Statistical issues in the study of fetal and neonatal growth

Eric Ochieng Ohuma

Nuffield Department of Obstetrics & Gynaecology
Green Templeton College

University of Oxford

Submitted in partial completion of the degree of

Doctor of Philosophy

Trinity Term 2016

To my dearest, lovely wife Ella and beautiful daughter Megan.

To my parents John and Margaret for their endless love, support and encouragement.

Acknowledgements

I wish to thank my main supervisor, Professor Douglas G. Altman, for his excellent mentorship, guidance, and support throughout this process. I will always be indebted and feel privileged to have had the opportunity to work alongside you and benefit from your wisdom, humility, and understanding. Many thanks to my second supervisor, Professor Josè Villar, for giving me the opportunity to work on this project and for the support accorded in pursuing my DPhil. Your belief in me is highly appreciated and I have benefited greatly from your wisdom, leadership, and enthusiasm.

I wish to thank the Nuffield Department of Obstetrics and Gynaecology (NDOG)'s, INTERGROWTH-21st Project at the University of Oxford for the departmental scholarship that funded my DPhil. Special thanks go to all members of the INTERGROWTH-21st Project. The experience and training that I have received have been instrumental in developing the perspective that I have on maternal and child health.

I would like to thank Dr Jennifer de Beyer for kindly agreeing to read and edit earlier drafts of my chapters and for proofreading my thesis. To Michael, Steve and Rachel for your friendship and the 'interesting' lunch time and pub conversations we have had over the years – always been a good laugh!

I am greatly indebted and thankful to my parents, John and Margaret Ohuma, for the great sacrifices you have made to support my education through difficult circumstances, enabling me to achieve my ambition to go to university. To my siblings, Lilian, George, Eunice, and Cosmas, thank you for your encouragement and support, and for always just being yourselves! To all my friends, thank you for your support and encouragement. I cannot list all of your names here, but you are always on my mind.

For his personal support and encouragement, I thank my father-in-law, Hayman Wheaton. Also thanks to my mother-in-law, Christine Wheaton, and sister-in-law, Lorna Wheaton, for all of your support, especially in babysitting Megan. Finally, I wish to thank my lovely wife Ella for the support, love, and encouragement throughout. Thanks so much for always stepping in to look after our beautiful daughter Megan during this period.

Statement of originality

This thesis is the result of research carried out between October 2012 and February 2016 in the Department of Obstetrics and Gynaecology at the University of Oxford. I declare that the material presented in this thesis has not been submitted previously for any degree in either this or any other university.

The thesis was integrated within the INTERGROWTH-21st Project, for which I was the medical statistician. I was responsible for all of the analyses undertaken and for writing up parts of the statistical methodology and results for the journal articles that have been published on this project. This thesis delves further into the details of the project, providing further unpublished information on the statistical methodology and results.

The work for this thesis is entirely my own, unless stated otherwise. The statistical software programmes used to carry out the analyses are the statistical programming language R discussed in Ihaka and Gentleman (1996) and STATA version 11.2 (StataCorp LP, College Station, Texas, USA). The copyright of this thesis rests with the author. No quotation from it should be published without prior written consent and any information derived from it should be acknowledged.

Abbreviations

AIC: Akaike information criterion
BCCG: Box-Cox Cole and Green distribution
BCPE: Box-Cox power exponential
BHC: Birth head circumference
BIC: Bayesian information criterion
BL: Birth length
BPD: Biparietal diameter
CI: Confidence interval
CRL: Crown-rump length
edf: equivalent degrees of freedom
FGLS: Fetal Growth Longitudinal Study
FHC: Fetal head circumference
FP: Fractional polynomial
GA: Gestational age
GAMLSS: Generalised additive models for location, scale, and shape
HC: Head circumference
INTERGROWTH-21st: The International Fetal and Newborn Growth Consortium for the 21st Century
IQR: Interquartile range
ISUOG: International Society of Ultrasound in Obstetrics and Gynaecology
LMP: Last menstrual period
MGRS: Multicentre Growth Reference Study
NCSS: Newborn Cross-Sectional Study
NDOG: Nuffield Department of Obstetrics and Gynaecology
NICE: National Institute for Health and Care Excellence
PPFS: Preterm Postnatal Follow-up Study
SD: Standard deviation
SE: Standard error
SSD: Standardised site difference
USQU: Ultrasound quality unit
WHO: World Health Organization

Details of ethics approval

The INTERGROWTH-21st Project was approved by the Oxfordshire Research Ethics Committee 'C' (reference: 08/H0606/139) and the research ethics committees of the individual participating institutions and corresponding health authorities where the project was implemented.

