAL TABLE 5. Agreement of methods for clinical postnatal gestational age assessment compared to early ultrasound (continuous measures, expressed in weeks) overall and within subgroups of neonates

	GA (weeks) mean (SD; range)	GA (weeks) median (IQR)	Mean bias[1] (95% Limits of agreement)	Lin's concordance Rho[2] (95% CI)	Correlation: mean and difference[3]	*p*-value[4] (Bradley-Blackwood F-test)
All neonates (N=106)						
Early ultrasound (reference), n=106	33.3 (3.9; 24.1-41.3)	33.6 (29.4-36.4)	Ref	Ref	-	-
Last menstrual period (all), n=81	34.2 (4.8; 23.0-46.6)	34.0 (30.6-37.3)	-0.70 (-7.85 to 6.44)	0.63 (0.49-0.74)	-0.29	0.01
Ballard total score, n=106	33.4 (3.9; 24.0-42.0)	34.0 (30.0-36.0)	-0.14 (-2.93 to 2.65)	0.93 (0.90-0.95)	0.02	0.59
Foot length, caliper, n=105	31.5 (4.2; 23.0-41.0)	31.0 (28.0-35.0)	1.85 (-2.24 to 5.95)	0.79 (0.71-0.84)	-0.16	<0.0001
End of bed assessment, n=87	33.1 (3.8; 26.0-40.0)	34.0 (30.0-36.0)	0.06 (-2.86 to 2.98)	0.92 (0.89-0.95)	-0.04	0.87
SGA neonates (N=25)						
Early ultrasound (reference), n=25	32.0 (4.0; 24.1-41.3)	31.4 (28.7-35.0)	Ref	Ref	-	-
Last menstrual period (all), n=19	33.9 (5.2; 23.0-42.1)	33.1 (29.4-36.9)	-1.78 (-8.45 to 4.89)	0.69 (0.40-0.86)	-0.31	0.04

	GA (weeks) mean (SD; range)	GA (weeks) median (IQR)	Mean bias[1] (95% Limits of agreement)	Lin's concordance Rho[2] (95% CI)	Correlation: mean and difference[3]	*p*-value[4] (Bradley-Blackwood F-test)
Ballard total score, n=25	31.7 (3.8; 24.0-40.0)	32.0 (28.0-34.0)	0.35 (-2.02 to 2.71)	0.95 (0.89-0.98)	0.21	0.22
Foot length, caliper, n=24	28.3 (3.3; 23.0-36.0)	29.0 (26.0-31.0)	3.78 (0.29 to 7.27)	0.57 (0.38-0.72)	0.46	<0.0001
End of bed assessment, n=21	31.1 (3.9; 26.0-39.0)	31.0 (27.0-34.0)	1.15 (-1.37 to 3.67)	0.90 (0.79-0.96)	0.02	0.003
AGA/LGA neonates (N=81)						
Early ultrasound (reference), n=81	33.7 (3.8; 25.0-41.1)	33.9 (30.6-37.0)	Ref	Ref	-	-
Last menstrual period (all), n=62	34.2 (4.6; 25.6-46.6)	34.2 (31.0-37.6)	-0.37 (-7.59 to 6.84)	0.60 (0.43-0.73)	-0.32	0.03
Ballard total score, n=81	34.0 (3.8; 26.0-42.0)	34.0 (32.0-36.0)	-0.29 (-3.14 to 2.56)	0.92 (0.88-0.95)	0.03	0.20
Foot length, caliper, n=81	32.4 (4.1; 25.0-41.0)	32.0 (29.0-35.0)	1.28 (-2.29 to 4.85)	0.85 (0.78-0.90)	-0.13	<0.0001
End of bed assessment, n=66	33.8 (3.6; 26.0-40.0)	34.0 (32.0-37.0)	-0.28 (-3.00 to 2.43)	0.93 (0.88-0.95)	0.07	0.22
LBW (<2500g) neonates (N=79)						

	GA (weeks) mean (SD; range)	GA (weeks) median (IQR)	Mean bias[1] (95% Limits of agreement)	Lin's concordance Rho[2] (95% CI)	Correlation: mean and difference[3]	*p*-value[4] (Bradley-Blackwood F-test)
Early ultrasound (reference), n=79	31.9 (3.4; 24.1-41.3)	32.7 (28.7-34.6)	Ref	Ref	-	-
Last menstrual period (all), n=61	33.5 (4.8; 23.0-46.6)	32.9 (29.1-36.6)	-1.23 (-8.32 to 5.86)	0.60 (0.43-0.72)	-0.44	<0.0001
Ballard total score, n=79	32.0 (3.3; 24.0-40.0)	32.0 (28.0-34.0)	-0.14 (2.80 to 2.52)	0.92 (0.88-0.95)	0.05	0.59
Foot length, caliper, n=78	29.7 (3.3; 23.0-36.0)	30.0 (27.0-32.0)	2.20 (-1.65 to 6.05)	0.68 (0.60-0.76)	0.10	<0.0001
End of bed assessment, n=67	31.8 (3.4; 26.0-39.0)	33.0 (28.0-34.0)	0.16 (-2.66 to 2.98)	0.91 (0.85-0.94)	-0.002	0.67
Normal birthweight (≥2500g) neonates (N=27)						
Early ultrasound (reference), n=27	37.3 (2.0; 32.3-41.1)	38.0 (36.4-38.4)	Ref	Ref	-	-
Last menstrual period (all), n=20	36.3 (3.9; 25.6-42.1)	37.7 (34.8-38.6)	0.91 (-5.61 to 7.42)	0.41 (0.09-0.66)	-0.62	0.01
Ballard total score, n=27	37.5 (2.1; 32.0-42.0)	38.0 (36.0-38.0)	-0.13 (-3.31 to 3.06)	0.69 (0.43-0.85)	-0.06	0.89
Foot length, caliper, n=27	36.5 (2.3; 31.0-41.0)	36.0 (35.0-38.0)	0.84 (-3.37 to 5.05)	0.48 (0.15-0.71)	-0.15	0.12

	GA (weeks) mean (SD; range)	GA (weeks) median (IQR)	Mean bias[1] (95% Limits of agreement)	Lin's concordance Rho[2] (95% CI)	Correlation: mean and difference[3]	*p*-value[4] (Bradley-Blackwood F-test)
End of bed assessment, n=20	37.4 (1.6; 33.0-40.0)	37.0 (37.0-38.0)	-0.26 (-3.48 to 2.95)	0.60 (0.26-0.81)	0.32	0.29
Very preterm (<32 weeks) neonates (N=36)						
Early ultrasound (reference), n=36	28.7 (1.7; 24.1-31.9)	28.7 (28.0-29.7)	Ref	Ref	-	-
Last menstrual period (all), n=26	30.3 (3.5; 23.0-41.4)	29.1 (28.6-31.4)	-1.33 (-7.03 to 4.36)	0.41 (0.16-0.60)	-0.69	<0.0001
Ballard total score, n=36	29.1 (2.0; 24.0-34.0)	28.0 (28.0-30.0)	-0.33 (-2.87 to 2.21)	0.74 (0.56-0.86)	2.36	0.11
Foot length, caliper, n=35	26.9 (2.1; 23.0-32.0)	27.0 (26.0-29.0)	1.77 (-1.27 to 4.81)	0.46 (0.25-0.62)	-0.25	<0.0001
End of bed assessment, n=31	28.8 (2.1; 26.0-33.0)	28.0 (27.0-30.0)	0.13 (-2.49 to 2.75)	0.74 (0.60-0.89)	-0.45	0.03
Term/late/moderate preterm neonates (N=70)						
Early ultrasound (reference), n=70	35.6 (2.3; 32.1-41.3)	35.0 (33.7-37.9)	Ref	Ref	-	-
Last menstrual period (all), n=55	36.0 (4.1; 25.6-46.6)	36.4 (32.9-38.6)	-0.40 (-8.12 to 7.31)	0.31 (0.10-0.50)	-0.54	<0.0001

	GA (weeks) mean (SD; range)	GA (weeks) median (IQR)	Mean bias[1] (95% Limits of agreement)	Lin's concordance Rho[2] (95% CI)	Correlation: mean and difference[3]	*p*-value[4] (Bradley-Blackwood F-test)
Ballard total score, n=70	35.7 (2.4; 32.0-42.0)	36.0 (34.0-38.0)	-0.04 (-2.94 to 2.87)	0.81 (0.72-0.89)	-0.06	0.86
Foot length, caliper, n=70	33.8 (3.0; 28.0-41.0)	34.0 (31.0-36.0)	1.89 (-2.66 to 6.44)	0.51 (0.36-0.63)	-0.34	<0.0001
End of bed assessment, n=56	35.5 (2.1; 31.0-40.0)	35.0 (34.0-37.0)	0.03 (-3.07 to 3.12)	0.74 (0.60-0.84)	0.15	0.53

