Diagnostic Accuracy of Neonatal Assessment for Gestational Age Determination:
A systematic review

Anne CC Lee[a,b], MD, MPH, Pratik Panchal[c], MD, MPH, Lian Folger[a], BA, Hilary Whelan[d], MD, Rachel Whelan[e], MPH, BA, Bernard Rosner[f], PhD, Hannah Blencowe[g,h], MRCPCH, MBChB Msc and Joy Lawn[g,h,i], MBBS, PhD

Affiliations:
[a] Department of Pediatric Newborn Medicine, Brigham and Women's Hospital, Boston, MA
[b] Harvard Medical School, Boston, MA
[c] Wits School of Public Health, University of Witwaterstrand, Johannesburg, South Africa
[d] Department of Pediatrics, University of Rochester Medical Center, Rochester, NY
[e] Community Partners International, Yangon, Myanmar
[f] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
[g] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine
[h] The Centre for Maternal, Adolescent, Reproductive, and Child Health (MARCH), London School of Hygiene and Tropical Medicine, London, UK
[i] Research and Evidence Division, UK AID

Address correspondence to: Dr. Anne CC Lee**,** Brigham and Women's Hospital, BB502A, 75 Francis Street, Boston, MA 02115, alee6@bwh.harvard.edu, 617-732-8343

Short Title: (55 characters or fewer with spaces) Accuracy of Neonatal Gestational Age Assessment

Financial Disclosure: The authors have no financial relationships relevant to this article to be disclosed.

Funding Source: This work was supported by the Bill & Melinda Gates Foundation through grant OPP1130198.

Conflicts of Interest: The authors have no conflicts of interest relevant to this article to disclose.

Systematic Review Registration: This systematic review was registered with the International prospective register of systematic reviews. PROSPERO Registration number: CRD42015020499

Abbreviations:
AGA: appropriate-for-gestational age; AVCL: anterior vascular capsule of the lens; BOE: best obstetric estimate; GA: gestational age; HIC: high-income countries; IUGR: intra-uterine growth restriction; LBW: low birth weight; LMIC: low- and middle-income countries; LMP: last menstrual period; NICU: neonatal intensive care unit; SGA: small-for-gestational age; US: ultrasound; WHO: World Health Organization

Table of Contents Summary: This systematic review and meta-analysis summarizes evidence on the diagnostic accuracy of neonatal clinical assessment for gestational age determination, with focus on low-middle income countries.

Contributors' Statement

Dr. Lee conceptualized and designed the study, coordinated and supervised data collection, completed secondary data extraction, drafted, reviewed, revised, and finalized the manuscript, and approved of the manuscript as submitted.

Dr. Panchal designed the database searches, carried out initial screening and data extraction for postnatal clinical exams, and conducted meta-analyses. He reviewed and revised the manuscript, and approved the final manuscript as submitted.

Ms. Folger screened and extracted data for anterior vascularity of the lens, helped write sections of the manuscript, and formatted, reviewed, and revised the manuscript, and approved the final manuscript as submitted.

Dr. Whelan undertook initial screening and data extraction for postnatal clinical exams, reviewed the manuscript, and approved the final manuscript as submitted.

Ms. Whelan coordinated and supervised data collection and data extraction, reviewed the extracted data, reviewed and revised the manuscript, and approved the final manuscript as submitted.

Dr. Rosner advised the statistical analysis of the data extracted, provided feedback on analyses, reviewed and revised the manuscript, and approved the final manuscript as submitted.

Drs. Blencowe and Lawn helped synthesize the data and data analysis, critically reviewed and revised the manuscript, and approved the final manuscript as submitted.

ABSTRACT
Context: An estimated 15 million neonates are born preterm annually**.** However, in low-and-middle-income countries (LMIC), the dating of pregnancy is frequently unreliable or unknown.
Objective: To conduct a systematic literature review and meta-analysis to determine the diagnostic accuracy of neonatal assessments to estimate gestational age (GA).
Data Sources: PubMed, EMBASE, Cochrane, Web of Science, POPLINE, and WHO Global Health Library databases.
Study Selection: Studies of live-born infants comparing individual clinical signs or neonatal scores/assessments for GA estimation with a reference standard.
Data Extraction: Two independent reviewers extracted data on study population, design, bias, reference standard, test method, agreement, validity, correlation, and inter-rater reliability.
Results: 4,956 studies were screened; 78 were included. We identified 19 newborn assessments for GA estimation (ranging 4-23 signs). Compared to ultrasound, the Dubowitz score dated 95% of pregnancies within ±2.6 weeks (n=7 studies), while the Ballard score overestimated GA (0.4 weeks), and dated pregnancies within ±3.8 weeks (n=9). Compared to last menstrual period, imprecision was greater [Dubowitz ± 2.9 weeks (n=6), Ballard ±4.2 weeks (n=5)]. Assessments with fewer signs tended to be less accurate. A few studies showed a tendency to overestimate GA in preterm infants and underestimate GA in growth-restricted infants.
Limitations: Poor study quality and few studies with an early ultrasound based reference.
Conclusions: Efforts in LMIC should focus on improving dating in pregnancy through ultrasound and improving validity in growth-restricted populations. In settings where ultrasound is not possible, increased efforts are needed to develop simpler, yet specific, approaches for newborn assessment, through new combinations of existing parameters, new signs, or technology.
PROSPERO Registration Number: CRD42015020499

Key Words: Gestational Age, Maturity, Preterm Birth, Small for Gestational Age, IUGR, Neonatal Assessment, Dubowitz, Ballard, Diagnostic Accuracy

INTRODUCTION

Of the estimated 14.9 million annual preterm births, 13.6 million (91%) occur in low-middle income countries (LMIC), defined by the World Bank as GNI per capita less than $12,475.[1,2] Preterm birth is the leading cause of under-5 child mortality globally, accounting for 1 million neonatal deaths annually, almost all occurring in LMIC.[3] In these settings, early recognition of the preterm infant may help the timely delivery of potentially life-saving interventions for the newborn, such continuous positive airway pressure or kangaroo mother care.

Ultrasound dating in early pregnancy is the most accurate method currently available to assess gestational age (GA), and is standard of care in high-income countries. In LMIC, pregnancy dating is challenging, and the GA of the infant is frequently unknown or inaccurate, due to several factors. Maternal recall of last menstrual period (LMP) is often unavailable or unreliable, particularly in populations with high rates of maternal illiteracy.[4,5] The shortage of health care providers in LMIC— currently estimated at 7.9 million[6]— contributes to the low coverage of antenatal care in these regions, especially for women in rural areas and the lowest income groups. In 2015, in sub-Saharan Africa and Southeast Asia, fewer than one third of mothers in the poorest household quintile received at least one antenatal care visit.[7] Furthermore, the time to first presentation to antenatal care is late, occurring, on average, at 5 months gestation.[8-11] Access to ultrasonography is low in LMIC, with fewer than 7% of pregnant women having access to ultrasound in sub-Saharan Africa.[4] Traditionally, sonography in late pregnancy is inaccurate for determining gestational age (±3-4 weeks).[12,13]

Clinical assessment of newborn maturity after birth has long been used as a proxy to estimate gestational age in the newborn. In 1966, Farr et al. described and defined a classification system for the development of a range of external physical characteristics.[14] In

1968, Amiel Tison described the maturation of the neonatal neurologic assessment.[15] Lily and Victor Dubowitz developed a scoring system for gestational age in 1970 based on 10 neurologic and 11 physical signs. The Dubowitz exam dated pregnancies within 5 days of last menstrual period (LMP) in their original study.[16] Since then, numerous simplified clinical assessments have been described in the literature.[17-20] The Ballard exam is one of the most commonly used. It is a simplified scoring system comprised of 11 signs[21] that was revised to the New Ballard assessment in 1991 to improve accuracy for early preterm infants.[22]

Clinical newborn assessment for GA dating has become less relevant in high-income settings, where coverage of early pregnancy ultrasound is high and uncertainty of pregnancy dating is less common than in LMICs. However, in LMICs, GA is frequently unknown, and furthermore, challenging to estimate when fetal growth restriction is prevalent. Accurate GA is required to identify babies that are preterm and small for gestational age, and provide them with effective interventions. The Every Newborn Action plan, launched in 2014, seeks to end preventable neonatal deaths and stillbirths by 2030.[23] Its measurement improvement roadmap[24] has identified improved GA measurement as a high-priority area to improve the epidemiology of preterm birth and small-for-gestational age, and to improve comparability of neonatal mortality estimates through stratification of neonatal deaths by GA and birthweight. In settings without widespread access to early ultrasound scan dating and where the accuracy of recalled LMPs is highly variable, clinical assessment of the newborn remains the commonest available tool to assess GA.

The aim of this systematic review is to: 1) identify individual neonatal signs and combined clinical scores or assessments that have been used to ascertain GA of the newborn, and

2) to assess the diagnostic accuracy and reliability of these methods to estimate GA compared to standard dating using a reference standard (ultrasound or LMP).

METHODS

Search Strategy

We conducted a systematic review of the published and grey literature, which was initially done in March 2015 and updated in June 2016 (Figure 1). The review was registered with the International prospective register of systematic reviews (PROSPERO CRD42015020499). The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement is available in Web Appendix 1. Databases searched included PubMed, EMBASE, Cochrane, Web of Science, POPLINE, and the WHO Global Health Libraries/regional databases (LILACS, IMEMRO, AIM) (Review protocol, Web Appendix 2). The detailed search terms and strategy are in Web Appendix 3.

Inclusion Criteria

There were no language restrictions. Abstracts of non-English articles were translated via Google Translate to determine eligibility, and if eligible, the full text was translated to English by fluent speakers. Articles were considered for inclusion if the study: 1) included live-born neonates, 2) compared at least two methods of GA estimation, one of which was a neonatal clinical assessment/scoring method or individual clinical physical signs (i.e. anterior vascularity of the lens, inter-mammillary distance, skin impedance, palmar creases), and 3) reported at least one statistic assessing correlation, agreement or validity of GA estimation. Prenatal assessments (i.e. symphysis fundal height, ultrasound) and neonatal anthropometrics (i.e. foot length) were reviewed separately and will be reported elsewhere.

Exclusion Criteria

We excluded studies that did not provide data informing the correlation, agreement or validity of the neonatal clinical assessment compared to a reference method of pregnancy dating (i.e. ultrasound or LMP). We excluded studies from specialized sub-populations (e.g. infants of diabetic mothers), individual case reports and duplicate studies.

Data Extraction

All articles were reviewed independently by two researchers and extracted into a standard Excel file. Any differences were resolved by a third independent reviewer. Data were extracted on the following study characteristics: population characteristics, study setting, study design, patient recruitment, reference standard method, test method, GA distribution, and statistics regarding agreement, validity, correlation, or inter-rater reliability. A full list of variables that were extracted is in Web Appendix 4.

Study Quality Assessment

Two independent reviewers assessed and graded the methodological quality of the studies of diagnostic accuracy per the Cochrane Diagnostic Test Accuracy working group recommendations using the QUADAS-2 (Quality Assessment of Diagnostic-Accuracy Studies – 2)[25] tool, which was modified to fit the context of this review (Web Appendix 2, Section 5). Any differences were resolved by joint review of the studies. Individual studies were evaluated for limitations and biases in the following domains: patient selection, test method, reference standard, and patient flow and timing. Studies were graded of highest quality for those which had a

reference standard GA of ultrasonography or best obstetric estimate (BOE) (including ultrasound confirmation of dating). While LMP may be considered "gold standard" in high-resource settings, where rates of early booking and literacy are high; in LMIC, LMP recall is less reliable due to low rates of literacy and late presentation for antenatal care.[13,26] We also assessed the generalizability of study results to LMIC.