Abstract

Eric Ochieng Ohuma | Green Templeton College | DPhil in Obstetrics & Gynaecology | Trinity Term 2016

Human growth begins at conception and continues into adult life. Growth is usually classified as normal or abnormal using the expected attained size at a given age. The statistical analysis of growth data has been of interest to many academics from a wide range of disciplines over the last century. The study of fetal and neonatal growth is complex, and the many remaining clinical questions require complex statistical methodology. In this thesis, I have focused on growth in the prenatal period, namely fetal and newborn growth.

The thesis highlights the potential pitfalls in methodology, statistical methods, and reporting of studies aimed at creating fetal charts. I propose a checklist for evaluating the methodological quality of studies that provides a rough guide of the minimum information that should be reported in these studies. This checklist is not intended to commend or discard studies, but rather to act as a consensus guideline to improve consistency and as a guide for evaluating similar studies for future research in human growth studies. The thesis assesses approaches for developing charts of attained size at a given age and the velocity gain (rate or speed of growth) of a fetus. It aims to address some of the statistical issues that relate to fetal and neonatal growth studies.

The thesis also focuses on fetal velocity. This work is novel as there are currently no fetal velocity standards in existence. The methodologies discussed have previously been applied to child growth data, but not to fetal data, to the best of my knowledge. In this thesis, I construct the first fetal growth velocity standards. Such standards were not yet available largely due to lack of appropriate, good quality longitudinal fetal data.

All of the work and research carried out as part of the thesis was embedded within the International Fetal and Newborn Growth Consortium for the 21^{st} Century (INTERGROWTH-21^{st} Project). The primary aims of the INTERGROWTH-21^{st} project were to produce international growth standards for practical clinical applications and to monitor trends in populations using three sets of data describing fetal growth (n=4,321 women), postnatal growth of preterm infants (n=201 preterm infants), and newborn size (weight, length, and head circumference) for gestational age (n=20,486 newborns). The design and conduct of these studies are detailed elsewhere and have already resulted in numerous publications, including four papers published in *The Lancet*. I was the responsible statistician for these publications and undertook all of the statistical analyses.

Contents

List of Figures

List of Tables

1

General Introduction

The study of human growth is complex as it begins at conception and continues into adult life. It is important to monitor and understand the process of human growth. In general, the purpose of growth monitoring in general is to detect poor nutrition (which is an indicator of the general health or nutrition of an individual or population) and growth disorders, such as growth hormone deficiency, and their possible consequences [1]. A growth reference chart can be used to monitor a fetus's growth to identify if it is at increased risk of intrauterine growth restriction [2]. This risk identification is not simple as it requires an in-depth understanding of the factors that can affect or distort a fetus's growth trajectory [3, 4, 5]. Anthropometric data, sometimes coupled with clinical data, are essential for identifying potential risk factors affecting normal growth [6].

Anthropometric assessment of fetal size is of interest in many disciplines, such as statistics, medicine, nutrition, education, and anthropology. The link between health and disease status in future adult life was first demonstrated by Barker and his colleagues [7, 8, 9] and led to the fetal origins hypothesis (often called Barker's

hypothesis) [10]. Barker's hypothesis emerged from epidemiological studies of birth and death records that revealed a high geographical correlation between rates of infant mortality and certain classes of later adult deaths, as well as an association between birthweight and rates of adult death from ischemic heart disease [7]. These observations led to the theory that undernutrition during the prenatal period was an important early origin of adult cardiac and metabolic disorders. This link was attributed to fetal programming that permanently shaped the body's structure, function, and metabolism, and contributed to adult disease. This theory stimulated interest in the fetal origins of adult disorders and generated a new field of research that led to the formation of an international society for the Developmental Origins of Health and Disease (DOHaD) (https://dohadsoc.org/).

In this thesis, I focus on growth in the prenatal period, which refers to the period between conception and birth. Fetal growth is influenced by many factors, many of them maternal, such as weight, height, weight gain during pregnancy, smoking, and age. Factors such as poor environmental conditions and the nutritional status of the mother can result in impaired fetal growth and preterm birth [6], which are the leading causes of neonatal and infant mortality worldwide [11, 12]. Recognition of pathological growth is dependent on the existence of reliable reference charts. A reference chart enables a fetus to be compared to the general population at a particular time. Inferences made from these comparisons are dependent on how accurately the sample population represents the general population. The distribution of a growth reference is usually summarised by selected centiles that, in the case of fetal dimensions, are symmetric about the median (50^{th} centile). The extreme centiles, such as the 3^{rd} and 97^{th} centiles, are subject to greater sampling error than the more central centiles, but are also the target for identifying those most at risk.