Abbreviations: GA, gestational age at birth; SD, standard deviation; IQR, inter-quartile range; CI, confidence interval
[1]"Mean bias": average difference in estimated gestational age between gold standard (early ultrasound) and each comparator (US – comparator); therefore, a positive mean bias indicates underestimation by the comparator, a negative mean bias indicates overestimation; [2] Correlation coefficient for gold standard and comparator; [3] Closer to zero is required to meet model assumptions; greater correlation implies that the relationship between the gold standard and comparator method varies over different values of gestational age; [4] F-test simultaneously tests for significant differences between means and variances (non-significant p-value indicates concordance)

SUPPLEMENTARY TABLE 6. Repeatability of measures used in assessment of gestational age (interrater agreement): two assessors at a single time point

Measure	Assessor 2[1]	Assessor 1[1]	Correlation coefficient (95% CI)[2]	Mean bias	95% Limits of agreement (±2 SD)	Sample size
Ballard Score (total)	26 (18-32)	26 (18-33)	0.89 (0.85-0.93)	-0.15	-8.97 to 8.68	n=103
Ballard Score (physical)	13 (8-16)	13 (8-17)	0.88 (0.83-0.92)	-0.28	-5.74 to 5.19	n=102
Ballard Score (neuromuscular)	13 (10-16)	13 (10-16)	0.77 (0.68-0.84)	0.10	-5.54 to 5.73	n=103
Gestation based on Ballard overall score (wks)[3]	34 (30-36)	34 (30-36)	0.88 (0.83-0.92)	-0.04 weeks	-3.8 to 3.7 weeks	n=103
"End of bed" gestational age (wks)	33.5 (29-36)	34 (31-36)	0.91 (0.86-0.94)	-0.5 weeks	-3.8 to 2.8 weeks	n=84
Anterior lens assessment			0.47 (benchmark interval, 0.2 to 0.4)[4]	n/a	n/a	n=48[5]
Grade 0 (~≥35 weeks' gestation)	24/64 (37%)	31/69 (45%)				
Grade 1 (~33-34 weeks' gestation)	13/64 (20%)	13/69 (19%)				
Grade 2 (~31-32 weeks' gestation)	12/64 (19%)	4/69 (6%)				
Grade 3 (~29-30 weeks' gestation)	3/64(5%)	8/69 (12%)				
Grade 4 (~27-28 weeks' gestation)	12/64 (19%)	13/69 (19%)				
Head circumference (cm)	31.0 (27.3-32.6)	30.3 (28.0-32.5)	0.93 (0.89-0.95)	0.2 cm	-2.5 to 2.9 cm	n=90

Mid-upper arm circumference (cm)	8.0 (6.7-9.8)	7.4 (6.2-9.2)	0.85 (0.78-0.89)	0.6 cm	-1.4 to 2.6 cm	n=98
Abdominal circumference (cm)	25.5 (22.5-29.0)	25.4 (22.4-29.0)	0.93 (0.90 – 0.95)	0.3 cm	-2.9 to 3.6 cm	n=98
Foot length by ruler (mm)	65.0 (57.0-74.0)	65.0 (56.5–72.0)	0.96 (0.94-0.97)	0.65 mm	-5.0 to 6.3mm	n=102
Foot length by calipers (mm)	64.5 (56.0-72.0)	63.5 (55.2-71.5)	0.97 (0.95-0.98)	0.45 mm	-4.7 to 5.6mm	n=101

Abbreviations: SD, standard deviation; CI, confidence interval; cm, centimeters; wks, weeks

[1] Continuous variables are summarized as median (interquartile range, IQR), categorical variables with n (column percentage); [2] Continuous measures' correlation expressed as Lin's correlation; categorical measures correlation reflect Krippendorf's alpha coefficient with probabilistic interval; [3] gestation calculated in 2-week intervals as per Ballard score recommendations; [4]probabilistic interval of 0.2 to 0.4 interpretable as "fair" agreement; [5] sample size of neonates with assessments by both assessors

SUPPLEMENTAL TABLE 7. Diagnostic accuracy of total Ballard score: to identify neonates <37 weeks' gestation, using Ballard score threshold <30

		GOLD STANDARD ALLOCATION OF GESTATIONAL AGE		
		< 37 weeks by ultrasound N=67	**≥ 37 weeks by ultrasound N=39**	**TOTAL N=106**
GESTATIONAL AGE ALLOCATION BY INDEX TEST (BALLARD SCORE)	**< 37 weeks by Ballard score (indicated by total score <30) N=83**	66/83 (79.5%)	17/83 (20.5%)	N=83 (83/106 = 78.3%)
	≥37 weeks by Ballard score (indicated by total score ≥30) N=23	1/23 (4.3%)	22/23 (95.7%)	N=23 (23/106 = 21.7%)
	Total N=106	N=67 (67/106 = 63%)	N=39 (39/106=37%)	**N=106**

[1]Obtained by average score from two assessors

SUPPLEMENTAL TABLE 8. Diagnostic accuracy of postnatal clinical methods to identify moderately preterm infants (<34 weeks) and extremely preterm infants (< 28 weeks) at different clinical cut-off values and pre-test probabilities (underlying prevalence estimates, sample-specific and hypothetical)

Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
MODERATE/LATE PRETERM, < 34 weeks								
Ballard total overall score	Score < 30	96.4 (87.7-99.6)	74.0 (59.7-85.4)	3.71 (2.32-5.94)	0.05 (0.01-0.19)	53% (43-63%)	80.6 (69.1-89.2)	94.9 (82.7-99.4)
						10% *(hypothetical)*	29.2 (20.5-39.7)	99.5 (97.9-99.9)
						50% *(hypothetical)*	78.8 (69.9-85.6)	95.4 (84.0-98.8)
	Score < 25	75.0 (61.6-85.6)	98.0 (89.4-99.9)	37.50 (5.36-263.0)	0.25 (0.16-0.40)	53% (43-63%)	97.7 (87.7-99.9)	77.8 (65.5-87.3)
						10% *(hypothetical)*	80.6 (37.3-96.7)	97.2 (95.7-98.2)
						50% *(hypothetical)*	97.4 (84.3-99.6)	79.7 (71.3-86.1)

Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
Ballard total neuro-muscular score	Score < 14	82.1 (69.6-91.1)	88.0 (75.7-95.5)	6.85 (3.20-14.60)	0.20 (0.12-0.36)	53% (43-63%)	88.5 (76.6-95.6)	81.5 (68.6-90.7)
						10% *(hypothetical)*	43.2 (26.2-61.9)	97.8 (96.2-98.7)
						50% *(hypothetical)*	87.3 (76.2-93.6)	83.1 (73.6-89.7)
Ballard total external (physical) score	Score < 14	89.3 (78.1-96.0)	86.0 (73.3-94.2)	6.38 (3.19-12.80)	0.13 (0.06-0.27)	53% (43-63%)	87.7 (76.3-94.9)	87.8 (75.2-95.4)
						10% *(hypothetical)*	41.5 (26.2-58.6)	98.6 (97.1-99.4)
						50% *(hypothetical)*	86.4 (76.1-92.7)	88.9 (78.9-94.5)
Birthweight	< 2000g	89.3 (78.1-96.0)	78.0 (64.0-88.5)	4.06 (2.39-6.89)	0.14 (0.06-0.30)	53% (43-63%)	82.0 (70.2-90.6)	86.7 (73.2-94.9)
						10% *(hypothetical)*	31.1 (21.0-43.4)	98.5 (96.8-99.3)
						50% *(hypothetical)*	80.2 (70.5-87.3)	87.9 (77.1-94.0)

Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
	< 1800g	83.9 (71.7-92.4)	84.0 (70.9-92.8)	5.25 (2.75-10.00)	0.19 (0.10-0.35)	53% (43-63%)	85.5 (73.3-93.5)	82.4 (69.1-91.6)
						10% *(hypothetical)*	36.8 (23.4-52.6)	97.9 (96.2-98.9)
						50% *(hypothetical)*	84.0 (73.3-90.9)	83.9 (73.9-90.6)
Head circumference	< 31cm	81.1 (68.0-90.6)	80.4 (66.1-90.6)	4.15 (2.28-7.56)	0.23 (0.13-0.42)	53% (43-64%)	82.7 (69.7-91.8)	78.7 (64.3-89.3)
						10% *(hypothetical)*	31.5 (20.2-45.6)	97.5 (95.6-98.6)
						50% *(hypothetical)*	80.6 (69.5-88.3)	81.0 (70.6-88.4)
	< 30cm	79.2 (65.9-89.2)	89.1 (76.4-96.4)	7.29 (3.15-16.90)	0.23 (0.14-0.40)	53% (43-64%)	89.4 (76.9-96.5)	78.8 (65.3-88.9)
						10% *(hypothetical)*	44.8 (25.9-65.2)	97.5 (95.8-98.5)
						50% *(hypothetical)*	87.9 (75.9-94.4)	81.1 (71.5-88.0)

Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
Foot length, calipers	< 70mm	94.5 (84.9-98.9)	58.0 (43.2-71.8)	2.25 (1.62-3.14)	0.09 (0.03-0.29)	53% (42-62%)	71.2 (59.4-81.2)	90.6 (75.0-98.0)
						10% *(hypothetical)*	20.0 (15.2-25.8)	99.0 (96.9-99.7)
						50% *(hypothetical)*	69.2 (61.8-75.8)	91.4 (77.5-97.0)
	< 65mm	87.3 (75.5-94.7)	80.0 (66.3-90.0)	4.36 (2.48-7.67)	0.16 (0.08-0.32)	53% (43-63%)	82.8 (70.6-91.4)	85.1 (71.7-93.8)
						10% *(hypothetical)*	32.7 (21.6-46.0)	98.3 (96.5-99.1)
						50% *(hypothetical)*	81.4 (71.3-88.5)	86.3 (75.6-92.7)
EXTREMELY PRETERM, < 28 weeks								
Ballard total overall score	Score < 15	87.5 (47.3-99.7)	86.7 (78.4-92.7)	6.60 (3.73-11.70)	0.14 (0.02-0.90)	8% (3-14%)	35.0 (15.4-59.2)	98.8 (93.7-100.0)
						10% *(hypothetical)*	42.3 (29.3-56.4)	98.4 (90.9-99.7)
						50% *(hypothetical)*	86.8 (78.9-92.1)	87.4 (52.6-97.8)

Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
	Score < 10	25.0 (3.2-65.1)	99.0 (94.4-100.0)	24.50 (2.48-242.0)	0.76 (0.51-1.13)	8% (3-14%)	66.7 (9.4-99.2)	94.2 (87.8-97.8)
						10% *(hypothetical)*	73.1 (21.6-96.4)	92.2 (88.8-94.7)
						50% *(hypothetical)*	96.1 (71.3-99.6)	56.9 (46.9-66.3)
Ballard total neuro-muscular score	Score < 10	87.5 (47.3-99.7)	81.6 (72.5-88.7)	4.76 (2.91-7.80)	0.15 (0.02-0.96)	8% (3-14%)	28.0 (12.1-49.4)	98.8 (93.3-100.0)
						10% *(hypothetical)*	34.6 (24.4-46.4)	98.3 (90.4-99.7)
						50% *(hypothetical)*	82.7 (74.4-88.6)	86.7 (51.0-97.6)
Ballard total external (physical) score	Score < 6	87.5 (47.3-99.7)	88.8 (80.8-94.3)	7.8 (4.21-14.4)	0.14 (0.02-0.88)	8% (3-14%)	38.9 (17.3-64.3)	98.9 (93.8-100.0)
						10% *(hypothetical)*	46.4 (31.9-61.6)	98.5 (91.1-99.8)
						50% *(hypothetical)*	88.6 (80.8-93.5)	87.7 (53.1-97.8)

Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
Birthweight	< 1000g	87.5 (47.3-99.7)	89.8 (82.0-95.0)	8.57 (4.51-16.30)	0.14 (0.02-0.87)	8% (3-14%)	41.2 (18.4-67.1)	98.9 (93.9-100.0)
						10% *(hypothetical)*	48.8 (33.4-64.4)	98.5 (91.2-99.8)
						50% *(hypothetical)*	89.6 (81.8-94.2)	87.8 (53.4-97.8)
	< 800g	37.5 (8.52-75.5)	94.9 (88.5-98.3)	7.35 (2.13-25.3)	0.66 (0.38-1.13)	8% (3-14%)	37.5 (8.52-75.5)	94.9 (88.5-98.3)
						10% *(hypothetical)*	45.0 (19.2-73.8)	93.2 (88.9-95.9)
						50% *(hypothetical)*	88.0 (68.1-96.2)	60.3 (47.0-72.2)
Head circumference	< 26cm	85.7 (42.1-99.6)	89.1 (80.9-94.7)	7.89 (4.08-15.2)	0.16 (0.03-0.98)	7% (3-14%)	37.5 (15.2-64.6)	98.8 (93.5-100.0)
						10% *(hypothetical)*	46.7 (31.2-62.9)	98.3 (90.1-99.7)
						50% *(hypothetical)*	88.7 (80.3-93.8)	86.2 (50.4-97.5)

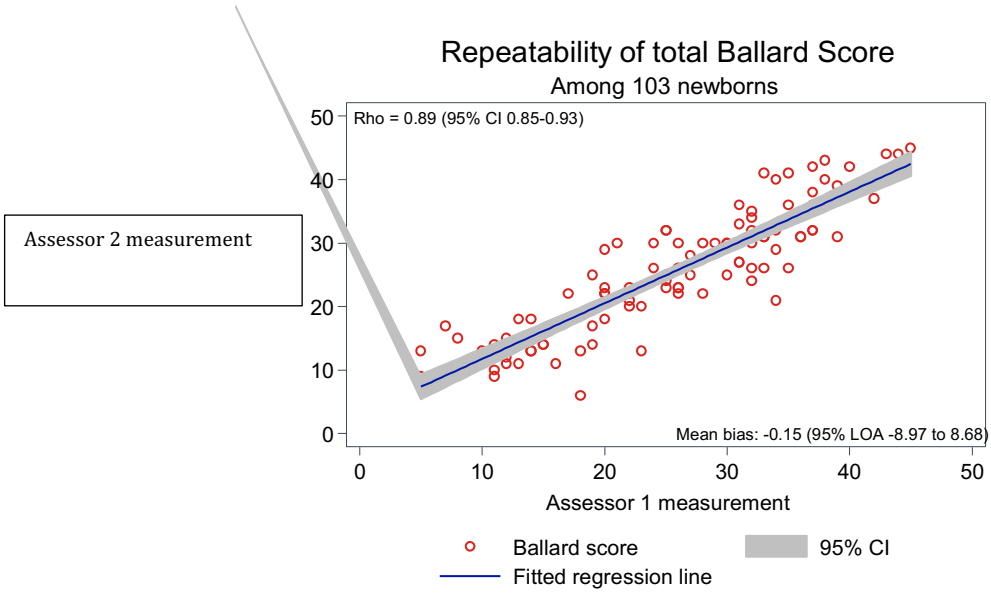
Method	Boundary	Sensitivity, % (95% CI)	Specificity, % (95% CI)	LRT+ (95% CI)	LRT- (95% CI)	Prevalence % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)
Foot length, calipers	< 52mm	100.0 (63.1-100.0)	91.8 (84.4-96.4)	12.1 (6.24-23.5)	0	8% (3-14%)	50.0 (24.7-75.3)	100.0 (-)
						10% *(hypothetical)*	57.4 (41.0-72.3)	100.0 (-)
						50% *(hypothetical)*	92.4 (86.2-95.9)	100.0 (-)
	< 50mm	50.0 (15.7-84.3)	95.9 (89.8-98.9)	12.10 (3.71-39.60)	0.52 (0.26-1.04)	8% (3-14%)	50.0 (15.7-84.3)	95.9 (89.8-98.9)
						10% *(hypothetical)*	57.4 (29.2-81.5)	94.5 (89.6-97.2)
						50% *(hypothetical)*	92.4 (78.8-97.5)	65.7 (48.9-79.3)

Abbreviations: CI, confidence interval; sensitivity = Pr(Test+|Disease+); specificity = Pr(Test-|Disease-); LRT, likelihood ratio (interpretation: +LR, positive likelihood ratio= true positives/false positives; -LRT, negative likelihood ratio=false negatives/true negatives); PPV, positive predictive value = Pr(Disease+|Test+), using the given prevalence as pretest probability; NPV, negative predictive value = Pr(Disease-|Test-), using the given prevalence as pretest probability
Prevalence estimates: (a) actual prevalence in sample, with 95% CI; (b) hypothetical prevalence of 10%; (c) hypothetical prevalence of 50%

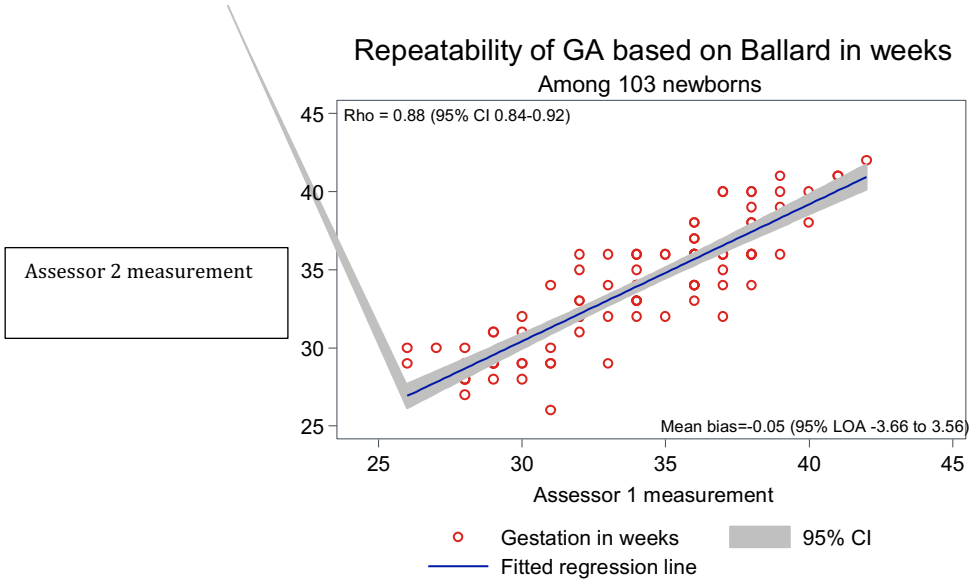
Supplemental Figures

SUPPLEMENTAL FIGURE 1. Repeatability of continuous measures: Scatter plots with regression lines, comparing two assessors at a single time point