Statistical Analysis

Stata 13 (StataCorp, College Station Texas) and R (R Foundation for Statistical Computing, Vienna, Austria) were used for analyses. The definition of a preterm birth was a live birth at <37 weeks of gestation. Studies were grouped by method of newborn assessment and reference standard. Simple descriptive statistics were used to report ranges and medians. The mean individual level differences between two methods of GA assessment were pooled using the metan package in Stata 13, which provided the pooled mean-difference estimate and 95% confidence interval. The variance and standard deviation around the pooled estimate was calculated using the formula[27]:

$$Variance_{pooled} = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

Studies also often reported the percent of test measures within ± 1-2 weeks of the reference standard measures. To summarize these data, percentages were logit transformed, and their standard errors calculated. Meta-analysis was conducted using the STATA metan command with a random-effects model. The Higgins I^2 statistic was calculated to assess heterogeneity. Forest plots were generated in R to summarize test diagnostic accuracy. Pooling of sensitivity and specificity separately fails to account for the inter-relatedness of the measures. Hierarchical bivariate models are recommended for meta-analysis[28] of these measures and were analyzed using MetaDisc® 1.4 and RStudio using the Mada package, and hierarchal summary

receiver operating characteristic curves were generated. Sub-group analyses were conducted by clinical method of assessment, reference standard type (ultrasound vs. LMP), and country income level (high income (HIC) vs LMIC). Correlation coefficients were not pooled given that many studies did not indicate the type of coefficient (i.e. Spearman or Pearson), and furthermore, methods for pooling of Spearman correlation coefficients have not been well described.[27]

RESULTS

A) NEONATAL CLINICAL ASSESSMENTS

Our searches for neonatal clinical assessment of GA yielded a total of 3,862 titles after de-duplication (Figure 1). 22 articles were identified by snowball searching the bibliographies of identified papers. 270 full-text articles were reviewed, and 66 included. Of these, 25 papers reported on the Dubowitz scoring system, 31 reported on the Original or New Ballard scoring system, and 25 reported on other clinical scores of assessing GA.

Overall Study Characteristics

The basic characteristics of all studies included in the review are shown in Web Appendix 5. The 66 studies with clinical assessment data were published between 1968 and 2016. Fewer than half of these studies were conducted in LMIC. The vast majority (n=62) were conducted in health facilities, while 4 studies were community-based. Nineteen of the studies were performed within NICUs on preterm or LBW populations. For the reference standard comparison, 31 studies had an ultrasound/BOE, 42 had LMP, and 3 used another neonatal clinical assessment. The level of health worker performing the assessment was a physician or nurse in the majority (68%) of

studies, and a community health worker/non-medical personnel in 3 (4.6%) studies; in the remaining studies, the level of health worker was not reported.

Study Quality

The overall QUADAS-2 summary assessment is shown in Web Appendix 6. In general, the quality of the studies was relatively low. There was a high risk of bias in over half of the studies related to patient selection, test method, or reference standard. The individual study QUADAS-2 data is available on request.

Neonatal Clinical Assessments/Scoring Systems

We identified nineteen different neonatal assessments or scoring systems (combining >1 individual clinical signs) for GA determination that were described in the literature from 1966 to 2014. Twelve were originally developed in high-income settings and 7 in LMIC (4 Africa, 2 Asia, 1 other). The reference standard from which the scoring systems were derived was US/BOE in only 2 studies, both from high-income settings. The assessments are shown in Table 1 by level of complexity. The most complex system, Amiel Tison[15], had 23 criteria, including a large number of complicated neurological signs. The simplest system, the Parkin[18], included only 4 external physical criteria. One simplified score was developed in Nigeria (Eregie) and also included physical anthropometrics (head circumference and mid-arm circumference).[19] The complexity of the assessment and required training are important considerations for feasibility in LMIC.

Individual External Physical Criteria/Signs

Table 2 shows 12 studies that reported the correlation of individual external physical criteria with GA. The correlation coefficients were generally higher for comparisons with a reference standard of LMP and the maturity of the external physical signs correlated positively with LMP GA. The median correlation coefficients ranged from 0.60 to 0.75 for most individual signs. Three studies used US/BOE GA estimates as their reference standard, and lower correlations were reported in 2 of these 3 studies; although the GA range of included infants did not include early preterm infants in both of these studies.[15,29] The physical characteristics with the highest median correlation coefficients were breast size, plantar skin creases, ear firmness and skin texture.

Individual Neuromuscular Signs

Ten studies reported correlation of individual neuromuscular criteria with GA (Table 2). The median correlation coefficients generally ranged from 0.52 to 0.70 in the studies with LMP reference. Correlation coefficients were lower in the same two studies with ultrasound-based dating, however these studies did not include early preterm infants.[15,29] The signs with the highest median correlation coefficients were ventral suspension, square window, and posture.

Validity of Neonatal Clinical Scores of Gestational Age

Studies that reported on the validity or agreement of neonatal clinical exams with a reference standard are shown in Table 3 (Dubowitz), Table 4 (Ballard), and Web Appendix 9 (other assessments).

1) Dubowitz Score

We identified a total of 26 studies that validated the Dubowitz Score (11 with ultrasound/BOE; 19 with LMP reference standard). In most studies, the neonatal assessment was performed by physicians or nurses. Ten studies were from LMIC.

US/BOE Reference Standard: Two studies reported the correlation of GA dating by Dubowitz scoring and BOE (r= 0.73 and 0.90, respectively). Seven studies reported a mean difference and standard deviation in GA between Dubowitz and ultrasound-based dating, ranging from -2.2 weeks (underestimation) to +0.7 weeks (overestimation). The pooled mean difference was not statistically different from the null hypothesis (i.e. difference=0), indicating no evidence of systematic bias (Table 5, Web Appendix 7a). The precision of the estimate is reflected in the standard deviation of the mean difference, which, at the individual study level, ranged from 0.52 to 1.94 weeks. The pooled standard deviation across the studies was 1.3 weeks, indicating that 95% of the differences in GA (Dubowitz score-US dating) fell within ± 2.6 weeks (n=7 studies). In the studies which reported upon the % agreement within weeks, the Dubowitz GA fell within 1 week of US based dates in 48% of infants (pooled estimate, n=3, 95% CI: 23%- 74%), and within 2 weeks in 75% of newborns (pooled estimate, n=3, 95% CI: 40% - 93%). One study reported a sensitivity of 61% and specificity of 99% to identify preterm infants <37 weeks.[30] Among studies in LMIC, there was no significant bias compared to ultrasound dating and the precision of GA dating by the Dubowitz score was similar in LMIC and HIC (Web Appendix 8).

In four studies, there was evidence of greater bias of Dubowitz scoring among preterm infants (Web Appendix 9). Four studies reported that the Dubowitz systematically overestimated GA in preterm infants by up to 2.6 weeks,[30-32] and more so among early preterm infants.[30-33]

LMP Reference Standard: The correlation of GA determined by Dubowitz scoring and GA determined by LMP was reported in 14 studies and was generally high, ranging from 0.41 to 0.94 (median= 0.89). The pooled mean difference in dating was 0.65 weeks (n=6, 95% CI: 0.01 - 1.30), indicating a systematic overestimation compared to LMP based GA (Table 5, Web Appendix 7a). 95% of the differences fell within ± 2.9 weeks of the mean. The GA determined by Dubowitz assessment fell within 1 week of LMP dates in 57% of newborns (n=4, 95% CI: 34% - 77%), and within 2 weeks in 87% (n=6, 95% CI: 70% - 95%). One study reported on the diagnostic accuracy of the Dubowitz to identify preterm infants (<37 weeks) (sensitivity 81.5%, specificity 98.6%).[34] Among LMIC studies (n=2), there was a tendency of the Dubowitz score to overestimate GA (0.48 wks), although the precision of the GA estimates were similar between HIC and LMIC (Web Appendix 8).

Two studies showed evidence that Dubowitz tended to overestimate GA in early preterm infants (Web Appendix 9).[35,36]

2) Ballard/New Ballard Score

We identified a total of 30 studies that assessed the validity of the Original (n=25) and/or New Ballard Score (n=13) (Table 4) (17 with ultrasound/BOE, 20 with LMP reference). The Original Ballard Score (1979)[21] was refined in 1991 to improve dating of extremely premature neonates. The signs assessed are the same in both versions, however, the New Ballard Score (1991)[22] includes expanded scoring categories for early preterm infants. Given the similarity of the assessments, results from studies that used either the Original or New Ballard were combined for the purpose of this analysis. Additionally, in the summary statistics and analyses, we only

included studies using the full Ballard. Ballard assessments were performed by medically-trained health workers (physicians, nurses or research assistants) in the majority of studies, and by community health workers in 2 studies. Fourteen studies were from LMIC.

US/BOE reference standard: Of 17 studies, 12 used the Original Ballard and 6 used the New Ballard. The correlation coefficients of GA determined by Ballard scoring vs. US/BOE ranged from 0.12 to 0.97 (median=0.85, n=7 studies). The mean difference in GA ranged from -0.41 weeks (underestimation) to +1.4 weeks (overestimation) in 9 studies. The pooled mean difference was 0.40 weeks (95% CI: 0.00-0.81) (Table 5, Web Appendix 7b), and while including zero, indicates a trend toward overestimation of GA. The pooled standard deviation across the studies was 1.9 weeks, indicating that 95% of the differences in GA by Ballard assessment vs. ultrasound dates fell within ± 3.8 weeks (n=9 studies, Table 5) of the mean. For the studies that reported upon agreement in weeks, Ballard score dates fell within 1 week of US dates in 34% (95% CI: 22% - 47%, n=3) of infants and within 2 weeks in 72% (95% CI: 53% - 85%, n=5) of newborns. The Ballard assessment had a pooled sensitivity of 64% (95% CI: 61% - 67%) and specificity of 95% (95% CI: 95% - 96%) for identifying preterm <37 week newborns (n=4 studies). Among LMIC studies, the trend of GA overestimation was similar to HIC studies. However, the imprecision of GA estimation was greater in LMIC compared to HIC studies (pooled standard deviation of 2.12 vs. 1.49) (Web Appendix 8).

Several studies reported evidence of greater bias in Ballard scoring among smaller babies (Web Appendix 9). Three studies reported that the Original Ballard assessment systematically overestimated GA by up to 2-3 weeks, in particular, among preterm infants[33,37,38], and generally, the trend was towards increasing bias in lower GAs. However, a study by Karl et al[39] in Papua

New Guinea found the opposite trend. Wariyar et al[38] reported that the New Ballard overestimated GA to a lesser degree than the Original Ballard in infants <30wks (1.6 vs. 3.4wks, respectively). Among SGA infants, two studies showed that GA was underestimated by the original Ballard.[29,37]

LMP reference standard: Thirteen studies used the Original Ballard, and 7 used the New Ballard Score. The correlation coefficients of Ballard GA and LMP GA ranged from 0.66 to 0.96 (median=0.85; n=13). The mean difference and standard deviation in GA was reported in 6 studies, ranging from 0.34 to 2.6 weeks (overestimation). The pooled mean difference was 0.70 weeks (95% CI: 0.36-1.04), indicating systematic overestimation (Table 5, Web Appendix 7b). 95% of mean differences fell within ± 4.2 weeks (n=5 studies) of the mean. Ballard GA fell within 1 week of LMP GA in 43.9% (95% CI: 23.9% - 66.1%; n=3) of newborns and within 2 weeks of LMP in 75.4% (95% CI: 70.3% - 79.8%; n=9) of newborns. The Ballard had a sensitivity of 84.1% (95% CI: 81.6% - 86.3%) and specificity of 83.5% (95% CI: 79.5% - 87.0%) for identifying preterm newborns (n=2 studies). Forest plots of the pooled sensitivity and specificity are shown in Figure 2. There were an inadequate number of studies to stratify analysis by LMIC vs. HIC.