Monitoring growth from conception poses many statistical challenges. For example, pregnancy is difficult to accurately date and fetal dimensions are subject to measure-

ment error and biological variation. Other issues include missing data, correlated repeat measurements, and the correct interpretation of the attained fetal size.

Reference charts are globally the most widely used growth charts, yet little work has been done to improve the methodology used to develop them. Many reference charts are currently available and in use. Wide variations exist between their centile values. These differences are largely attributable to methodological heterogeneity rather than racial [13], gender [14], or other biologic and demographic determinants [15], as the charts are based on different populations with great differences in sample selection, methodology, and statistical modelling methods [16, 17]. The use of a suboptimal methodology to produce a fetal size curve is likely to affect its ability to discriminate between healthy and compromised fetuses. These charts directly affect the identification of at-risk fetuses and newborns in need of treatment and nutritional strategies, so must be as accurate as possible.

The World Health Organization (WHO) Multicentre Growth Reference Study (MGRS) developed a set of child growth standards from birth to five years old that can be applied to different populations [18]. The International Fetal and Newborn Growth Consortium for the 21^{st} Century (INTERGROWTH-21^{st}) Project used a similar approach to construct standards to be used from the prenatal period to birth.

This thesis focuses on addressing statistical issues that relate to fetal and neonatal growth globally. The content is embedded within the INTERGROWTH-21^{st} Project, which is a large-scale, population-based, multi-centre project involving health institutions from eight geographically diverse countries (Brazil, China, India, Oman, Kenya, UK, USA, and Italy). The project aims were to assess fetal, newborn, and preterm growth under optimal conditions, using a similar approach to the MGRS [19]. This approach selects women regarded to be "healthy", educated, affluent, and living in areas with minimal environmental constraints on growth [20].

The INTERGROWTH-21^{st} Project (http://www.intergrowth21.org.uk/) has three

major components, which were designed to create: (a) longitudinally derived, prescriptive, international fetal growth standards using both clinical and ultrasound measures, called the Fetal Growth Longitudinal Study (FGLS); (b) preterm, postnatal growth standards for those infants born $\geq 26^{+0}$ but $<37^{+0}$ weeks of gestation in the longitudinal cohort; and (c) newborn size standards for birthweight, birth length (BL), and birth head circumference (BHC) according to gestational age (GA) constructed from all of the newborns delivered at the eight study sites over a period of approximately 12 months, this study was referred to as the Newborn Cross-Sectional Study (NCSS) [20]. To ensure that the ultrasound measurements were accurate and reproducible, the study centres used uniform methods, identical ultrasound equipment, a standardised methodology for taking fetal measurements, and employed locally accredited ultra-sonographers who underwent standardisation training and monitoring. All of the included women had a reliable estimate of GA confirmed by ultrasound measurement of the fetal crown-rump length (CRL) in the first trimester [21].

The design and conduct of these three components of the INTERGROWTH-21st Project studies have been detailed elsewhere [12, 20, 22, 23]. The FGLS and NCSS components have already resulted in numerous publications, including three papers published in *The Lancet* for which I was the responsible statistician and undertook all of the statistical analyses [12, 24, 25, 26].

The INTERGROWTH-21st team conducted systematic reviews to: (a) identify all published studies that aimed to construct charts to predict GA from CRL, charts of fetal biometry, and newborn size charts; (b) rank all of the studies included in the reviews according to a set of predefined methodological quality criteria; (c) evaluate how well these predefined methodological criteria discriminated between studies based on the ranking; and (d) propose a checklist of predefined methodological criteria for future studies of human growth. This chapter gives a brief overview

of the thesis, then summarises the findings from the three systematic reviews on charts for pregnancy dating, fetal biometry, and neonatal size [27, 28, 16]. I was one of the three team members that developed the original checklist to be used for the pregnancy dating review and this checklist was further modified slightly for the remaining two reviews (review of fetal size and neonatal size charts) as necessary. I contributed in all aspects of the three reviews by reading all the papers included in the review, extracted and scored all the statistical method sections of the papers and did all the analyses for the three reviews.

1.1 Thesis overview

The focus of this thesis is fetal growth in the prenatal period and newborn growth. Chapter 1 gives a general introduction that is relevant to the study of human growth studies. I summarise the findings and discuss some of the results from three systematic reviews of reference charts for pregnancy dating and fetal and newborn size conducted prior to the INTERGROWTH-21st Project. The main purpose of conducting the reviews was to evaluate the charts currently in clinical use and assess their methodological quality so that their general applicability could be judged. The gaps identified in the literature for studies constructing such charts, especially statistical methodology and reporting gaps, were informative for ensuring that the INTERGROWTH-21st Project avoided similar problems.