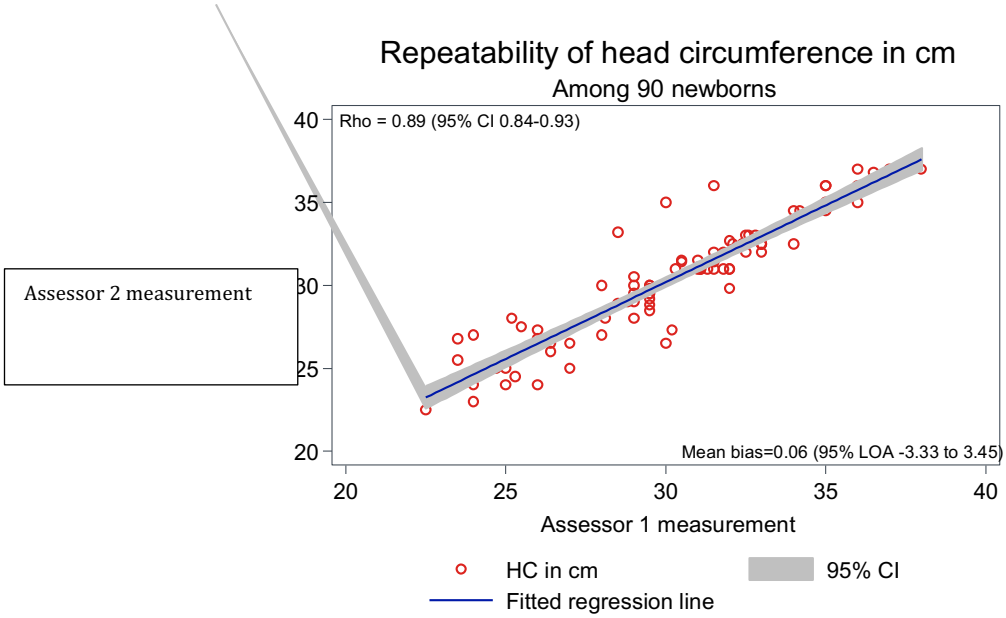
S Figure 1 (a) Interrater agreement for total Ballard score



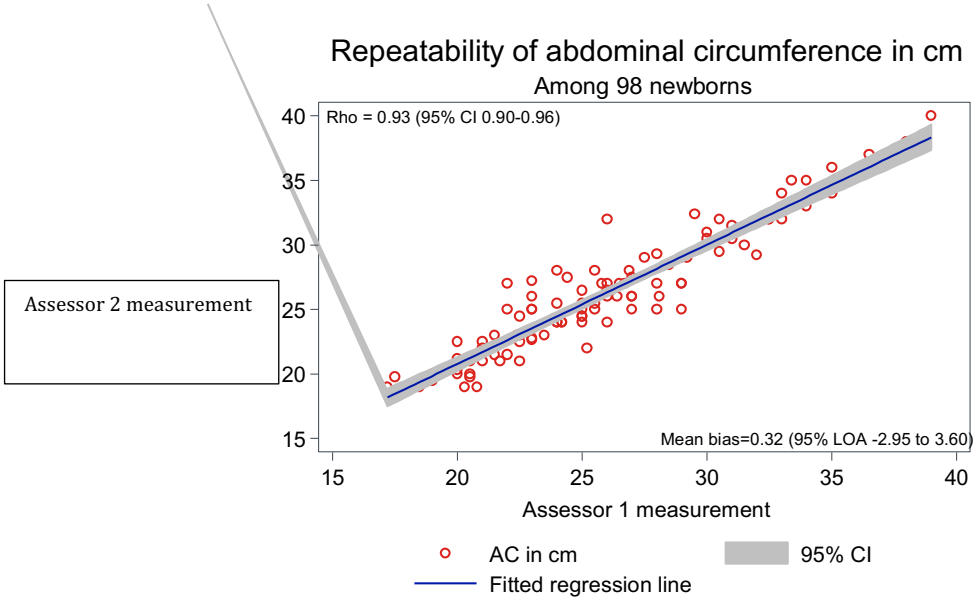
S Figure 1 (b) Interrater agreement for gestational assessment based on Ballard score (weeks)



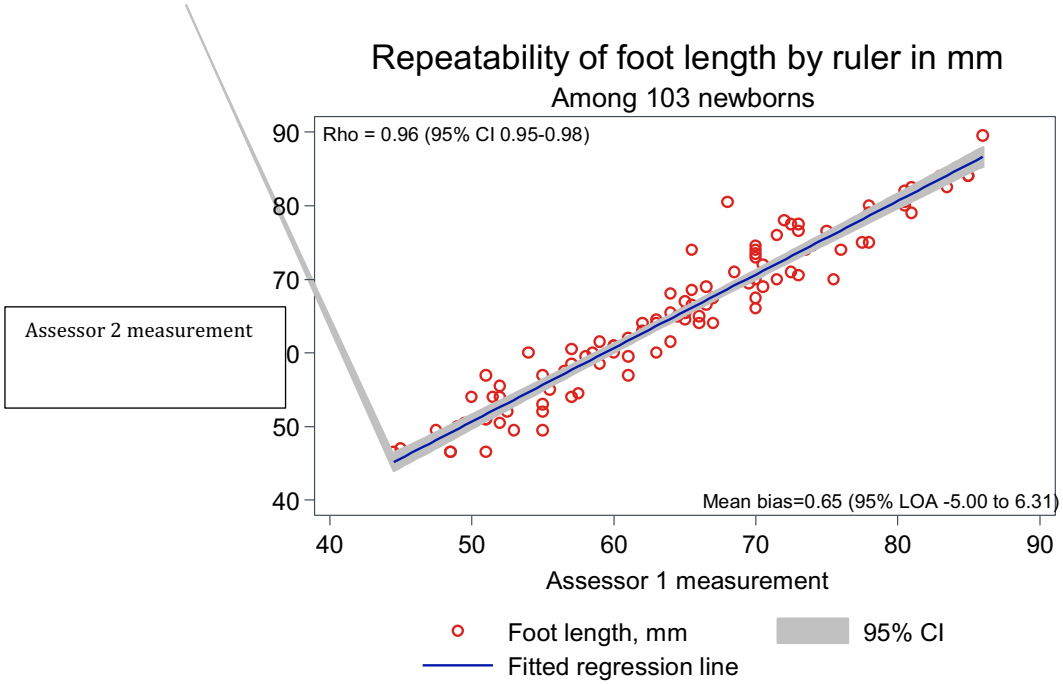
S Figure 1 (c) Interrater agreement for head circumference (cm)



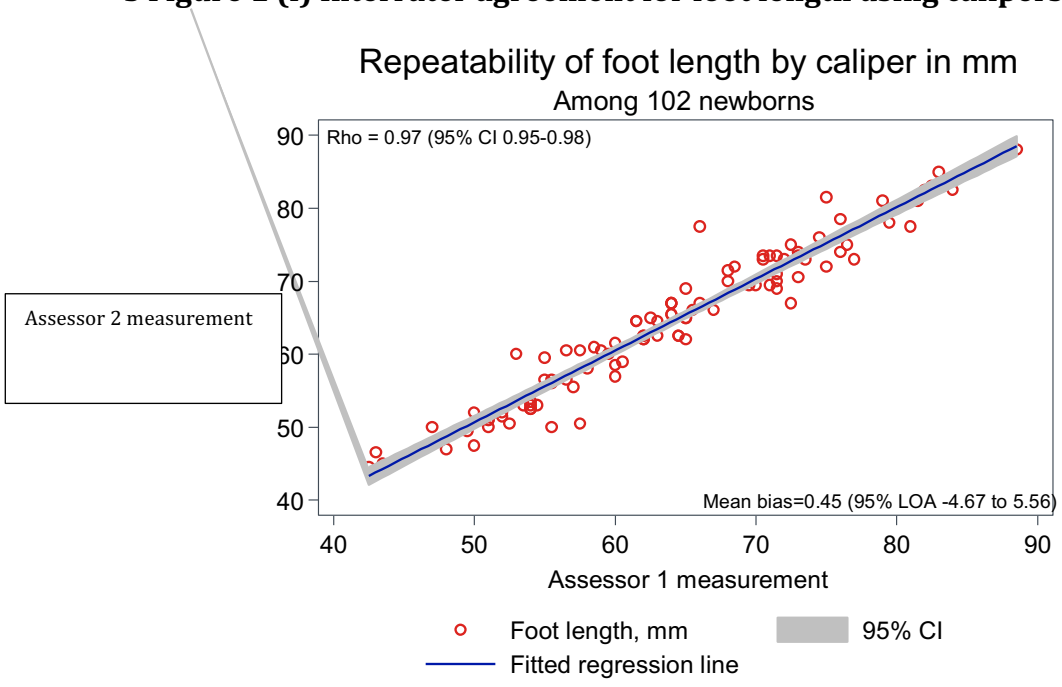
S Figure 1 (d) Interrater agreement for abdominal circumference (cm)