Two studies demonstrated overestimation of GA among preterm infants by the Original Ballard exam,[40,41] but one study used the External Ballard only (Web Appendix 9).[41] In addition, two studies found that the Original Ballard performed differently for SGA infants: Baumann et al. reported that the correlation of Ballard with GA was lower among SGA infants compared to those AGA.[42] Constantine et al. showed that for SGA babies, the bias for GA dating was 1-1.5 weeks lower than for non-SGA infants.[40]

3) Other Clinical Assessments

Eighteen studies were identified which reported on the validity of other clinical methods of GA assessment (i.e. Eregie[29,35,43], Capurro[17,29,44-47], Parkin[29,36,38,48,49], Bhagwat[29,50], Tuncer[51,52], Finnstrom[53], Narayanan[54], Robinson[38,55]) (Web Appendix 10). Many of the methods were simplified assessments with fewer characteristics than the Dubowitz or Ballard clinical assessments. Fourteen studies were performed in LMIC settings. The Eregie assessment was developed in Nigeria and found to have high correlation with LMP based GA[43], however the performance was only fair in a South Asian study using ultrasound as a reference standard.[29] The Capurro is a simplified 7 sign assessment developed in South America, and 5 studies in LMIC were identified comparing Capurro dating to ultrasound-based dates (Table 5; Web Appendix 8).[29,44-47] The pooled sensitivity for the Capurro to identify preterm births using an ultrasound reference standard was low at 42.7% (95% CI: 35.6% - 50.0%), and the pooled specificity was 96.7% (95% CI: 95.7% - 97.5%) (n=3 studies).

4) Inter-rater Agreement

Web Appendix 11 shows the studies that reported on inter-rater agreement. Ten studies reported upon the agreement of GA estimates when the newborn clinical assessments were performed by two different assessors, and all studies found high rates of inter-rater agreement. In three studies, the kappa for the classification of preterm births ranged from 0.73 to 0.93, in the good-excellent range.[22,56,57] The GA estimates determined by two different raters were also highly correlated, with correlation coefficients (R) of 0.71 and 0.95 in two studies.[22,58] Four studies showed that the mean difference in scores between raters was not significant.[31,59-61]

B) ANTERIOR VASCULARITY OF LENS

The literature searches for anterior vascular capsule of the lens (AVCL) assessments yielded a total of 344 unique manuscripts (Figure 3), of which 27 full text articles were reviewed and 10 papers met inclusion criteria.

Overall Study Characteristics

The individual study characteristics are shown in Table 6. The studies were generally of smaller sample size (N= 30-356), and the assessments were performed by physicians in tertiary health facilities. The latest study was published in 1993. Three studies were from LMIC.

Study Quality

The overall QUADAS-2 summary assessment is shown in Web Appendix 12. In general, the quality of the AVCL studies was poor. The majority of studies had high risk of bias related to patient selection and the reference standard. Individual study QUADAS-2 data is available upon request.

Correlation of Grading of Anterior Vascular Capsule of the Lens with Gestational Age

Ten studies reported upon the correlation of the disappearance of the AVCL with GA, in the GA range <35 weeks (Table 6). Hittner[62] first found that as the infant matures in gestation, the anterior vascular capsule disappears in stages. In Grade 4, the entire anterior surface of the lens is vascularized (27-28 weeks gestation), and the vascularity reduces as GA increases. Grade 1 indicates a small number of vessels contributing to the periphery (~33-34 weeks), and Grade 0

indicates no vasculature (≥34 weeks). Of note, the reference standard in this original study was the Dubowitz assessment. Nine of the 10 studies used the AVCL grading system described by Hittner et al (1977). Three studies were conducted in LMIC (2 South Asia, 1 Africa). The examination was performed by a physician in all studies, and pupil dilation was performed prior to the assessment in 3 studies. In almost all studies, the exam was performed within the first 72 hours of life. Most studies were performed in NICU settings and included only preterm and/or low birth weight (LBW) infants. Only two studies included infants of all gestational ages. An ultrasound/BOE-based date was available in only two studies.

Two studies presented data on the average GA determined by Hittner's AVCL grading system.[33,63] Nine of the 10 studies reported a correlation coefficient of AVCL grading with GA. The correlation coefficients (R) for preterm and/or LBW populations ranged from -0.84 to -0.96, with a median of -0.88 (n=7 values). For the two studies that analyzed all-GA populations, the degree of correlation was lower (R= -0.64 and -0.45, respectively).[53,54] Three studies analyzed results for SGA preterm newborns, and among these studies, the median correlation coefficient was -0.77 (range: -0.68 to -0.91).[42,62,64]

C) INTERMAMILLARY DISTANCE

Searches for inter-mammilary distance yielded 320 unique studies. From these, 2 studies were identified that reported on the correlation of inter-mamillary distance with GA. In one study from Switzerland, inter-mamillary distance was correlated with LMP-based GA (r=0.62)[65]; whereas a study in India reported low correlation with neonatal clinical assessment-based GA.[66]

D) OTHERS

Searches for skin impedance and palmar creases yielded 109 and 321 unique studies, respectively. However, no articles addressed the validity, correlation or agreement with a reference standard GA estimate.

DISCUSSION

Accurate GA determination is a key public health priority to help target and reduce preterm birth related morbidity and mortality in LMIC. The Every Newborn Action plan has prioritized improving GA measurement as a high-priority area to improve the epidemiology of preterm birth and small-for-gestational age.[23] In our systematic literature review, we identified 19 different newborn assessments which have been used for GA dating. The most commonly reported and validated in the literature were the Dubowitz and Ballard scores. The Dubowitz score dated 95% of GA estimates within ± 2.6 weeks of ultrasound dating and was unbiased. The Ballard score tended to overestimate GA by 0.4 weeks compared to ultrasound, and dated 95% of infants within ±3.8 weeks of this mean. Newborn clinical assessments tended to overestimate GA among preterm infants, and therefore may misclassify preterm infants as full-term. They also tended to underestimate GA in growth-restricted babies. Simplified assessments tended to be less accurate. While several studies showed promise of the anterior vascularity of lens to classify GA <34-35 weeks, there were few studies assessing AVCL compared to an ultrasound-based reference standard.

Study quality was a major limitation of the studies identified in the review. Approximately half of the studies included had a high risk of bias from patient selection, reference standard diagnosis, or test measurement. Many of the original validation studies were

from the 1970s when LMP was the "gold standard" for pregnancy dating, and ultrasound was not widely available. Most hospital-based studies were performed in NICUs or among low birth weight babies, and thus prone to selection as well as measurement biases (lack of blinding). Fewer than half of the studies were based in LMIC, and studies in HIC may not be generalizable to these settings, due to differences in the prevalence in SGA, preterm birth, and health worker availability and training.

The majority of individual physical and neurologic signs that have been used in different scoring systems had fair to moderate correlation with GA, with a median correlation coefficient of 0.6. Skin opacity was the most weakly correlated, and is perhaps the most affected by timing of the assessment after birth. While neurologic signs may be more affected by neonatal morbidity (birth asphyxia, neonatal infection, maternal medications, etc.), the correlation coefficients for most neurologic signs were in a similar range to the physical criteria. In two studies[15,29] that excluded early-moderate preterm infants, the correlation of clinical signs with GA was lower, suggesting that the criteria maybe more correlated with GA at lower GA, but less discriminating for late preterm and full term infants. In interpreting correlation, it should be emphasized that correlation is not equivalent to agreement or validity. A higher correlation coefficient simply indicates that the rank order of scores for a particular sign may be associated with relative increases in GA. Thus, this does not equate to agreement in GA dating or diagnostic accuracy in identifying preterm births.

A critical consideration in LMIC is the validity of the assessment in populations with high rates of fetal growth restriction, or SGA. Distinguishing whether a small baby is preterm, SGA, or both, is a challenge in these settings. Most neonatal assessments were designed to measure infant maturity, as opposed to gestational length. SGA infants may act less mature

during a neonatal clinical assessment. Three studies have shown that among growth restricted infants (SGA), neonatal clinical exams tend to systematically underestimate GA.[29,37,40] Thus, improving the validity of the neonatal assessment to estimate GA in growth restricted populations is a critical research need in LMIC.[54,62,67]

The disappearance of the anterior vascular capsule of the lens (AVCL), or pupillary membrane, was found to correlate with increasing GA in 10 studies. The overall quality of the studies was low, with relatively small sample size and lower quality reference standard GA dating (LMP and/or clinical assessment). AVCL may show promise in LMIC with high rates of fetal growth restriction, considering that in the original Hittner[62] study, the grading correlated relatively well with GA even among growth-restricted/SGA infants. However, one study by Baumann et al.[42] reported lower correlation of AVCL grading with GA among SGA infants. An important factor is that the AVCL completely disappears after ~34wks GA; thus, it may not distinguish GA above 34 weeks. Other considerations include that assessment of the AVCL requires specialized skills and instruments (ophthalmoscope), which may limit the feasibility and scalability of the AVCL assessment in LMIC.

Several factors should be considered in interpreting and generalizing the validity of clinical methods of GA determination in different settings. For example, comparing the Ballard Score to an ultrasound reference standard, the imprecision was greater in LMIC studies (n=5) than HIC (n=4) (HIC: ± 3.0 weeks; LMIC: ± 4.2 weeks). The Dubowitz score performed similarly in LMIC and HIC settings, though the number of studies was small for comparison. The validity of a clinical assessment may vary with the level of medical training of the assessor.[29,68] Most LMIC studies identified used physicians, nurses or midwives, and there were few studies with front line health workers. The validity of the newborn assessment has primarily

been studied in the facility/hospital-based setting, and the few studies in home-based settings had poorer performance.[29,68] Certain factors may improve the validity in the hospital setting, including the timing of assessment sooner after birth, more controlled environment, and lighting. The development of some characteristics may vary by ethnicity. For example, plantar creases have been reported to progress differently in African American populations.[69] Skin color also varies between ethnicities, and the interpretation or scoring of certain signs related to skin coloring may vary between populations. Gestational diabetes is more common in specific populations (Asian and African American)[70] and may affect the maturity assessment. Finally, the performance of an assessment may also be affected by the GA ranges in which it is tested. For example, many of the scoring systems were developed and validated in NICU populations with larger numbers of early preterm infants. The performance and validity of the assessments may be different in a general population where there is a larger representation of late preterm and near-term infants.

Feasibility and scalability are critical factors in considering the use of the newborn assessment for GA dating in LMIC. As shown in this review, there is a positive correlation between the number of parameters and accuracy of a GA assessment. Yet, there is likely to be a negative correlation between number of parameters (especially neurological) and feasibility of use. While the Dubowitz assessment had the best accuracy of the newborn clinical assessments, the assessment is complex (21 signs), may take 15-20 minutes to complete, and includes more difficult-to-train neurologic criteria. In South Asia and Sub-Saharan Africa, approximately half of births occur outside of hospital facilities, and community-based health workers or traditional birth attendants may be the first point-of-contact for newborns. These health workers may not have medical training, skills, or time required to adequately perform the assessment. The

duration of the assessment, as well as the feasibility of training, standardization, and quality control are critical considerations in evaluating a method of GA assessment that may be scaled up in LMIC.