The results of the three systematic reviews revealed great heterogeneity in the existing literature, especially regarding statistical methodology and reporting. These results, informed the work presented in Chapter 2, on the design and methodological issues that must be considered when constructing human fetal and neonatal size and growth charts. The aim of this work was to promote understanding and result in improvements in the overall conduct of such studies in the future.

Chapter 3 evaluates methods for assessing whether the fetal and newborn data from

the eight INTERGROWTH-21st Project sites were similar enough to be combined to construct international growth charts.

In Chapters 4 and 5, I review some of the statistical methodology and analytical approaches for cross-sectional and longitudinal studies. Using the INTERGROWTH-21st Project data for demonstration purposes, I apply these analysis methods and discuss other statistical considerations when analysing these types of data. The methods in these chapters were used to develop the new international standards for fetal size [24] and newborn size [25] published by the INTERGROWTH-21st Project.

Chapter 6 discusses and evaluates an uncommon feature of data truncation encountered while modelling CRL as a function of GA to develop an equation for pregnancy dating.

Chapter 7 moves from the concept of attained size charts to growth velocity charts. I discuss two different methodologies for constructing fetal growth velocity charts i.e. velocity increments and velocity gain approaches. This chapter also discusses regression to the mean and how to account for it in relation to growth velocity.

In Chapter 8, I provide general concluding remarks on the issues discussed within the scope of the thesis, indicate the strengths and limitations of the work discussed, and discuss areas for future work.

1.2 Methodology used to conduct the systematic reviews

The three systematic reviews of published reference charts were conducted and reported using the checklist proposed by the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group [29]. All of the major electronic databases, MEDLINE, PUBMED, EMBASE, CINAHL, and secondary reference sources were systematically searched.

Once the search strategies had been developed, two reviewers screened the titles and abstracts of all identified citations. For all three reviews, eligible studies were those that developed a new equation, rather than just compared existing equations. The full-text versions of eligible studies were independently assessed by the same reviewers. The reference lists of the retrieved full-text articles were examined for additional, relevant citations.

A list of criteria for assessing the methodological quality of studies that develop reference charts was developed before the systematic reviews were conducted. The criteria were discussed and agreed between three researchers, independently of the researchers who performed the data abstraction. The agreed methodological quality criteria were similar for the three reviews, with a few modifications applied to each review as appropriate. The list of the criteria used is provided in Appendix A for pregnancy dating charts, Appendix B for fetal biometry charts, and Appendix C for newborn reference charts. The criteria were divided into three main domains: study design, statistical methods, and reporting methods.

All of the included studies were assessed against each agreed criterion and assigned a binary score of 0 or 1, corresponding to a high or low risk of bias classification, respectively. If there was insufficient information available to make a judgment about some items, then they were scored as 'not evaluable'. Disagreements were resolved either by consensus or consultation with a third reviewer. The overall quality score was defined as the sum of low risk of bias marks over the total number of criteria for each review. All non-evaluable items were neither awarded a score of zero or one and were therefore not counted in the computation of the overall percentage quality score for the respective studies. Domain-specific scores for the three domains were also computed. Medians (interquartile range [IQR] and range) were calculated as the summary measure of the distribution of scores.

1.3 Review of pregnancy dating reference charts

The first systematic review concerned pregnancy dating reference charts. A reliable estimate of GA is essential as it allows appropriate scheduling of a woman's antenatal care, informs obstetric management decisions, allows the expected delivery date to be estimated, and facilitates the correct interpretation of any fetal size assessment [28]. Abnormal fetal growth patterns such as growth restriction or macrosomia may be missed or incorrectly diagnosed if the GA is unknown or incorrect.

Traditionally, the GA of a fetus is calculated from the first day of the last menstrual period (LMP). This is about two weeks before ovulation on average, but can vary by up to six days in either direction. Therefore 40 weeks is the most frequent GA at birth, but in fact represents a true fetal age of 38 weeks [30]. Gestational lengths from 37 to 42 weeks are regarded as term, whereas babies born earlier than 37 weeks are considered to be preterm.

The use of LMP alone for pregnancy dating should be interpreted with caution as up to 50% of women are uncertain of their dates, have an irregular cycle, have recently stopped using the oral contraceptive pill, are lactating, or did not have a normal LMP [31]. The LMP method has been shown to be unreliable, even in women with a known menstrual history [32, 33].