S Figure 1 (e) Interrater agreement for foot length using ruler (mm)

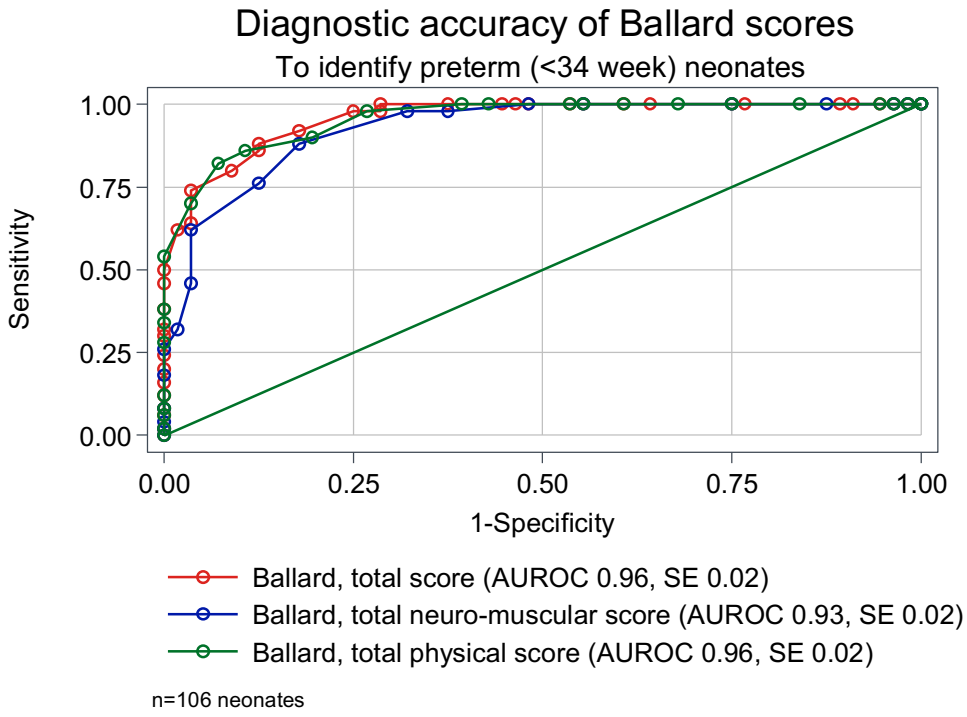


S Figure 1 (f) Interrater agreement for foot length using calipers (mm)

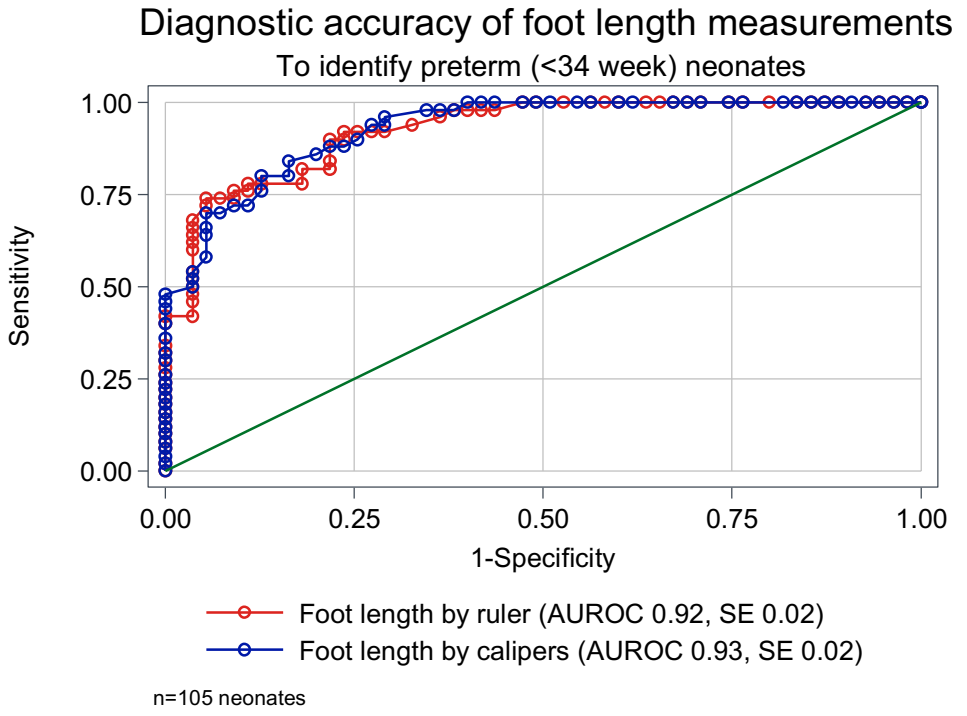


SUPPLEMENTAL FIGURE 2. Receiver operating curves for measures to identify neonates ≥ 34 vs < 34 weeks' gestational age