Finally, when evaluating methods of GA assessment in LMIC, the clinical, research and programmatic objectives should be weighed. For the clinician, the primary objective is to identify preterm infants requiring special care, and individual level misclassification may result in missed opportunities for intervention. A measurement tool with high sensitivity is desired in order to identify all preterm infants, perhaps at the expense of specificity. A very simple tool based on a single parameter, such as foot size or another anthropometric parameter may be suitable to meet these needs. On the other hand, for research purposes, a more precise and continuous measurement of GA may be desirable. Given the inaccuracy of clinical GA scores, for clinical research requiring precise GA dating, early pregnancy ultrasound should be used. At the population level, inaccuracy and imprecision in GA dating may result in biased estimates of preterm birth rates and epidemiologic associations with preterm birth.[71] Determining the optimal precision (i.e. a 95% CI of +/-1, 2, vs. 3 weeks) and diagnostic accuracy is critical to choosing an appropriate method of GA measurement for LMIC. Research priorities for improving GA determination in LMIC are shown in Figure 4.

CONCLUSION

Improving GA dating is a key priority for improving the measurement of the global burden of disease of preterm birth and SGA, as well as the delivery of effective interventions to improve the survival and development of these high-risk populations. As part of the Metrics Group of the Every Newborn Action Plan, we have conducted the first systematic review and

meta-analysis assessing the diagnostic accuracy of published scoring systems for neonatal gestational age assessment. In general, neonatal assessments with more parameters tended to be more accurate. Notably the Dubowitz score, with 21 signs including neurological assessment, was found to be most accurate (±2.6 weeks). The Ballard exam, with 12 signs, over-estimated GA by 0.4 weeks, and had wider limits of agreement (±3.8 weeks). Both assessments tended to overestimate GA in preterm infants, and underestimate GA in growth-restricted babies. The assessment of the anterior vascular capsule of the lens (AVCL) correlated well with GA below 35 weeks; however, the assessment requires an ophthalmoscope. Feasibility is a critical consideration in LMIC, and the complexity of scoring, training required, time to conduct the assessment and specialized equipment are challenges to scale up. Additional high-quality studies are needed in LMIC to determine the accuracy of neonatal assessment compared to an early ultrasound reference, particularly in settings with SGA, as well as to explore the feasibility of implementation of complex GA assessments. This work also underlines the importance of increasing the focus on improving the coverage of accurate GA assessment through early pregnancy ultrasound scans and innovations to improve GA assessment in late pregnancy, such as novel ultrasound approaches. In settings where early ultrasound is not possible, increased efforts and innovation are urgently needed to develop simpler, yet specific, approaches for clinical GA assessment of the newborn, either through new combinations of existing parameters, new signs, or technology.

Acknowledgements

We would like to acknowledge the team of students who were also part of the gestational age working group in the BWH global newborn health lab (Chelsea Clark, Keiko Chen). We would also like to thank the BWH Department of Newborn Medicine and Dr. Terrie Inder for

their support of this work. Finally, we would like to thank the following individuals for their assistance in translating foreign articles: Madeline Gilbert, Alison Leschen, Felix Bergmann, and Lina Driouk.

References

1. Blencowe H, Cousens S, Oestergaard MZ, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet* 2012; **379**(9832): 2162-72.
2. World Bank. World Bank Country and Lending Groups. *The World Bank: Data: Country Classification*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups. 2017. Accessed May 21, 2017.
3. Liu L, Johnson HL, Cousens S, et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* 2012.
4. Aliyu AD, Kurjak A, Wataganara T, et al. Ultrasound in Africa: what can really be done? *J Perinat Med* 2015; **44**(22).
5. Savitz DA, Terry JW, Jr., Dole N, Thorp JM, Jr., Siega-Riz AM, Herring AH. Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination. *Am J Obstet Gynecol* 2002; **187**(6): 1660-6.
6. World Health Organization. Global health workforce shortage to reach 12.9 million in coming decades. http://www.who.int/mediacentre/news/releases/2013/health-workforce-shortage/en/. November 11, 2013. Accessed May 28 2017.
7. UNICEF. Antenatal Care: Current Status and Progress, 2017. https://data.unicef.org/topic/maternal-health/antenatal-care/. Updated April 2017. Accessed April 17 2017.
8. Abou-Zahr CL, Wardlaw TM, World Health Organization Dept. of Reproductive Health and Research. Antenatal care in developing countries: promises, achievements and missed opportunities: an analysis of trends, levels and differentials, 1990-2001. Geneva: World Health Organization and UNICEF, 2003.
9. Central Statistical Agency (Ethiopia) and ICF International. Ethiopia Demographic and Health Survey 2011. Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ICF International; 2012. Available at: https://dhsprogram.com/pubs/pdf/FR255/FR255.pdf
10. Bucher S, Marete I, Tenge C, et al. A prospective observational description of frequency and timing of antenatal care attendance and coverage of selected interventions from sites in Argentina, Guatemala, India, Kenya, Pakistan and Zambia. *Reprod Health* 2015; **12 Suppl 2**: S12.
11. Matthews Z, Mahendra S, Kilaru A, Ganapathy S. Antenatal Care, Care-seeking and Morbidity in Rural Karnataka, India: Results of a Prospective Study. *Asia-Pacific Population Journal* 2001; **16**(2): 11-28.
12. Committee opinion no 611: method for estimating due date. *Obstet Gynecol* 2014; **124**(4): 863-6.
13. Blencowe H, Cousens S, Chou D, et al. Born too soon: the global epidemiology of 15 million preterm births. *Reprod Health* 2013; **10 Suppl 1**: S2.
14. Farr V, Mitchell RG, Neligan GA, Parkin JM. The definition of some external characteristics used in the assessment of gestational age in the newborn infant. *Dev Med Child Neurol* 1966; **8**(5): 507-11.
15. Amiel-Tison C. Neurological evaluation of the maturity of newborn infants. *Arch Dis Child* 1968; **43**(227): 89-93.

16. Dubowitz LM, Dubowitz V, Goldberg C. Clinical assessment of gestational age in the newborn infant. *J Pediatr* 1970; **77**(1): 1-10.
17. Capurro H, Konichezky S, Fonseca D, Caldeyro-Barcia R. A simplified method for diagnosis of gestational age in the newborn infant. *J Pediatr* 1978; **93**(1): 120-2.
18. Parkin JM, Hey EN, Clowes JS. Rapid assessment of gestational age at birth. *Arch Dis Child* 1976; **51**(4): 259-63.
19. Eregie CO. Assessment of gestational age: modification of a simplified method. *Dev Med Child Neurol* 1991; **33**(7): 596-600.
20. Bhagwat VA, Dahat HB, Bapat NG. Determination of gestational age of newborns--a comparative study. *Indian Pediatr* 1990; **27**(3): 272-5.
21. Ballard JL, Novak KK, Driver M. A simplified score for assessment of fetal maturation of newly born infants. *J Pediatr* 1979; **95**(5 Pt 1): 769-74.
22. 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; **119**(3): 417-23.
23. World Health Organization (WHO), United Nations International Children's Emergency Fund (UNICEF). Every Newborn: An Action Plan to End Preventable Deaths (ENAP). World Health Organization. Geneva; 2014.
24. World Health Organization (WHO). Every Newborn Action Plan Metrics. WHO technical consultation on newborn health indicators, Dec 3-5, 2014; Ferney Voltaire, France; 2014.
25. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; **155**(8): 529-36.
26. Kramer MS, McLean FH, Boyd ME, Usher RH. The validity of gestational age estimation by menstrual dating in term, preterm, and postterm gestations. *JAMA* 1988; **260**(22): 3306-8.
27. Rosner B. Fundamentals of Biostatistics. 8 ed: Cengage Learning; 2016.
28. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 10: The Cochrane Collaboration; 2010.
29. Lee AC, Mullany LC, Ladhani K, et al. Validity of Newborn Clinical Assessment to Determine Gestational Age in Bangladesh. *Pediatrics* 2016; **138**(1).
30. Moore KA, Simpson JA, Thomas KH, et al. Estimating Gestational Age in Late Presenters to Antenatal Care in a Resource-Limited Setting on the Thai-Myanmar Border. *PLoS One* 2015; **10**(6): e0131025.
31. Shukla H, Atakent YS, Ferrara A, Topsis J, Antoine C. Postnatal overestimation of gestational age in preterm infants. *Am J Dis Child* 1987; **141**(10): 1106-7.
32. Robillard PY, De Caunes F, Alexander GR, Sergent MP. Validity of postnatal assessments of gestational age in low birthweight infants from a Caribbean community. *J Perinatol* 1992; **12**(2): 115-9.
33. Sanders M, Allen M, Alexander GR, et al. Gestational age assessment in preterm neonates weighing less than 1500 grams. *Pediatrics* 1991; **88**(3): 542-6.
34. Raghu MB, Patel YS, Gupta K. Estimation of gestational age in Zambian newborn infants. *Ann Trop Paediatr* 1981; **1**(4): 245-7.
35. Sunjoh F, Njamnshi AK, Tietche F, Kago I. Assessment of gestational age in the Cameroonian newborn infant: a comparison of four scoring methods. *J Trop Pediatr* 2004; **50**(5): 285-91.

36. Vogt H, Haneberg B, Finne PH, Stensberg A. Clinical assessment of gestational age in the newborn infant. An evaluation of two methods. *Acta Paediatr Scand* 1981; **70**(5): 669-72.
37. Alexander GR, de Caunes F, Hulsey TC, Tompkins ME, Allen M. Validity of postnatal assessments of gestational age: a comparison of the method of Ballard et al. and early ultrasonography. *Am J Obstet Gynecol* 1992; **166**(3): 891-5.
38. Wariyar U, Tin W, Hey E. Gestational assessment assessed. *Arch Dis Child Fetal Neonatal Ed* 1997; **77**(3): F216-20.
39. Karl S, Li Wai Suen CS, Unger HW, et al. Preterm or not--an evaluation of estimates of gestational age in a cohort of women from Rural Papua New Guinea. *PLoS One* 2015; **10**(5): e0124286.
40. Constantine NA, Kraemer HC, Kendall-Tackett KA, Bennett FC, Tyson JE, Gross RT. Use of physical and neurologic observations in assessment of gestational age in low birth weight infants. *J Pediatr* 1987; **110**(6): 921-8.
41. Verhoeff FH, Milligan P, Brabin BJ, Mlanga S, Nakoma V. Gestational age assessment by nurses in a developing country using the Ballard method, external criteria only. *Ann Trop Paediatr* 1997; **17**(4): 333-42.
42. Baumann C, Huppi P, Amato M. [Prenatal and postnatal determination of gestational age of small newborn infants]. *Z Geburtshilfe Perinatol* 1993; **197**(3): 135-40.
43. Eregie CO. A new method for maturity determination in newborn infants. *J Trop Pediatr* 2000; **46**(3): 140-4.
44. Oliveira S, Kimura A, Riesco M. Evaluation of the gestational age through prenatal and postnatal data. 4th World Congress of Perinatal Medicine. Buenos Aires, Argentina; 1999. p. 1091-4.
45. Pereira AP, Dias MA, Bastos MH, da Gama SG, Leal Mdo C. Determining gestational age for public health care users in Brazil: comparison of methods and algorithm creation. *BMC Res Notes* 2013; **6**: 60.
46. Laveriano WRV. [Reliability of the post natal gestational assessment: Capurro test compared with ultrasound at 10+0 to 14+2 weeks of gestation]. *Revista Peruano Ginecologia y Obstetrica* 2015: 115-8.
47. Neufeld LM, Haas JD, Grajeda R, Martorell R. Last menstrual period provides the best estimate of gestation length for women in rural Guatemala. *Paediatr Perinat Epidemiol* 2006; **20**(4): 290-8.
48. Karunasekera KA, Sirisena J, Jayasinghe JA, Perera GU. How accurate is the postnatal estimation of gestational age? *J Trop Pediatr* 2002; **48**(5): 270-2.
49. Sreekumar K, d'Lima A, Nesargi S, Rao S, Bhat S. Comparison of New Ballards score and Parkins score for gestational age estimation. *Indian Pediatr* 2013; **50**(8): 771-3.
50. Bindusha S, Rasalam CS, Sreedevi N. Gestational age assessment of newborn- clinical trial of a simplified method. *Transworld Medical Journal* 2014; **2**(1): 24-8.
51. Tuncer M, Yilgor E, Erdem G. A new, simple three-step method for determining gestational age. *The Turkish Journal of Pediatrics* 1981; **23**(2): 85-97.
52. Cevit O, Bayram B, Toksoy HB, Gultekin A, Gokalp A. Gestational age assessment in preterm neonates weighing less than 2500 grams. *J Trop Pediatr* 1998; **44**(1): 57-8.
53. Finnstrom O. Studies on maturity in newborn infants. II. External characteristics. *Acta Paediatr Scand* 1972; **61**(1): 24-32.
54. Narayanan I, Dua K, Gujral VV, Mehta DK, Mathew M, Prabhakar AK. A simple method of assessment of gestational age in newborn infants. *Pediatrics* 1982; **69**(1): 27-32.