Pregnancy dating by measuring the fetal CRL using an ultrasound scan is widely accepted as the most reliable method for dating pregnancy between 9^{+0} to 13^{+6} weeks gestation, or the first trimester [34]. First trimester scans are also recommended for confirming viability and determining the number of fetuses [35, 36]. More correct ultrasound dating in the first trimester may improve first trimester screening for chromosomal abnormalities [37] and improve classification of fetuses as either preterm, term, or post-term [38, 39]. Dating a pregnancy in the first trimester rather than the second trimester reduces the number of unnecessary inductions of

labour by improving the accuracy of the predicted pregnancy outcome [40, 41, 42].

Robinson *et al.* reported the first equation developed for pregnancy dating based on CRL in 1975 [43]. Since then, several equations have been reported and numerous validation studies have been published. Many dating charts are thus now in use [27]. These charts have been developed using different populations, population recruitment strategies, and statistical methodologies, among other variations. The lack of consensus over which of these formulas should be used makes harmonising clinical practice difficult. The lack of accurate GA estimates means that preterm birth and small-for-GA rates are often inaccurate, particularly in geographical regions at greatest risk of these conditions [44, 45].

The systematic review of pregnancy dating charts identified 29 studies from 14 countries that aimed to construct charts to predict GA from CRL [27]. The studies were scored, using a set of 29 criteria, as having a low or high risk of bias based on the study design (12 items) and statistical (6 items) and reporting (11 items) methods used. This produced a wide range of scores, showing that the quality of the studies was variable (median = 15, range = 5-21): 9 studies scored >15/29 and 6 studies scored <12/29. As the charts were developed using different populations and different methodologies, it is unsurprisingly that their estimated GAs at any given CRL varied widely [27]. The studies' inclusion criteria, exclusion criteria, and maternal demographic characteristics generated the highest potential for bias. None of the studies had a systematic method for checking the quality of their ultrasound measurements (Figure 1.1). The four studies with the highest scores (lowest risk of bias) satisfied 18 or more of the 29 checked criteria. They also had less variation between their GA estimates than the variation between the remaining lower-scoring studies [36, 46, 43, 47] (Figure 1.2).

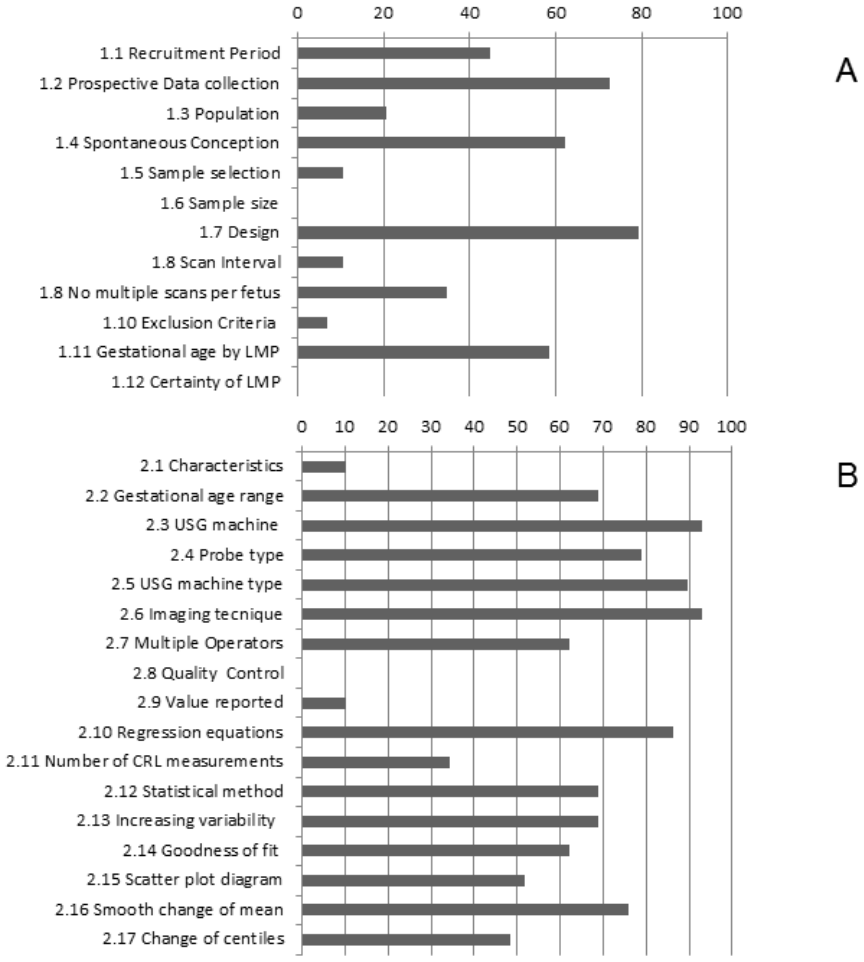


Figure 1.1: Aggregated methodological assessment of reporting quality of the studies included in the systematic review of pregnancy dating charts expressed as a percentage of the studies with a low risk of bias for each (A) study design and (B) reporting and statistical method criterion.