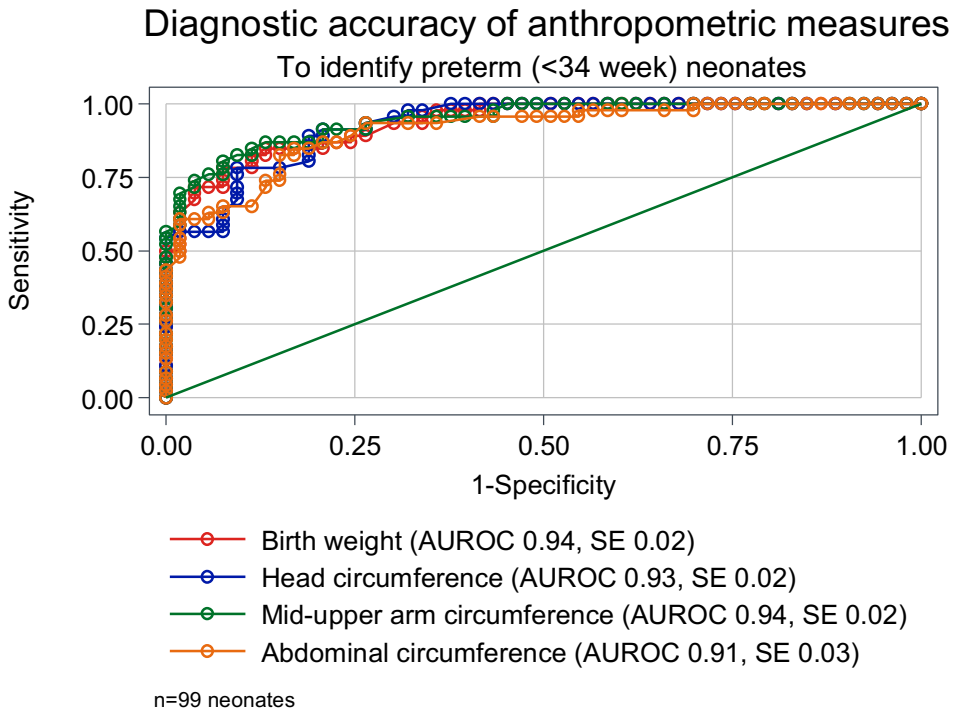
S Figure 2 (a) Ballard scores



S Figure 2 (b) Foot length

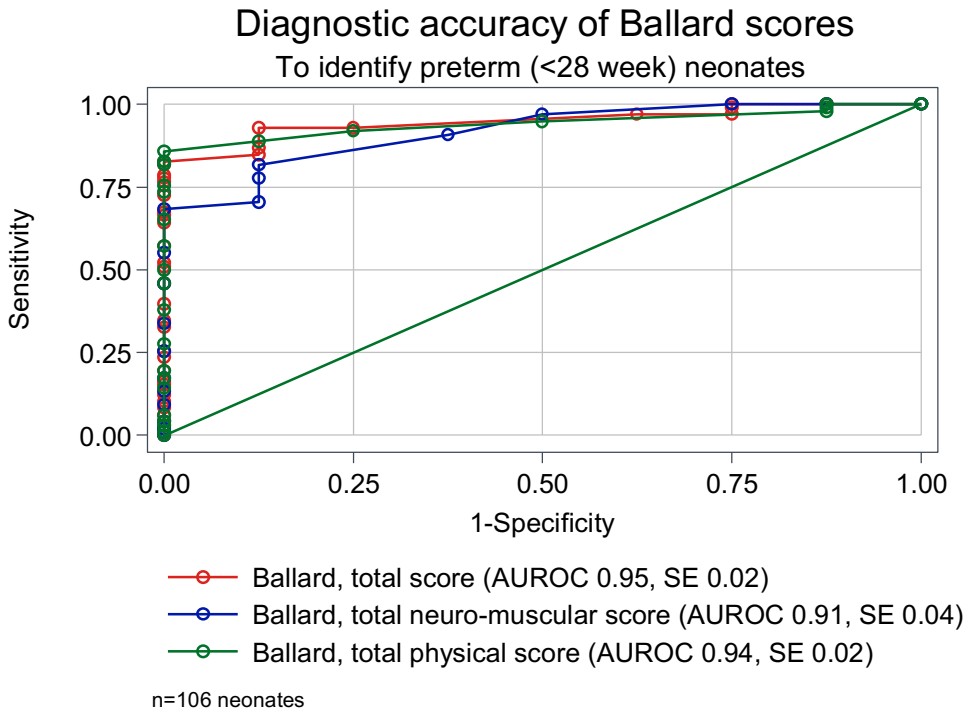


S Figure 2 (c) Anthropometric measures

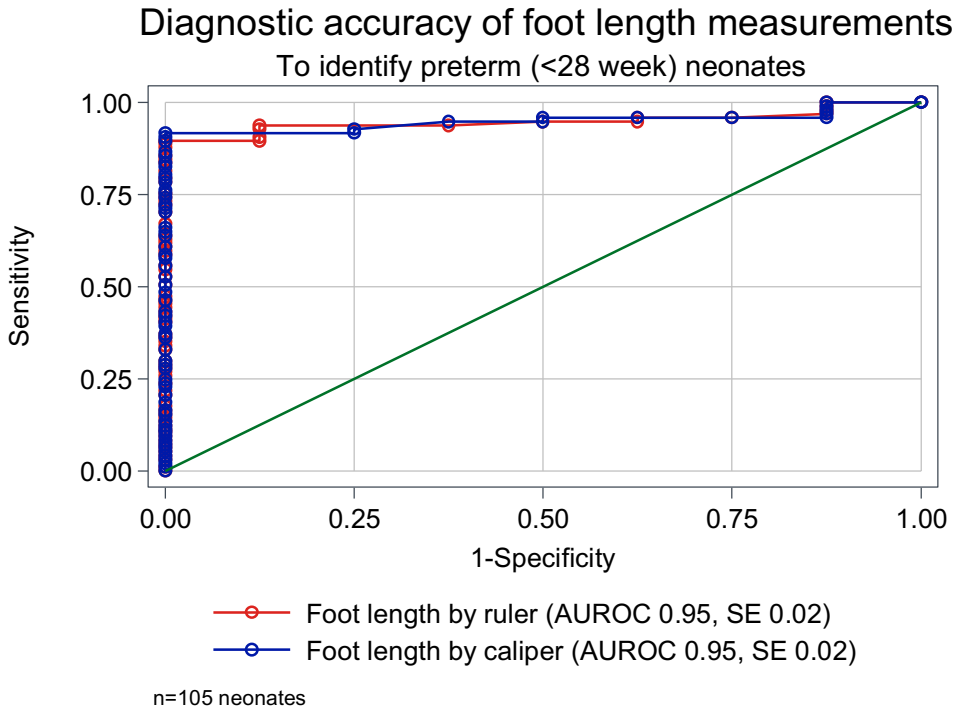


SUPPLEMENTAL FIGURE 3. Receiver operating curves for measures designed to identify neonates ≥ 28 vs <28 weeks' gestational age

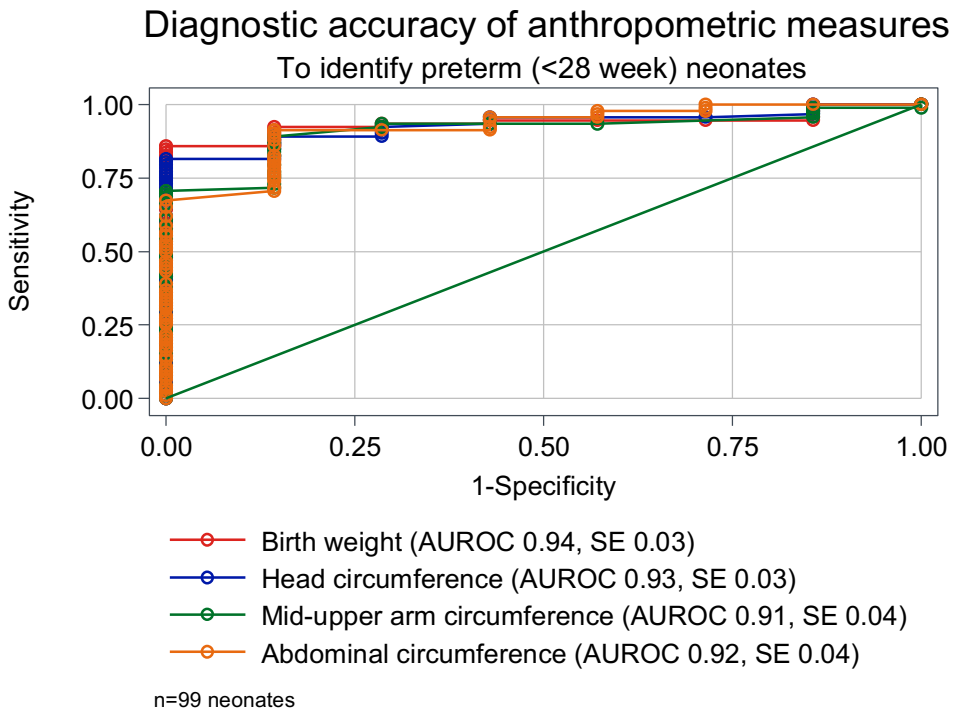
S Figure 3 (a) Ballard scores



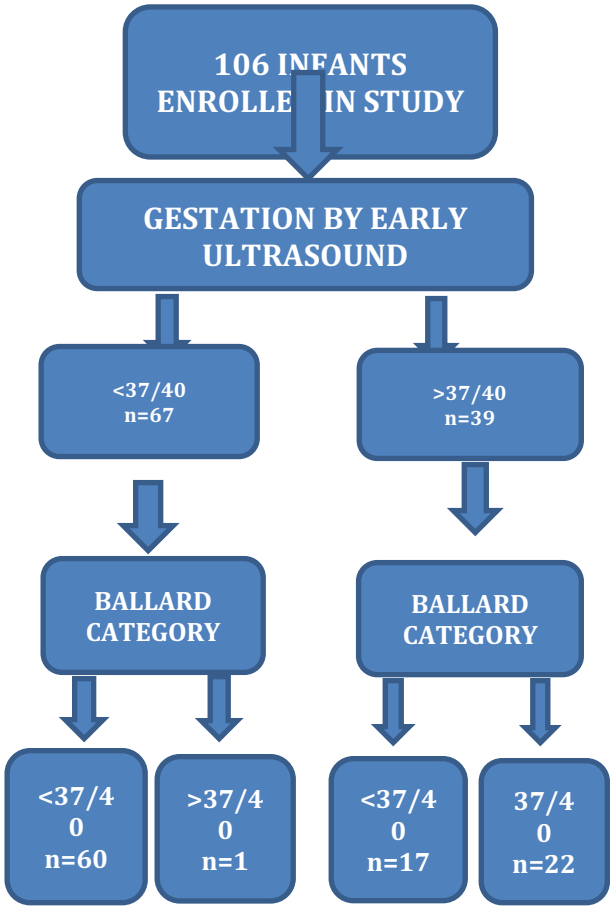
S Figure 3 (b) Foot length



S Figure 3 (c) Anthropometric measures



SUPPLEMENTAL FIGURE 4. Study Flow Diagram



Appendix 1

Instructions to Authors from Journal of Tropical Paediatrics Website:

Preparation of manuscripts

Manuscripts should be legibly typed, using double spacing throughout, with 25 mm margins at each side. Please note the following word count, this does not include the title page, abstract and references:

Original paper - 3000
Brief report - 1000
Case report -1000
Research Letter - 500
Clinical Review 3000
Book review - 500
Editorials - 1000
News from Region - 1000

Regular full length papers should be divided into the following sequence of sections, and each section should begin on a new page:

Title page
Summary
Text
Acknowledgements
References
Legends to figures
Tables.

Number each page at the top right corner consecutively, beginning with the title page. Please avoid footnotes; use instead, parentheses within brackets. Underline only words which should appear in italic. Clearly identify unusual or handwritten symbols and Greek letters. Differentiate between the letter O and zero, and the letters I and l and number 1. Mark the position of each figure and table in the margin. SI units should be used for scientific measurements.