55. Serfontein GL, Jaroszewicz AM. Estimation of gestational age at birth. Comparison of two methods. *Arch Dis Child* 1978; **53**(6): 509-11.
56. Moraes CL, Reichenheim ME. [Validity of neonatal clinical assessment for estimation of gestational age: comparison of new Ballard score with date of last menstrual period and ultrasonography]. *Cad Saude Publica* 2000; **16**(1): 83-94.
57. Lee ACC, Uddin J, Shah R, et al. Validation of Community Health Worker Clinical Assessment of Gestational Age in Rural Bangladesh. Pediatric Academic Societies, Washington, DC; 2013.
58. Aslan Y, Yildirian A, Sen Y, Erduran E, Kasim S, Gedik Y. Assessment of Gestational Age in Healthy Neonates by Auxiliary Health Personnel Using a Simple Scoring System. *T Klin J Med Res* 2000; **18**: 121-4.
59. Sasidharan K, Dutta S, Narang A. Validity of New Ballard Score until 7th day of postnatal life in moderately preterm neonates. *Arch Dis Child Fetal Neonatal Ed* 2009; **94**(1): F39-44.
60. Gagliardi L, Scimone F, DelPrete A, et al. Precision of gestational age assessment in the neonate. *Acta Paediatr* 1992; **81**(2): 95-9.
61. Smith LN, Dayal VH, Monga M. Prior knowledge of obstetric gestational age and possible bias of Ballard score. *Obstet Gynecol* 1999; **93**(5 Pt 1): 712-4.
62. Hittner HM, Hirsch NJ, Rudolph AJ. Assessment of gestational age by examination of anterior vascular capsule of the lens. *J Pediatr* 1977; **91**(3): 455-8.
63. Sasivimolkul W, Siripoonya P, Tejavej A. Gestational age assessment by the examination of the anterior vascular capsule of the lens. *J Med Assoc Thai* 1986; **69 Suppl 2**: 38-45.
64. Hittner HM, Gorman WA, Rudolph AJ. Examination of the anterior vascular capsule of the lens: II. Assessment of gestational age in infants small for gestational age. *J Pediatr Ophthalmol Strabismus* 1981; **18**(2): 52-4.
65. Amato M, Huppi P, Claus R. Rapid biometric assessment of gestational age in very low birth weight infants. *J Perinat Med* 1991; **19**(5): 367-71.
66. Thawani R, Dewan P, Faridi MM, Arora SK, Kumar R. Estimation of gestational age, using neonatal anthropometry: a cross-sectional study in India. *J Health Popul Nutr* 2013; **31**(4): 523-30.
67. Skapinker R, Rothberg AD. Postnatal regression of the tunica vasculosa lentis. *J Perinatol* 1987; **7**(4): 279-81.
68. Taylor RA, Denison FC, Beyai S, Owens S. The external Ballard examination does not accurately assess the gestational age of infants born at home in a rural community of The Gambia. *Ann Trop Paediatr* 2010; **30**: 197-204.
69. Damoulaki-Sfakianski E, Robertson A, Gordero L. Skin creases on the sole of the foot as a physical index of maturity: comparison between Caucasian and Negro infants. *Pediatrics* 1972; **50**(3): 483-5.
70. Fujimoto W, Samoa R, Wotring A. Gestational Diabetes in High-Risk Populations. *Clinical Diabetes* 2013; **31**(2): 90-4.
71. Martin JA, Hamilton BE, Osterman MJK, et al. Births: Final data for 2013. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention (CDC), 2015.
72. Feresu SA, Gillespie BW, Sowers MF, Johnson TR, Welch K, Harlow SD. Improving the assessment of gestational age in a Zimbabwean population. *Int J Gynaecol Obstet* 2002; **78**(1): 7-18.

73. Nicolopoulos D, Perakis A, Papadakis M, Alexiou D, Aravantinos D. Estimation of gestational age in the neonate: a comparison of clinical methods. *Am J Dis Child* 1976; **130**(5): 477-80.
74. Farr V. Estimation of gestational age by neurological assessment in first week of life. *Arch Dis Child* 1968; **43**(229): 353-7.
75. Kollee LA, Leusink J, Peer PG. Assessment of gestational age: a simplified scoring system. *J Perinat Med* 1985; **13**(3): 135-8.
76. Klimek R, Klimek M, Rzepecka-Weglarz B. A new score for postnatal clinical assessment of fetal maturity in newborn infants. *Int J Gynaecol Obstet* 2000; **71**(2): 101-5.
77. Allan RC, Sayers S, Powers J, Singh G. The development and evaluation of a simple method of gestational age estimation. *J Paediatr Child Health* 2009; **45**(1-2): 15-9.
78. Robinson RJ. Assessment of gestational age by neurological examination. *Arch Dis Child* 1966; **41**(218): 437-47.
79. Roberts CJ, Hibbard BM, Evans DR, et al. Precision in estimating gestational age and its influence on sensitivity of alphafetoprotein screening. *Br Med J* 1979; **1**(6169): 981-3.
80. Vik T, Vatten L, Markestad T, Jacobsen G, Bakketeig LS. Dubowitz assessment of gestational age and agreement with prenatal methods. *Am J Perinatol* 1997; **14**(6): 369-73.
81. Awoust J, Keuwez J, Levi S. Comparision between three methods for assessment of fetal age. *Journal of Foetal Medicine* 1982; **2**(1): 11-5.
82. Rosenberg RE, Ahmed AS, Ahmed S, et al. Determining gestational age in a low-resource setting: validity of last menstrual period. *J Health Popul Nutr* 2009; **27**(3): 332-8.
83. Mitchell D. Accuracy of pre- and postnatal assessment of gestational age. *Arch Dis Child* 1979; **54**(11): 896-7.
84. Latis GO, Simionato L, Ferraris G. Clinical assessment of gestational age in the newborn infant. Comparison of two methods. *Early Hum Dev* 1981; **5**(1): 29-37.
85. Hertz RH, Sokol RJ, Knoke JD, Rosen MG, Chik L, Hirsch VJ. Clinical estimation of gestational age: rules for avoiding preterm delivery. *Am J Obstet Gynecol* 1978; **131**(4): 395-402.
86. Jaroszewicz AM, Boyd IH. Clinical assessment of gestational age in the newborn. *S Afr Med J* 1973; **47**(44): 2123-4.
87. Dawodu A, Qureshi MM, Moustafa IA, Bayoumi RA. Epidemiology of clinical hyperbilirubinaemia in Al Ain, United Arab Emirates. *Ann Trop Paediatr* 1998; **18**(2): 93-9.
88. Scher MS, Barmada MA. Estimation of gestational age by electrographic, clinical, and anatomic criteria. *Pediatr Neurol* 1987; **3**(5): 256-62.
89. Dombrowski MP, Wolfe HM, Brans YW, Saleh AA, Sokol RJ. Neonatal morphometry. Relation to obstetric, pediatric, and menstrual estimates of gestational age. *Am J Dis Child* 1992; **146**(7): 852-6.
90. Wylie BJ, Kalilani-Phiri L, Madanitsa M, et al. Gestational age assessment in malaria pregnancy cohorts: a prospective ultrasound demonstration project in Malawi. *Malar J* 2013; **12**: 183.
91. Thi HN, Khanh DK, Thu Hle T, Thomas EG, Lee KJ, Russell FM. Foot Length, Chest Circumference, and Mid Upper Arm Circumference Are Good Predictors of Low Birth Weight and Prematurity in Ethnic Minority Newborns in Vietnam: A Hospital-Based Observational Study. *PLoS One* 2015; **10**(11): e0142420.
92. Mackanjee HR, Iliescu BM, Dawson WB. Assessment of postnatal gestational age using sonographic measurements of femur length. *J Ultrasound Med* 1996; **15**(2): 115-20.

93. Alexander GR, Hulsey TC, Smeriglio VL, Comfort M, Levkoff A. Factors influencing the relationship between a newborn assessment of gestational maturity and the gestational age interval. *Paediatr Perinat Epidemiol* 1990; **4**(2): 133-46.
94. Alexander GR, de Caunes F, Hulsey TC, Tompkins ME, Allen M. Ethnic variation in postnatal assessments of gestational age: a reappraisal. *Paediatr Perinat Epidemiol* 1992; **6**(4): 423-33.
95. Ahn Y. Assessment of gestational age using an extended New Ballard Examination in Korean newborns. *J Trop Pediatr* 2008; **54**(4): 278-81.
96. Guillory C, Carsia-Prats JA, Hittner HM, Rudolph J. Effect of prenatal steroid administration on the anterior vascular capsule of the lens (AVCL) in preterm infants. *Pediatric Research* 1980; **14**(4).
97. Krishnamohan VK, Wheeler MB, Testa MA, Philipps AF. Correlation of postnatal regression of the anterior vascular capsule of the lens to gestational age. *J Pediatr Ophthalmol Strabismus* 1982; **19**(1): 28-32.

Figure Legends

Figure 1. Neonatal Clinical Assessment: Flow Diagram

Diagram of the screening process to identify studies for inclusion in neonatal assessment review; adapted from the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analysis; Moher et al., 2009).