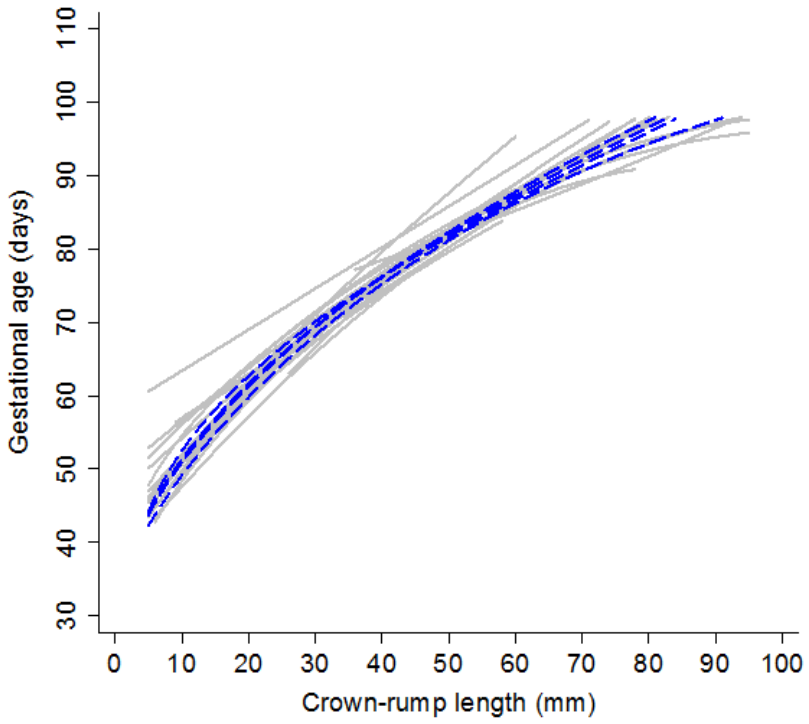


Figure 1.2: Gestational age charts from 24 of the 29 included studies. The four studies with the highest methodological quality scores are shown in blue [43, 46, 47, 36] and the remaining 20 are shown in grey. For three of the studies, [32, 48, 49], the chart was based on data from tables in the study publication. Data are not shown for five of the included studies. Two of these studies [50, 51] did not provide an equation or table. The figures in the study publications of the remaining three studies, [52, 53, 54] could not be reproduced from the equations given and no tables were provided.

1.4 Review of fetal biometry reference size charts

The second systematic review investigated fetal biometry reference size charts. Screening for disturbances in fetal growth is one of the main purposes of antenatal care. As biomarkers of fetal growth restriction have little clinical utility at present, [55] screening relies on routine measurement of uterine fundal height, complemented by ultrasound measurement of fetal size in women with a relevant history, pregnancy complications, or clinical evidence of fetal growth restriction. Prenatal evaluation of fetal size was made possible in the 1960's by the introduction of combined A and B-mode ultrasonography in obstetrics [56]. Second and third trimester fetal biometry

with real-time 2D ultrasound is now common practice in developed countries and is one of the most common medical investigations undertaken. However, a Cochrane review of routine ultrasonography in low-risk pregnancies has not demonstrated any benefit in perinatal outcomes [57].

Intrauterine growth restriction remains a leading cause of perinatal loss, accounting for at least one fifth of stillbirths in the UK; failure to diagnose and lack of treatment are the most likely explanations. Fetal size reference charts are used to assess prenatal size. Despite the widespread use of ultrasound, concerns have been expressed about the low detection rates of abnormal fetal growth in routine practice [57, 58], even when ultrasound is used mostly in high-risk subpopulations. However, these observations should be interpreted with caution given the large number of locally derived reference charts available, the wide variation in the methodologies used to create such charts [28], and the lack of suitable international standards like those for monitoring child growth [18]. Large variations are also seen in the cut-off points (e.g., 3^{rd}, 5^{th}, or 10^{th} centiles) commonly used for determining or making decisions whether fetal growth is abnormal, even within the same population or region. The use of such a variety of charts and cut-off points [28, 27] in clinical decision-making about fetal growth patterns inevitably leads to diagnostic confusion, difficulties comparing outcomes across populations, and unnecessary anxiety for mothers and their families.