References

Number references consecutively in the order in which they are cited in the text. Published articles and those in press (state the journal which has accepted them) may be included. References should include (in the following order) author's names, editors (books only) paper title in full, journal/book title, name and address of publisher (books only), year, volume number and inclusive page numbers. Personal communication should be authorized by those involved, in writing, and unpublished data should be cited as (unpublished data). Papers in preparation or submitted for publication should not be in the reference list. They should be cited in the text as follows: H. G. Jones, unpublished results/submitted for publication/in preparation (as appropriate).

Style in the reference section should be as follows:

1. Kennedy T, Jones R. Effect of obesity on esophageal transit. Am J Surg 1985;149:177–81.
2. Long HC, Blatt MA, Higgins MC et al.. Medical Decision Making. Boston: Butterworth-Heinemann, 1997. Full text - International Conference on Natural Orifice Transluminal Endoscopic Surgery
3. Manners T, Jones R, Riley M. Relationship of overweight to haitus hernia and reflux oesophagitis. In: Newman W (ed). The Obesity Conundrum. Amsterdam: Elsevier Science, 1997,352–74.
4. Hou Y, Qiu Y, Vo NH et al. 23-O derivatives of OMT: highly active against H. influenzae. In: Programs and Abstracts of the Forty-third Interscience Conference on Antimicrobial Agents and Chemotherapy, Chicago, IL, 2003. Abstract F-1187, p.242. American Society for Microbiology, Washington, DC, USA.

5. Public Health Laboratory Service. Antimicrobial Resistance in 2000: England and Wales. http://www.hpa.org.uk./infections/topics_az/antimicrobial_resistance/amr.pdf (7 January 2004, date last accessed).

Tables

Tables should be typed on separate sheets, and numbered consecutively. Tables should be self-explanatory and include a brief descriptive title. Footnotes to tables indicated by lower case letters are acceptable, but they should not include extensive experimental detail. Cite each table in the text in consecutive order.

Illustrations

All illustrations must be cited in the text in consecutive order. Each figure should be labelled clearly with the figure number. Also indicate clearly the top margin of the figure. Figures should be submitted in the desired final size so that reduction can be avoided. The type area of a page is 206 (height) mm x 150 mm (width); a single column is 71 mm (width). Figures must be at a minimum resolution of 600 d.p.i. for line drawings (black and white) and 300 d.p.i. for colour or greyscale.

Photographs - Photographs should be of sufficiently high quality with respect to detail, contrast, and fineness of grain to withstand the inevitable loss of contrast and detail inherent in the printing process. Indicate the magnification by a rule on the photographs.

Line drawings - These should be clear, sharp prints, suitable for reproduction as submitted. Ensure that the size of lettering is in proportion with the overall dimensions of the figure.

Figure legends - These should be added at the bottom of the manuscript in numbered order. Define all symbols and abbreviations used in the figure.

Lay summary

Authors of all article types are encouraged to submit a lay summary as part of the article, in addition to the main text abstract. The lay summary should clearly summarize the focus and findings of the article for non-expert readers, and will be published as part of the article online and in PDF. The lay summary should be submitted for peer review as part of the main manuscript file, under the heading ‘Lay summary’, before the article’s main text. The lay

summary should be no longer than 200 words. As with a main abstract, avoid citations and define any abbreviations.

Funding

Details of all funding sources for the work in question should be given in a separate section entitled 'Funding'. This should appear before the 'Acknowledgements' section.

The following rules should be followed:

- The sentence should begin: 'This work was supported by ...'
- The full official funding agency name should be given, i.e. 'National Institutes of Health', not 'NIH' (full RIN-approved list of UK funding agencies) Grant numbers should be given in brackets as follows: '[grant number xxxx]'
- Multiple grant numbers should be separated by a comma as follows: '[grant numbers xxxx, yyyy]'
- Agencies should be separated by a semi-colon (plus 'and' before the last funding agency)
- Where individuals need to be specified for certain sources of funding the following text should be added after the relevant agency or grant number 'to [author initials]'.

An example is given here: 'This work was supported by the National Institutes of Health [AA123456 to C.S., BB765432 to M.H.]; and the Alcohol & Education Research Council [hfygr667789].

Oxford Journals will deposit all NIH-funded articles in PubMed Central. See Depositing articles in repositories – information for authors for details. Authors must ensure that manuscripts are clearly indicated as NIH-funded using the guidelines above.

Crossref funding data registry

In order to meet your funding requirements authors are required to name their funding sources, or state if there are none, during the submission process. Information on the CHORUS initiative.

Language editing

Language editing, if your first language is not English, to ensure that the academic content of your paper is fully understood by journal editors and reviewers, is optional. Language editing does not guarantee that your manuscript will be accepted for publication. Information on the language editing service. Several specialist language editing companies offer similar services and you can also use any of these. Authors are liable for all costs associated with such services.

Submission appeals/Editorial concerns

The mechanism to appeal editorial decisions and/or express concerns about the editorial process is to write directly to the editorial office at tropej.editorialoffice@oup.com

Research misconduct

The journal runs the iThenticate plagiarism detection software program on all manuscripts sent for peer-review and will reject any submissions found to have any similarities between previously published pieces. The publication handles allegations of research misconduct in accordance with the Committee on Publication Ethics (COPE) flowcharts.

Author Self-Archiving Policy

For information about this journal's policy, please visit our Author Self-Archiving policy page .

Availability of Data and Materials

Where ethically feasible, the *Journal of Tropical Pediatrics* strongly encourages authors to make all data and software code on which the conclusions of the paper rely available to readers. We suggest that data be presented in the main manuscript or additional supporting files, or deposited in a public repository whenever possible. For information on general repositories for all data types, and a list of recommended repositories by subject area, please see Choosing where to archive your data.

Data Citation

The *Journal of Tropical Pediatrics* supports the Force 11 Data Citation Principles and requires that all publicly available datasets be fully referenced in the reference list with an accession number or unique identifier such as a digital object identifier (DOI). Data citations should include the minimum information recommended by DataCite:

- [dataset]* Authors, Year, Title, Publisher (repository or archive name), Identifier

*The inclusion of the [dataset] tag at the beginning of the citation helps us to correctly identify and tag the citation. This tag will be removed from the citation published in the reference list.

ORCID

Journal of Tropical Pediatrics requires submitting authors to provide an ORCID iD at submission to the journal. More information on ORCID and the benefits of using an ORCID iD is available. If you do not already have an ORCID iD, you can register for free via the ORCID website.

Preprint Policy

Authors retain the right to make an Author's Original Version (preprint) available through various channels, and this does not prevent submission to the journal. For further information see our Online Licensing, Copyright and Permissions policies. If accepted, the authors are required to update the status of any preprint, including your published paper's DOI, as described on our Author Self-Archiving policy page.

Appendix 2

Reviewers Comments from initial Submission

Title: Manuscript ID JTP-2020-228-OP - A comparison of the accuracy of various methods of postnatal gestational age estimates, including Ballard score, Foot-length, Vascularity of the Anterior Lens, Last Menstrual Period and a clinician's non structured estimate. Version:1

Date: 22 June 2010 Reviewer's Report Thank you for the opportunity to review this manuscript in which the authors attempt to assess the validity of alternative and less-expensive methods of gestational age and preterm birth assessment in a peri-urban hospital in South Africa. General comments: While I appreciate that writing style differs, I feel that the manuscript currently reads as a 'collection of facts' that do not culminate in justification for this study. The introduction could be improved upon.

There are already so many studies investigation so different methods of GA assessment/ maturity in newborns. The authors need to therefore show why their study is important – e.g. is it due to a lack of population (country)-specific data The authors present a lot of data in the results that is difficult to follow. I also find the tables crowded with a lot of information. The discussion should begin with a statement on the key findings of the study in relation to the study objectives. I have not checked the journal requirements, but most biomedical journals do not permit sub-headings in the discussion.

The discussion mostly repeats the results without critically reviewing the most clinically significant positive or negative findings from this study in relation to findings from similar studies in the region or other LMICs.

Specific Major Comments A) Introduction 1. "Improving methods of ascertaining gestational age has been identified as a research priority by the World Health Organisation (3). In developing countries relatively few mothers have access to early (dating) ultrasounds(3)." This selected reference (3) does not apply to any of the statements for which it has been used. It is also from the year 2000 which makes it rather old for an issue which is considered a research priority.
2. "The accuracy of postnatal gestational age estimates has been extensively studied in the era before early ultrasound scanning became routine." Please include appropriate reference(s) here.
3. "These studies were at high risk of bias and confounding as the methods of gestational age estimate were evaluated against an imperfect standard: the last menstrual period." Again, please include reference(s) here appropriate to the validity of the LMP
4. "It also categorises which babies need admission to a neonatal unit, which babies need methylxanthines to prevent apnoea of prematurity" – reference please
5. References (5) and (6) are the same.
6. "The paper highlighted the need to specifically look at SGA babies and further recommended research into the use of lens vascularity as a potential method (6) This second part of this statement is inaccurate. The authors of the referenced systematic review state : "An important consideration is that the avascular capsule of the lens (AVCL) completely disappears after ~34 weeks' GA; thus, it may not help with GA dating >34 weeks.