Figure 2. Forrest Plots of the Ballard Exam Sensitivity/Specificity for Identifying Preterm Births Compared to Ultrasound (A,B) and Last Menstrual Period (C,D)

LMP= last menstrual period

Figure 3. Anterior Vascular Capsule of the Lens: Flow Diagram

Diagram of the screening process to identify studies for inclusion in AVCL review; adapted from the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analysis; Moher et al., 2009). AVCL= anterior vascular capsule of the lens

Figure 4. Research Priorities to Improve Gestational Age Dating in LMIC

LMIC= low-and-middle-income countries, SGA= small-for-gestational-age, GA= gestational age

Table 1. Scoring Systems by Level of Complexity

Clinical Scoring System/Name	# of criteria	Physical Criteria	Neuromuscular Criteria	Other Criteria	Reference Standard	Original Manuscript Accuracy	Study Setting	Sample Size	Year
Amiel Tison[15]	23	Skin color, skin opacity, skin texture, oedema, lanugo, skull hardness, ear form, ear firmness, genitals, breast size, nipple formation, plantar creases	Return to flexion of forearms, scarf sign, popliteal angle, foot dorsiflexion, righting reaction, raise-to-sit, back-to-lying, finger grasp and response to traction, non-nutritive sucking, crossed extension, vision fix and track		LMP	Individual correlation coefficients available per criteria & a regression equation.	France/Paris, Port-Royal-Baudelocque Hospital	397	1999
Feresu[72]	22	Edema, skin texture, skin color, skin opacity, lanugo, plantar creases, nipple formation, breast size, ear form, ear firmness, genitals	posture, square window, dorsiflexion of foot, arm recoil, leg recoil, popliteal angle, heel-t0-ear, scarf sign, head lag, ventral suspension	Birth weight	LMP	Dubowitz Score - r= 0.81; Revised Ballard Score - r=0.8	Maternity Unit, Harare Central Hospital; Harare, Zimbabwe	364	2002
Dubowitz[16]	21	Edema, skin texture, skin color, skin opacity, lanugo, plantar creases, nipple formation, breast size, ear form, ear firmness, genitals	Posture, square window, ankle dorsiflexion, arm recoil, leg recoil, popliteal angle, heel-to-ear, scarf sign, head lag, ventral suspension		LMP	95 CI: 2.0 weeks	NICU, Jessop Hospital for Women, Sheffield, England	167	1970
Sunjoh[35]	21	Edema, skin texture, skin color, skin opacity, lanugo, plantar creases, nipple formation, breast size, ear form, ear firmness, genitals	posture, square window, dorsiflexion of foot, arm recoil, leg recoil, popliteal angle, heel-t0-ear, scarf sign, head lag, ventral suspension		LMP	Combined Dubowitz & Farr; mean difference - 0.5 (+/- 1.31) weeks, r=0.94	Mother & Child Center, National Social Insurance & Central Hospitals; Yaounde, Cameroon	358	2004
Dubowitz & Farr (from Nicolopoulos[73])	17	skin texture, skin color, skin opacity, lanugo, plantar creases, nipple formation, breast size, ear form, ear firmness	posture, square window, dorsiflexion of foot, popliteal angle, heel-t0-ear, scarf sign, head lag, ventral suspension		LMP	r = 0.878	Athens/Greece, Alexandra Maternity Hospital	710	1976
Finnstrom[53]	12	Breast size, nipple formation, skin opacity, scalp hair, hair-forehead border, eyebrows, ear cartilage, fingernails, xiphoid process, external genitalia, plantar skin creases, pupillary membrane			LMP	r = 0.84 for 5 external characteristics (nipple formation, plantar skin creases, breast size, scalp hair, ear cartilage)	Sweden/Umea, tertiary care hospital	174	1972
Ballard[21]	12	Skin color, Lanugo, Plantar creases, Breast size, Ear firmness, Genitals	Posture, square window (wrist), arm recoil, popliteal angle, scarf sign, heel-to-ear		LMP & Clinical Data	r = 0.852 (based on 224 infants)	NICU, Cincinnati General Hospital, Cincinnati, USA	252	1979
Ballard (New Ballard Score, NBS)[22]	12	Skin, lanugo, plantar crease, breast maturity, Eye/ear, genitals	Posture, square window (wrist), arm recoil, popliteal angle, scarf sign, heel to ear		BOE	r = 0.97	NICUs and nurseries, 4 major medical centers, Cincinnati, Ohio, USA	530	1991

Clinical Scoring System/Name	# of criteria	Physical Criteria	Neuromuscular Criteria	Other Criteria	Reference Standard	Original Manuscript Accuracy	Study Setting	Sample Size	Year
Farr[74]	10		Spontaneous motor activity, reaction of pupils to light, rate of sucking, closure of mouth when sucking, stripping action of the tongue, resistance against passive movement, recoil of forearms, plantar grasp, pitch of cry, intensity of cry		LMP	Accurate +/- 1 wk: 61%	Aberdeen, Scotland	82	1968
Tuncer[51]	8	Skin texture, ear form, firmness, breast size & nipple formation, plantar creases, facial appearance	Posture, arm recoil, scarf sign		LMP	r = 0.945 (assessed by neonatologists)	Hacettepe University, NICU, Ankara, Turkey	100	1981
Eregie[19]	8	Skin texture, ear form, breast size, genitalia	Posture, scarf sign	Head circumference, mid-arm circumference	Dubowitz	Accurate within +/- 2 weeks: 92%	University teaching Hospitals, Benin, Nigeria	262	1991
Capurro[17]	7	Skin texture, nipple formation, ear form, breast size, plantar creases	Scarf sign, head lag		LMP	r = 0.9 (std. error of estimation= 8.4 days)	Montevideo, Uruguay	115	1978
Kollee[75]	7	Skin color, skin texture, plantar creases, breast size, ear firmness, nail length		AVCL	NS	± 19.9 days (95% CI)	Catholic University, Nijmegen, The Netherlands.	229	1985
Klimek[76]	6	Lanugo, plantar creases, breast	Posture, angle forearm to arm, pulling an elbow to the body		LMP	r = 0.72 (comparison b/w Klimek & Ballard)	Tertiary care hospital, Krakow, Poland	800	2000
Simplified Dubowitz (from Allan 2009[77])	6	Breast size, skin texture, ear bending (substituted from ear firmness because some Aboriginal babies have less ear cartilage)	Square window, popliteal angle, scarf sign		US (27 were high quality 1st trimester)	Mean difference: 0.4 wks (95% LOA: -2.8-1.9)	Royal Darwin & Darwin Private Hospitals, Northern Territory, Australia	98	2009
Narayanan[54]	5	Skin color, ear form, plantar skin crease, breast formation, skin texture		AVCL	LMP	95% CI of GA estimation - 11 days	Kalawati Saran Children's Hospital, New Dehli, India	356	1982
Robinson 1966[78] (from Serfontein 1978[55])	5		Pupil reaction, traction, glabellar tap, neck-righting, head-turning		Dubowitz	95 CI: +/- 1 wk; r = 0.85	South Africa, "cape colored babies"	73	1966
Parkin[18]	4	Skin texture, breast size, edema, plantar skin creases, nail length, nail texture, ear firmness, skull hardness, lanugo hair, genitalia			LMP	95 CI: 18.1 days	University hospital, Newcastle, England	392	1976
Bhagwat et al[20] (from Bindusha 2014[50])	4	Skin texture, breast size, ear firmness, genitalia)			LMP	Mean difference: -0.58 weeks; r = 0.91	Government Medical College, Thiruvananthapuram, Kerala, India	1000; GA 28-37 wks, preterm with Apgar scores >6	2014

Abbreviations: NS=not stated, BOE= best obstetric estimate, US= ultrasound, LMP= last menstrual period, AVCL= anterior vascular capsule of the lens

Table 2. Correlation of Individual Physical or Neuromuscular Criteria with Gestational Age

	Amiel-Tison (1999)[15]	Lee (2016)[29]	Ballard (New Ballard) (1991)[22]	Parkin (1976)[18]	Dubowitz and Farr (Nicolopoulos 1976)[73]	Raghu (1981)[34]	Feresu (2002)[72]	Dubowitz and Farr (Sunjoh 2004)[35]	Finnstrom (1972)[53]	Ballard (1979)[21]	Tuncer (1981)[51]	Narayanan (1982)[54]	Summary Across all Studies Median (Min, Max)
N (sample size)	397	710	530	392	710	160	364	358	174	252	220	356	
Population Description	Port-Royal-Baudelocque Hospital, Paris, France	Community setting, Sylhet district, Bangladesh	NICUs & nurseries; 4 major medical centers, Cincinnati, USA	Hospital, University of Newcastle upon Tyne, England	Alexandra Maternity Hospital and private maternity clinics; Athens, Greece	Premature Unit, University Teaching Hospital, Lusaka, Zambia	Maternity unit, Harare Central Hospital; Harare, Zimbabwe	Mother and Child Centre, National Social Insurance Hospital & Central Hospital; Yaounde, Cameroon	University Hospital, Umea, Sweden	NICU/ nursery, Cincinnati General Hospital, Cincinnati, USA	NICU, Hacettepe University, Ankara, Turkey	Kalawati Saran Children's Hospital, New Dehli, India	
Gestational Age range included	37-41 weeks	34-42 weeks	20-44 weeks	25.2-45.2 weeks	28-44 weeks	NS	24-45 weeks	25-44 weeks	32.1-34 weeks	26-44 wks, 760-5460g	27-41 weeks	26-44.4 weeks	
Reference Standard	**BOE**	**US**	**BOE**	LMP	LMP	LMP	LMP	LMP	LMP	LMP	LMP	LMP	
Physical Criteria													
Skin colour	0.19	0.05		0.78	0.76	0.52	0.45	0.8	0.48	0.84		0.74	0.63 (0.05, 0.84)
Ear form	0.11	0.02	0.73		0.76	0.64	0.57	0.72	0.41	0.84	0.62		0.63 (0.02, 0.84)
Ear firmness	0.18	0.03		0.78	0.76	0.65	0.53	0.72				0.85	0.69 (0.03, 0.85)
Plantar skin creases	0.34	0.02	0.72	0.76	0.77	0.56	0.64	0.76	0.65	0.79	0.64	0.87	0.69, (0.02, 0.87)
Breast size	0.25		0.8	0.75	0.76	0.66	0.57	0.76	0.62	0.89	0.66	0.81	0.75 (0.25, 0.89)
Nipple formation	0.19	0.14			0.72	0.62	0.55	0.75	0.68				0.62 (0.14, 0.75)
Skin texture	0.28	0.14	0.75	0.72	0.77	0.59	0.57	0.8			0.65	0.77	0.69 (0.14, 0.80)
Genitalia	0.17	0.02	0.82	0.66	0.65	0.36	0.62	0.63	0.43	0.67			0.63 (0.02, 0.82)
Lanugo hair	0.2	-0.01	0.81	0.62	0.73	0.55	0.49	0.71		0.77			0.62 (-0.01, 0.81)
Edema	0.16			0.59	0.64	0.67	0.22	0.41					0.50 (0.16, 0.67)
Skin opacity	0.09	0.02			0.72	0.22	0.35	0.7	0.48				0.35 (0.02, 0.72)
Nail Texture				0.57									0.57 (0.57, 0.57)
Nail Length				0.51									0.51 (0.51, 0.51)
Facial Appearance											0.77		0.77 (0.77, 0.77)
Skull hardness	0.15												0.15 (0.15, 0.15)

	Amiel-Tison (1999)	Lee (2016)	Ballard (New Ballard) (1991)	Parkin (1976)	Dubowitz and Farr (Nicolopoulos 1976)	Raghu (1981)	Feresu (2002)	Dubowitz and Farr (Sunjoh 2004)	Finnstrom (1972)	Ballard (1979)	Tuncer (1981)	Narayanan (1982)	Summary Across all Studies Median (Min, Max)
Neuromuscular Criteria													
Posture		0.12	0.82	0.75	0.72	0.31	0.65	0.76		0.69	0.48		0.69 (0.12, 0.82)
Square Window			0.79	0.21	0.73	0.58	0.64	0.69		0.7			0.69 (0.21, 0.79)
Scarf Sign	0.23	0.08	0.82	0.67	0.72	0.51	0.63	0.72		0.71	0.41		0.65 (0.08, 0.82)
Popliteal angle	0.23	0.05	0.74	0.48	0.76	0.39	0.63	0.7		0.77			0.63 (0.05, 0.77)
Arm recoil	0.19	0.07	0.71	0.62	0.65	0.29	0.55	0.56		0.61	0.36		0.56 (0.07, 0.71)
Heel to ear		0.04	0.81	0.51	0.76	0.5	0.59	0.66		0.72			0.63 (0.04, 0.81)
Leg Recoil				0.59	0.55	0.3	0.47	0.52					0.52 (0.30, 0.59)
Ventral Suspension				0.59	0.72	0.42	0.7	0.71					0.70 (0.42, 0.72)
Head Lag				0.47	0.71	0.36	0.59	0.65					0.59 (0.36, 0.71)
Ankle Dorsiflexion	0.21			0.37	0.74	0.47	0.59	0.66					0.53 (0.21, 0.74)
Non-nutritive sucking reflex	0.24												0.24 (0.24, 0.24)
crossed extension	0.16												0.16 (0.16, 0.16)
Vision: Fix and track	0.1												0.10 (0.10, 0.10)
Righting reaction	0.07												0.07 (0.07, 0.07)
Raise to sit	0.15												0.15 (0.15, 0.15)
Back to lying	0.03												0.03 (0.03, 0.03)
Finger grasp and response to traction	0.11												0.11 (0.11, 0.11)