The systematic review of the methodology used in ultrasound studies that aimed to create charts of fetal size identified 83 studies from 32 countries [28]. The studies were scored, using a set of 25 criteria, as having a low or high risk of bias based on the study design (10 items), statistical methods (7 items), and reporting methods (8 items) used. The frequency of low bias scores in each of the three groups of methodological criteria are presented in Figures 1.3, 1.4, and 1.5. Again, the highest potential for bias was noted in the inclusion and exclusion criteria, the ultrasound

quality control measures, and the sample size calculation. Only six of the studies reported their sample size calculations [28]. Pregnancy dating by LMP was the most popular method, 44%, followed by LMP confirmed with a non-CRL ultrasound, 26%, LMP confirmed with CRL, 9%, and ultrasound alone, 9%. Five studies did not state which method was used for dating and another five used other methods.

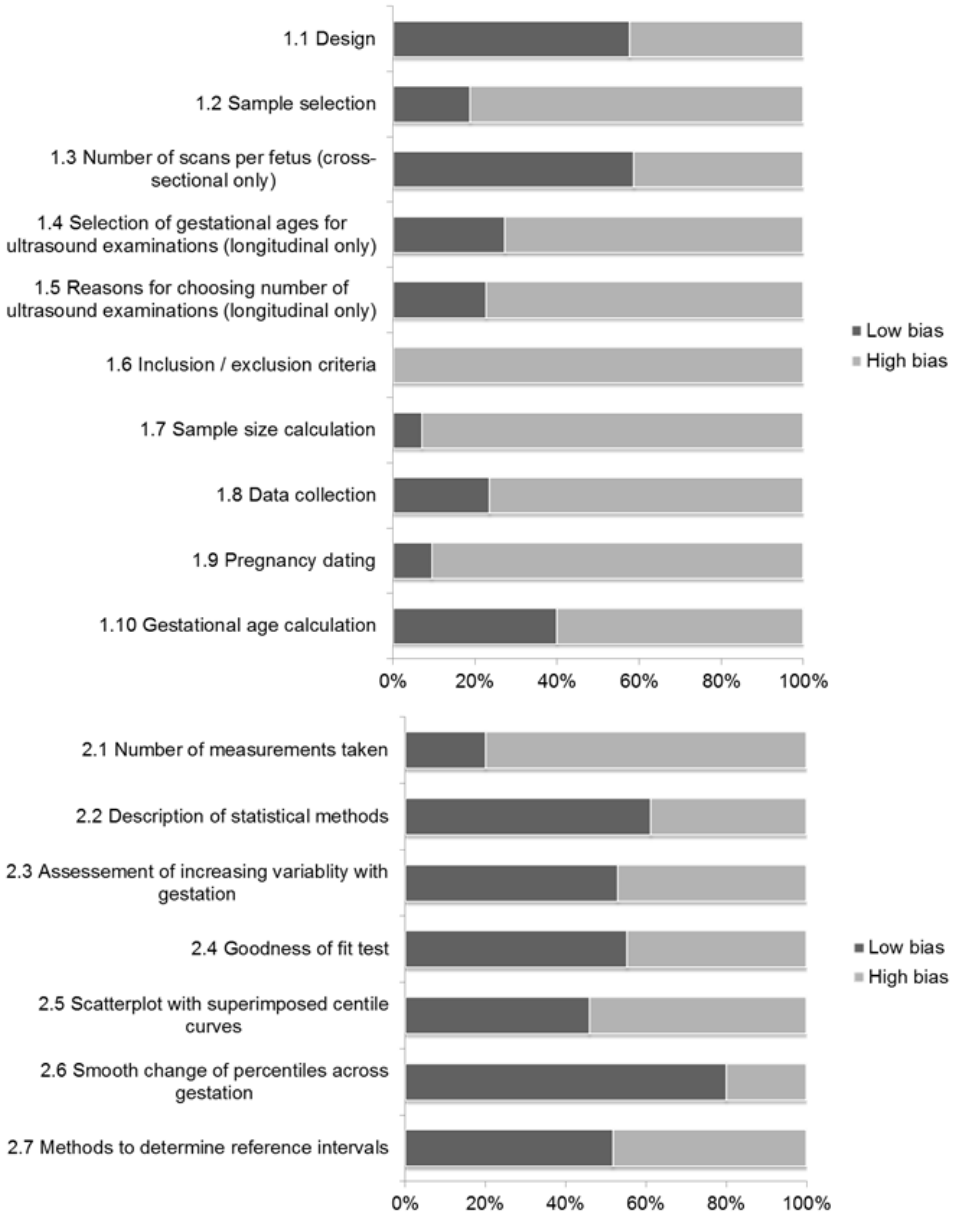


Figure 1.4: Aggregated methodological assessment of reporting quality of the studies included in the systematic review of fetal biometry reference charts expressed as a percentage of the studies with a low or high risk of bias for each reporting criterion.