Furthermore, the AVCL exam requires specialized skills with an ophthalmoscope, which may limit the feasibility and scalability in LMIC."
7. "We used an ultrasound scan before 20 weeks as the clinical reference standard in this prospective, hospital based study." This statement belongs to the methods section

B) Methods
1. Please check reference (9) it is incomplete.
2. More information on the study setting would improve understand of the context. How many deliveries per year in the hospital? What is the standard practice for fetal USS assessment during ANC? Do you have any idea of what proportion of women attend ANC and have an USS done? What are the inpatient neonatal services? Was neonatal GA assessment the standard practice before the study? If yes, who usually performs this?
3. Convenience sampling was used aiming to recruit every consecutive eligible baby who was identified when the researchers were present on the unit six days a week. Why was convenience sampling used? Which unit do the authors refer to?
4. Mothers with newborn infants aged less than 48 hours were invited to participate in the study if they had had a dating ultrasound scan performed before 20 weeks of gestation and had either delivered at Groote Schuur, or were transferred there in time for the baby to be examined before 48 hours of age. I'm unable to follow the process of recruitment. Would the mothers have a copy of the USS report and be expected to bring it along with them or would this information be on the antenatal card?
5. If the authors used an ultrasound scan before 20 weeks as the clinical reference standard in this Study, it is important that the USS have been performed reliably in the first instance. For the women who were transferred to Groote Schur, must the USS have been performed at Groote Schur or does this not matter?
6. In accordance with international guidelines, we assigned early ultrasound-based estimation of gestational age (GA) at birth as the clinical reference standard for this study (8,10). The guidelines referenced are the Canadian (8) and American (9) guidelines. Is there any particular reason why these were chosen above others?
7. last menstrual period (LMP; overall and limited to women who were confident of the last date) – one of the limitations in the use of LMP is recall bias. How confident were the authors themselves of the women's confidence? Did these women have documented evidence of there LMP (e.g journal??)
8. Please spell out FLC and FLR in full at first use
9. Maternal interviews, data abstraction and clinical assessments were completed by study staff, consisting of three junior medical clinicians who had completed their internship and had been working in the neonatal unit for at least six months. Specialized neonatologists provided structured, standardized training in foot-length measurement, Ballard scoring and anterior lens assessment.
10. Given that you have numerous definitions (e.g. SGA, AGA/LGA etc), it would be in order to have a sub-section on definitions.
11. Lens assessment requires specialized skills with an ophthalmoscope, and this skill is not acquired overnight. How long was the training by the neonatologists? Did any of the neonatologists carry out an independent assessment of the lenses of study participants as a quality control measure
12. "…..using non-stretchable measuring tape following standardized procedures." What do the authors mean by "standardized procedures"?
13. "Finally, study clinicians assessed the vascularity of anterior lens (anterior lens analysis, ALA) using direct ophthalmoscopy, without dilatation. Grading of lens vascularity was based on previously published guidelines(13)." The authors of the reference paper (Hirsch et at.)

dilated the pupils of study participants. How were the study clinicians able to adequately visualise and grade the lens without dilatation?
14. Something is missing here: "….gestational age by ultrasound was dichotomized into (a) ≥37 vs 10 provide stronger evidence of the presence of a diagnosis (here, being preterm < 34 weeks' gestation (supplemental figure 2, supplemental table 3)…." What do the authors mean by 'moderate to good diagnostic performance'?
D) Results 1. The vascularity completely disappears after 35 weeks of gestation, so the lens vascularity is not a useful tool for differentiating preterm from term infants. This is already known in the literature and so I do not see the rationale for including it in the study or its usefulness.
2. However, despite using junior clinicians, interrater variability was low for most of the clinical measures, suggesting that our findings may be generalisable to non-hospital settings provided junior clinicians have access to adequate training. I disagree with this viewpoint. The authors did not compare the findings by the junior doctors which those by experienced senior clinicians/neonatologists.
3. High interrater variability in ALA suggested a high risk of misclassification bias, thereby precluding meaningful evaluation of diagnostic accuracy for this potentially useful clinical measure. This statement appears to contradict an earlier statement (see Discussion comment 1 above); kindly explain the potential usefulness of lens examination.

Level of interest: An article whose findings are important to those with closely related research interests
Quality of written English: Acceptable Statistical review: A statistical review is needed.
Declaration of competing interests: I declare that I have no competing interest

Appendix 3
Data Acquisition Tool

Data Acquisition Form

Researcher initials

Date……..
DOB…… Time……..
Mother's Name and hospital
number……..
1st Ultrasound: At what gestation was the ultrasound done? ……GA at birth?
……
LMP of mother……
How sure of her dates is she? Very sure/intermediate/unsure

Baby: hospital Number …….. Singleton Y/N Male/Female

End of Bed Assessment:

Birth weight……
MUAC……Head Circumference……..Abdominal
Circumference……
Foot length with **ruler** measurement I,……..Foot Length Measurement 2
……..
Foot length, **calipers:** foot measurement I……Foot measurement
2……
Anterior lens examination grade:
……..
Please perform Ballard Score on attached paper

Appendix 4

Consent Form

Summary

The Analysis of Gestational Estimates study is a study that is happening at Groote Schuur Hospital by University of Cape Town researchers.

Not all babies come out after nine months of pregnancy. There are various methods doctors use to try and work out whether the baby was delivered at the right time, or whether they came out too early. For babies who were delivered too early, doctors also try to estimate how early. Was it just a week or two, or was it a few months?

In this study we are trying to work out which are best ways of working out how premature (early) the baby's delivery was.

They will do this by reviewing the mother and baby's notes and doing a short examination (looking at and touching) of the baby, including measuring the length of the baby's foot.

After the study is finished, the researchers will put all the information together and write an article about how accurate the different methods of working out how early the baby came out are. This article will be published in a medical journal.

The Researchers

This study is being done by The University of Cape Town Department of Neonatology. The main researcher is Dr Alex Stevenson who is a paediatrician. He is being supervised by two senior neonatalogists, Dr Lloyd Tooke and Dr. Yaseen Joolay.

Why are we doing this study?

When a baby comes out too early, it is called a premature baby. The number of weeks the baby did spend in the mother's womb is called the gestational age. If we know the gestational age, it can tell us how premature a baby is.

This study will try to find out which are the best methods for working out gestational age and how reliable they are, especially in babies who didn't grow as well as they should have when in the mother's womb.

What will happen?

If you consent to being part of the study, the researcher will look at your notes and find out when you had your first ultrasound scan during this pregnancy. They will also write down the weight and length and head circumference of the baby at birth.

Then they will look at the baby and measure how long the baby's foot is.

They will also feel how well some of the baby's joints move, including the baby's wrist, knee, hip and shoulder.

All this information will be written on a piece of paper and stored in a secure place.

At the end of the study all the information will be put together in an anonymous, confidential way. For example it will say that 20 babies were premature, but it won't say who those babies were. All the information in the summary will be confidential so no-one reading the study will be able to work out who took part in the study.

The information collected may be shared with other researchers in other countries doing similar research. This may also be published in other scientific journals.

Risks

Because this study only involves the researcher looking at your notes and looking at the baby, along with briefly feeling the baby's joints and measuring the length of the foot, this is a low risk study.

There is a small risk that the baby may get cold during the brief examination. The researchers will keep the baby well covered where possible and will stop the examination if the baby gets cold.

Whenever we touch babies there is a risk of transmitting germs to the baby. To prevent this the researchers will wash their hands before touching the baby and also clean the hands with a proven antiseptic spray. The ruler for measuring the babies' feet will also be cleaned.

The researchers understand that it is important for mothers and babies to bond. They will not interrupt a mother who is spending precious time with their baby and will not interrupt breastfeeding.

Benefits

By participating in this study mothers can know that they are helping doctors to develop better ways to identify the gestational age of babies.

There are no other direct benefits to the mothers or babies.

The study will not pay any money to mothers or families participating in this study.

Participation is voluntary

Please note that it is entirely voluntary for parents to allow (or refuse) their baby's participation in this study. You may refuse to have your baby participate in the study. Even if you initially agree, but then later change your mind, you can withdraw your baby from the study at any time. Whatever choice you make will not affect the care your baby received- they will receive the best care available regardless of whether they participate or not

Questions/Complaints

If you have any questions about this research you can discuss with the researcher conducting the research or you can contact:

The Department of Neonatology

Old Main Building , Groote Schuur Hospital, Observatory 7925

Tel 02I 4046025 email Secretary Mrs. Abass Gabeba.abass@uct.ac.za

You can also contact the Human research Ethics Committee

Old Main Building of Groote Schuur Hospital, Floor E53, Room 46, Observatory, 7925 .
Tel +27 21 650 3002

Conclusion and Consent

By Signing below you are agreeing that you understand what he study is about, what is going to happen during the study, and that you are happy for your baby to be included in this study.

Name of Mother or legal guardian Signature of Mother or legal guardian

……. …….
Name of Researcher Signature of Researcher

……. ……..
Name of Witness Signature of Witness

…….. ……..

Date……..