Abbreviations: BOE= best obstetric estimate; US= ultrasound; LMP= last menstrual period

Table 3. Agreement and Validity of the Dubowitz Assessment

Author	Year	Study Setting (NICU/clinic/hospital/ community, district/city, country)	GA of cohort	Sample Size	Assessment Version (Total, Physical/ External, Neurologic)	AGREEMENT						VALIDITY			
						Correlation coefficient (R) with reference GA	Mean difference (weeks)	SD of the mean difference (wks)	Bland Altman 95% LOA [±1.96 SD] (LL,UL) [wks]	% within 1 wk	% within 2 wks	Sensitivity preterm <37wk (%) (95%CI)	Specificity preterm <37wk (%) (95% CI)	<37 wk PPV	<37 wk NPV
ULTRASOUND															
High Income Countries															
Allan[77]	2009	Royal Darwin Hospital & Darwin Private Hospital, Tiwi Northern Territory, Australia	29.6-41.7 wks	98	Total		0.10	1.10	(-2.3, 2.0)						
Roberts[79]	1979	University Hospital of Wales, Cardiff, Wales	NS	118	Total					68.6	89.8				
Vik[80]	1997	Trondheim and Bergen, Norway	All GA	970	Total		-0.20	1.12	(-2.3, 2.1)						
Awoust[81]	1982	Brugman University Hospital, Brussels, Belgium	NS	130	Total		0.50	1.04							
Sanders[33]	1991	NICU, The Johns Hopkins Hospital, Baltimore, MD, USA	<1500gm, >20 wks	110	Total	0.73	3.00			18.2	39.1				
Wariyar[38]	1997	Newcastle, UK	32-42 wks	347	Total		0.71	1.17	(-1.57, 3.0)						
			<30 wks	105	Total		2.86	2.48	(-2.0, 7.71)						
Robillard[32]	1992	Neonatology Dept, Guadalupe, French West Indies	<2500g	384	Total		0.64	1.94		61.0	82.0				
Shukla[31]	1987	New York University-affiliated hospitals; New York, USA	Preterm <38 weeks, AGA	25	Total	0.90					48.0				
Low/Middle Income Countries (LMIC)															
Moore[30]	2015	Refugee/migrant antenatal clinics, Thai-Myanmar border	All GA	250	Total		2.57[a]	1.04[a]	(0.49, 4.65)[a]			61	99		
Rosenberg[82]	2009	Special Care Nursery, Shishu Hospital, Dhaka, Bangladesh	≤33 weeks	355	Total		0.56	0.52	(-1.57, 0.47)						
Karunasekera[48]	2002	North Colombo Teaching Hospital, Ragama, Sri Lanka	35-42 weeks	200	Total		-2.18	1.43							
					External		-0.45	2.39							
LMP															
High Income Countries															
Ballard[21]	1979	NICU/nursery, Cincinnati General and Children's Hospital; Ohio, USA	NS	224	Total	0.85									
Capurro[17]	1978	Tertiary Care Center; Montevideo, Uruguay	NS	115	Total	0.91									
Mitchell[83]	1979	Newborn Nursery, Guy's Hospital, London, England	NS	20	Total	0.41									
Nicolopoulos[73]	1976	Alexandra Maternity Hospital Athens, Greece	28-44 weeks	710	Total	0.91									
					External	0.88									
					Corrected neuro	0.85									
Roberts[79]	1979	Antenatal clinics at University Hospital Wales, Cardiff, Wales	NS	118	Total					67.8	79.6				

Author	Year	Study Setting (NICU/clinic/hospital/ community, district/city, country)	GA of cohort	Sample Size	Assessment Version (Total, Physical/ External, Neurologic)	Correlation coefficient (R) with reference GA	Mean difference (weeks)	SD of the mean difference (wks)	Bland Altman 95% LOA [±1.96 SD] (LL, UL) [wks]	% within 1 wk	% within 2 wks	Sensitivity preterm <37wk (%) (95%CI)	Specificity preterm <37wk (%) (95% CI)	<37 wk PPV	<37 wk NPV
Vogt[36]	1981	Tertiary Care Center, Norway	All GA	242	Total						90[b]				
Vik[80]	1997	Trondheim& Bergen, Norway	All GA	970	Total		-0.40	1.43	(-3.2, 2.4)						
Latis[84]	1981	Neonatal unit, L. Mangiagalli Institute of Obstetrics and Gynaecology; Milano, Italy	27-42 weeks	92	Total		0.44	1.62			80.7				
Dubowitz[16]	1970	Newborn nursery, Special Care Nursery & Premature Nursery, Jessop Hospital for Women; Sheffield, England	All gestational ages	167	Total	0.93					95.0				
					External	0.91									
					Neurologic	0.89									
Allan[77]	2009	Royal Darwin & Darwin Private Hospitals, Northern Territory, Australia	29.6-41.7 wks	56	Total		0.30	0.92	(-1.5, 2.1)						
Hertz[85]	1978	Antenatal Unit, Cleveland Metropolitan General Hospital; Ohio, USA	All GA	126	Total	0.86									
Sanders[33]	1991	NICU, Johns Hopkins Hospital; Baltimore, Maryland, USA	<1500g, >20 wks	110	Total	0.68	2.80	2.1		23.6	46.3				
Low/Middle Income Countries (LMIC)															
Feresu[72]	2002	Maternity Unit, Harare Central Hospital; Harare, Zimbabwe	All GA	364	Total	0.81									
					External	0.77									
					Neurologic	0.79									
Sunjoh[35]	2004	Neonatology services, Mother and Child Centre, National Social Insurance & Central Hospitals; Yaounde, Cameroon	25-44 weeks	358	Total	0.94	0.50	1.31			93.0				
Tuncer[51]	1981	NICU, Hacettepe University Hospital, Ankara, Turkey	27-41 weeks	120	Total	0.88									
Cevit[52]	1998	Tertiary Care Center, Sivas, Turkey	28-38 weeks; <2500g	91	Total	0.85	0.30			60.4	98.9				
Jaroszewicz[86]	1973	Tyberberg Hospital, Cape Town, South Africa	NS	100	Total	0.9									
Dawodu[87]	1977	Maternity Units, University College & Oluyoro Catholic Hospitals; Ibadan, Nigeria	29-43 weeks	100	Total	0.90	0.38	1.41	(-2.39,3.15)	74.0	94.0	81.5	98.6	95.7	93.5
Raghu[34]	1981	Premature Unit, University Teaching Hospital, Lusaka, Zambia	NS	160	Total	0.90									
					External	0.82									
					Neurologic	0.80									

[a] For a 34-week newborn with weight-for-age Z score (WFAz) of 0. There was evidence of a significant trend across gestational age; mean bias decreased by 0.35 weeks per week increase in newborn GA.
[b] Percent within ±3 weeks of LMP (reference) GA.

An empty cell indicates that the data was not available for that paper. Abbreviations: NS= not stated, GA= gestational age, AGA= appropriate-size-for-gestational age, SD= standard deviation, LOA= limits of agreement, LL=lower limit, UL=upper limit, CI= confidence interval, PPV= positive predictive value, NPV=negative predictive value

Table 4. Agreement and Validity of the Ballard Assessment

Author	Year	Study Setting (NICU/clinic/hospital/ community, district/city, country)	GA of cohort	Sample Size	Assessment Version [Ballard (1979), New Ballard (1991); Physical/ External, Neuro]	AGREEMENT						VALIDITY			
						Correlation coefficient (R) with reference GA	Mean difference (weeks)	SD of mean diff. (wks)	Bland Altman 95% LOA [±1.96 SD] (LL, UL) [wks]	% within 1 wk	% within 2 wks	Sensitivity preterm <37wk (%) (95%CI)	Specificity preterm <37wk (%) (95% CI)	<37 wk PPV	<37 wk NPV
ULTRASOUND															
High Income Countries															
Scher[88]	1987	NICU, Magee-Women's Hospital; Pittsburg, Pennsylvania, USA	23-30 wks by LMP[a]	24	Ballard		1.35	2.62	(-3.79, 6.49)	56.5	69.6				
Alexander[37]	1992	Medical University Hospital; Charleston, S Carolina, USA	28-44 wks by Ballard	4193	Ballard	0.79						72.2	97.1	83.2	94.6
Sanders[33]	1991	NICU, Johns Hopkins Hospital; Baltimore, Maryland, USA	<1500g; <37 weeks	110	Ballard	0.69	2.70			22.7	45.4				
Smith[61]	1999	Hermann Hospital, Houston, Texas, USA	<2500g; 85% preterm	82	Ballard	0.86	1.40	1.15			85				
Dombrowski[89]	1992	Hutzel Hospital, Detroit, Michigan, USA	24-43 weeks	38,318	Ballard						85.4				
Gagliardi[60]	1992	NICUs, 7 tertiary care centers; Milano, Italy	<37 wks; <2500g	227	Ballard		-0.21	1.76		20.5	40.4				
Wariyar[38]	1997	Newcastle, UK	32-42 wks	347	Ballard		0.57	1.31	(-2.0, 3.14)						
			<30wks	105	Ballard		3.43	1.97	(-0.43, 7.29)						
			<30wks	105	New Ballard		1.57	1.75	(-1.86, 5.0)						
Ballard[22]	1991	NICU/nursery, 4 medical cntrs, Cincinnati, Ohio, USA	All GA; 20-44 wks	530	New Ballard	0.97	0.15	1.46							
Amato[65]	1991	Neonatal Dept, University of Berne; Switzerland	All preterm, LBW	38	Ballard (Physical)										
Low/Middle Income Countries (LMIC)															
Karl[39]	2015	8 health facilities, Madang municipality, Papua New Guinea	25.5-43.7 wks; 900g-4250g	623	Ballard	0.35	0.86	2.41	(-3.86, 5.57)			39.0	92.0	21.0	97.0
				668	(External)	0.33						58.0	81.0	14.0	97.0
				668	(Neuro)	0.39			(-3.57, 6.57)			23.0	93.0	14.0	96.0
Rosenberg[82]	2009	Special Care Nursery, Shishu Hospital, Dhaka, Bangladesh	Preterm, all <33 wks	355	Ballard		-0.41	1.08	(-0.7, 1.51)						
Lee[29]	2016	Community setting, Sylhet district, Bangladesh	33-45 wks	710	Ballard	0.12	-0.40	2.22	(-4.7, 4.0)	32.0	64	15.0	87.0	9.0	92.0
Moraes[56] (translated)	2000	Maternity unit, Instituto Fernandes Figueira, Rio de Janeiro, Brazil	NS	116	New Ballard							57.0 (41.0-73.0)	97.0 (90.0-99.0)		
Sreekumar[49]	2013	Level III NICU & postnatal wards, St. Johns Hospital, Bengaluru, India	24-41.2 weeks	284	New Ballard		-0.04								
Wylie[90]	2013	Ndirande Antenatal Care Clinic, Blantyre, Malawi	All GA	177	New Ballard		0.80	2.19	(-3.5, 5.1)						
Taylor[68]	2010	Nurse-trekking teams & community medical station, Keneba, The Gambia	All GA	80	Ballard (External)		-2.23	1.56	(-5.3, 0.82						
Thi[91]	2015	Hoa Binh General Hospital, Hoa Binh province, Vietnam	30-42 wks by US	391	New Ballard	0.90									

Author	Year	Study Setting (NICU/clinic/hospital/ community, district/city, country)	GA of cohort	Sample Size	Assessment Version (Original, New Ballard (NB); Physical/ External, Neuro)	Correlation coefficient (R) with reference GA	Mean difference (weeks)	SD of mean diff.	Bland Altman 95% LOA [±1.96 SD] (LL,UL) [wks]	% within 1 wk	% within 2 wks	Sensitivity preterm <37wk (%) (95%CI)	Specificity preterm <37wk (%) (95% CI)	<37 wk PPV	<37 wk NPV
LMP															
High Income Countries															
Baumann[42] (translated)	1993	University Clinic-Bern, Bern, Switzerland	27-35 wks AGA	60	Ballard (Total)	0.91									
			28-36 wks SGA	29		0.66									
			27-35 wks AGA	60	Ballard (External)	0.83									
			28-36 wks SGA	29		0.66									
			27-35 wks AGA	60	Ballard (Neuro)	0.65									
			28-36 wks SGA	29		0.66									
Constantine[40]	1987	8 states (AK, NY, MA, FL, PA, TX, WA, CN), USA	All GA	1246	Ballard	0.81	0.60	2.18				85	81	89	75
					(Physical)	0.83	-0.1	2.14				92	74	87	87
					(Neuro)	0.71	1.4	2.72				70	84	89	60
Scher[88]	1987	NICU, Magee-Women's Hospital, Pittsburgh, Pennsylvania, USA	23-30 weeks by LMP	24	Ballard		1.42	2.32	(-3.13, 5.96)	45.8	62.5				
Mackanjee[92]	1996	NICU, St. Joseph's Health, London, Ontario, Canada	23-33 weeks; <1500g	47	Ballard	0.87	1.50	1.50							
Dombrowski[89]	1992	Hutzel Hospital, Detroit, Michigan, USA	24-46 weeks	38,818	Ballard						69.9				
Alexander[93]	1990	Medical University Hospital, Charleston, S Carolina, USA	20-45 weeks	10,794	Ballard	0.76	0.48			52.7	80.3				
Ballard[21]	1979	NICU/nursery, Cincinnati Gen. & Children's Hospitals; Ohio, USA	26-44 weeks, 760-5460g	224	Ballard	0.85									
Alexander[94]	1992	Medical University of South Carolina Hospital, Charleston, South Carolina, USA	28-44 wks; all black population	3480	Ballard	0.82	0.53				68.2				
			28-44 wks; all white population	2091	Ballard	0.86	0.17				70.6				
Ballard[22]	1991	NICU/nursery, 4 medical centers, Cincinnati, Ohio, USA	20-44 weeks	578	New Ballard	0.96					88.0				
Ahn[95]	2008	Neonatal units, University hospital, Incheon, S. Korea	All GA, 773-4870g	213	New Ballard[b]	0.85	0.46[c]								
Sanders[33]	1991	NICU, Johns Hopkins Hospital; Baltimore, Maryland, USA	<1500g; <37 weeks	110	Ballard	0.66	2.60	2.2		28.2	51.0				

Author	Year	Study Setting (NICU/clinic/hospital/ community, district/city, country)	GA of cohort	Sample Size	Assessment Version (Original, New Ballard [NB]; Physical/ External, Neuro)	Correlation coefficient (R) with reference GA	Mean difference (weeks)	SD of mean diff.	Bland Altman 95% LOA [±1.96 SD] (LL,UL) [wks]	% within 1 wk	% within 2 wks	Sensitivity preterm <37wk (%) (95%CI)	Specificity preterm <37wk (%) (95% CI)	<37 wk PPV	<37 wk NPV
Low/Middle Income Countries (LMIC)															
Cevit[52]	1998	Tertiary Care Center, Sivas, Turkey	Preterm 28-38 wks; <2500g	91	Ballard		0.10			59.3	98.9				
Feresu[72]	2002	Maternity Unit, Harare Central Hospital; Zimbabwe	24-45 weeks	364	Ballard	0.80									
					(Physical)[d]	0.75									
					(Neuro)[d]	0.74									
Sunjoh[35]	2004	Mother & Child Centre Hospitals; Cameroon	25-44 weeks	358	New Ballard	0.93	0.34	1.52			86.0				
Bindusha[50]	2014	Tertiary care hospital, Kerela, India	28-37 weeks	1000	New Ballard	0.92	0.31					<36 wk: 85.6	<36 wk: 94.6	<36 wk: 98.0	<36 wk: 53.6
Sasidharan[59]	2009	Level III NICU, medical institute, Northern India	29-35 weeks	129	New Ballard						100.0				
Moraes[56] (translated)	2000	Maternity unit, Instituto Fernandes Figueira, Rio de Janeiro, Brazil	NS	140	New Ballard							68.0 (49.0 - 82.0)	92.0 (85.0 - 96.0)		
Thi[91]	2015	Hoa Binh General Hospital, Hoa Binh province, Vietnam	30-43 wks by LMP	282	New Ballard	0.81									
Taylor[68]	2010	Nurse-trekking teams and community medical station, Keneba, The Gambia	All GA	76	Ballard (External)		-2.2	3.3	(-8.67, 4.27)						
Verhoeff[41]	1997	Chikwawa District Hospital & Montfort Hospital, Southern Region, Malawi	All GA; literate mothers	76	Ballard (External)		0.87								

[a] All infants in this study died; all deaths occurred after the assessments.
[b] This study used an "Extended New Ballard" scoring system to estimate gestational age (simply the standard NB score extended to be used to estimate a greater GA range, which was calculated mathematically).
[c] For infants <39 wks GA. Mean difference= -0.58 wks for infants >39 wks GA.
[d] This study used a "revised" version of the physical and neurological portions of the Ballard assessment.

An empty cell indicates that the data was not available for that paper. Abbreviations: NS= not stated, GA= gestational age, SD= standard deviation, LOA= limits of agreement, LL=lower limit, UL=upper limit, CI= confidence interval, PPV= positive predictive value, NPV=negative predictive value, AGA= average-for-gestational-age, SGA=small-for-gestational-age

Table 5. Pooled Data for Agreement and Validity of Neonatal Clinical Assessments

			AGREEMENT							VALIDITY		
			Mean Difference			% within 1 week		% within 2 weeks		Sensitivity		Specificity
Assessment Type	# of studies identified	Reference Standard	N	Pooled Difference	Pooled Std. Dev.	N	Pooled % (95% CIs)	N	Pooled % (95% CIs)	N	Pooled Sensitivity (%) (95% CIs)	Pooled Specificity (%) (95% CIs)
Dubowitz	9	US/BOE	7	0.02 (-0.51, 0.55)	1.27	3	48.1 (23.4, 73.8)	3	74.5 (40.4, 92.7)	1	61	99
	20	LMP	6	0.65 (0.01, 1.30)	1.45	4	56.5 (33.7, 76.8)	6	87.1 (69.8, 95.2)	1	81.5	98.6
Ballard	14	US/BOE	9	0.40 (0.00, 0.81)	1.90	3	33.5 (22.3, 46.8)	5	72.0 (53.1, 85.3)	4	64.1 (60.8, 67.4)	95.1 (94.5, 95.7)
	18	LMP	5	1.25 (0.64, 1.87)	2.10	3	43.9 (23.9, 66.1)	9	75.4 (70.3, 79.8)	2	84.1 (81.6, 86.3)	83.5 (79.5, 87.0)
Parkin	3	US/BOE	3	-0.17 (-0.26, -0.08)	1.97	0		0				
Eregie	2	LMP	1			0		2	93.4 (91.3, 95.1)			
Capurro	4	US/BOE	2	0.11 (-0.02, 0.23)	1.96	2	40.1 (34.7, 45.8)	3	79.2 (65.3, 88.6)	3	42.7 (35.6, 50.0)	96.7 (95.7, 97.5)

Abbreviations: US/BOE= ultrasound or best obstetric estimate; LMP= last menstrual period; CI= confidence interval

Table 6. Correlation of Anterior Vascular Capsule of the Lens (AVCL) with Gestational Age

Author	Year	Study Setting (hospital, district/country)	Population	Sample Size (N)	Reference Standard	Time of assessment after birth	Correlation coefficient (R) with reference gestational age	Gestational Age A) Range, or B) Mean (standard deviation) [N] Grade[a] 0	Grade 1	Grade 2	Grade 3	Grade 4
Finnstrom[53]	1972	University Hospital, Umea, Sweden	All GA	174	LMP	From birth up to 60 hours	0.45[b]					
Hittner[62]	1977	Jefferson Davis Hospital, Houston, USA	Preterm (27-34wks)	100	LMP & Dubowitz	Within 30 h	-0.88					
			Sub-population: Preterm SGA	12	LMP & Dubowitz	Within 30 h	-0.91					
Guillory[96]	1980	Hospital, Houston, TX	Preterm	43	LMP & Dubowitz	"Soon after birth"	-0.88					
						24 h after birth	-0.86					
Hittner[64]	1981	Tertiary Care Facility, Houston, USA	"Preterm SGA"	33	Dubowitz	Within 24 h	-0.77		A) ≥33wks [n=24[c]]	A)31-34wks [n=7]	A) 33 wks [n=1]	A) 28 wks; [n=1]
Krishnamohan[97]	1982	NICU, University of Connecticut Hospital & Hartford Hospital, Connecticut, USA	Preterm (28-32 wks)	30	Ballard, within 2 weeks of LMP	Within 24 h	-0.94					
Narayanan[54]	1982	Kalawati Saran Children's Hospital, New Dehli, India	All newborns; all gestational ages	356	LMP, or OB estimate if available	Within 48 h	-0.64					
			Sub-population: <35 wks GA	184	Same as above	Within 48 h	-0.96					
Sasivimolkul[63]	1986	Ramithibodi Hospital, Bangkok, Thailand	Low birth weight (LBW)	80	Ballard & LMP	24-48 h	-0.839	B) 36.3 (1.86) [n=43]	B) 34.0 (2.1) [n=13]	B) 32.4 (1.4) [n=12]	B) 29.9 (0.4); [n=7]	B) 27.8 (0.8) [n=5]
			Sub-population: LBW ≥34 wks	40	Ballard & LMP	24-48 h	-0.88					
Skapinker[67]	1987	Johannesburg Hospital, Johannesburg, South Africa	Preterm <35 wks	58	Ballard	Within 36 h	-0.84					
Sanders[33]	1991	NICU, Johns Hopkins Hospital, Baltimore, Maryland, USA	Preterm AND Birthweight <1500g	89	BOE (US was available for 92% of women)	Within 72 h		B) 32.4	B) 30.4	B) 29.8	B) 28.7	B) 26.7
Baumann[42] (translated)	1993	University Clinic- Bern, Bern, Switzerland	<34 wks AGA	60	US	NS	-0.92 ± 0.04 (CI: 0.81-0.97)					
			<34 wks SGA	29	US	NS	-0.68 ± 0.09 (CI: 0.49-0.82)					

[a]The AVCL grading system is as described in Hittner et al, 1977[62]

[b]Finnstrom used the Harnack and Oster (1958) grading system, a classification system with grades 1-3, in which 1=most vascularity and 3= no vascularity. Therefore, the correlation between disappearance of the AVCL and increasing GA is noted as positive under this classification system, but would be negative by the Hittner grading system.

[c]N=24 for both **Grades 1 & 0 combined**; the GA range stated (≥ 33wks) comprises infants that scored **either** a 1 or 0.

An empty cell indicates that the data was not available for that paper. Abbreviations: NS= not stated, CI= confidence interval, LMP= last menstrual period, OB estimate= obstetric estimate, BOE= best obstetric estimate, US= ultrasound