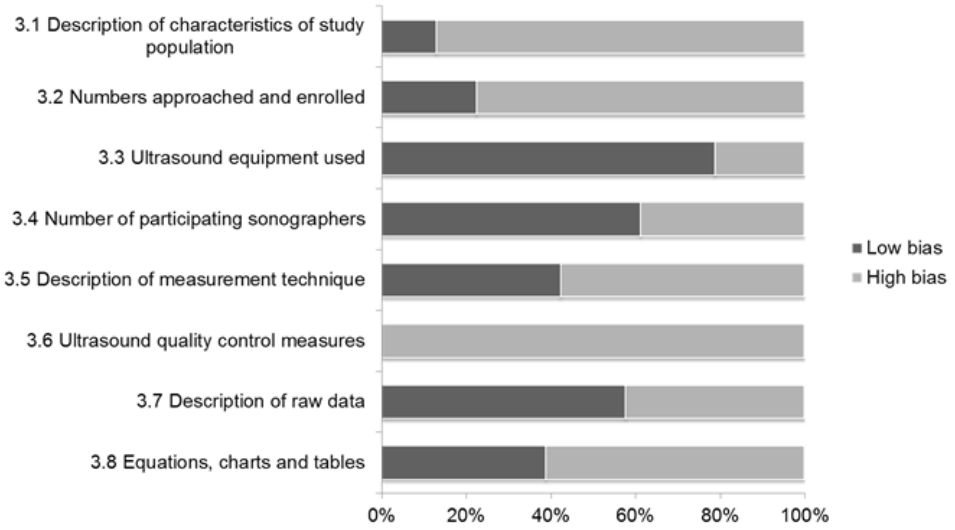


Figure 1.5: Aggregated methodological assessment of reporting quality of the included studies: Reporting methods (percentage of studies with a low or high risk of bias)

Of the 83 studies included in the review, 48%, had ultrasound examinations performed by multiple sonographers, whereas, 22%, of the studies used a single sonographer. Ultrasound quality control measures were reported very rarely, as shown in Figure 1.6. None of the studies reported the use of an image scoring method for quality assurance and only 5% of the studies standardised the participating sonographers in some way. Only 5% of the studies reported blinding their sonographers to the actual measurement recorded during the examination.

Study results were reported in the form of tables, equations, or charts, as demonstrated in Figure **??**. Although tables of median values (82%) and percentile ranges (78%) were common, only half of these tables contained mathematically predicted values following analysis. The remainder presented raw centiles. The mean or median was expressed mathematically with an equation in 60% of the studies. The standard deviation (SD) was mathematically expressed in 39% of the studies, either as a fixed number or as a function of gestation. Printed charts of the median and centile curves were seen in most of the publications.

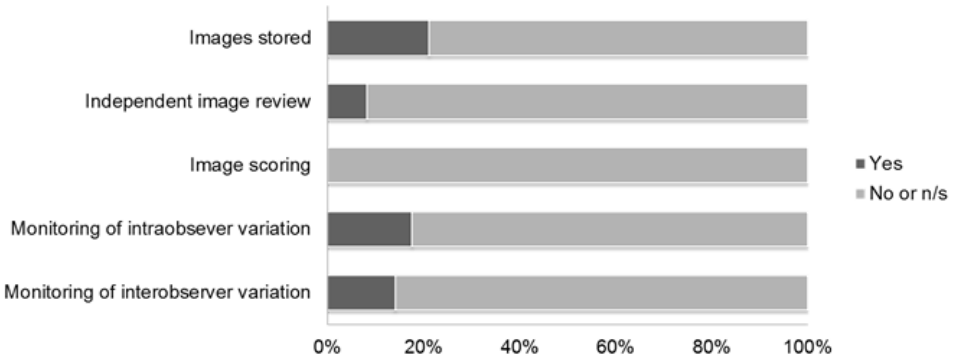


Figure 1.6: Ultrasound quality assurance measures used in the studies included in the fetal biometry reference charts systematic review

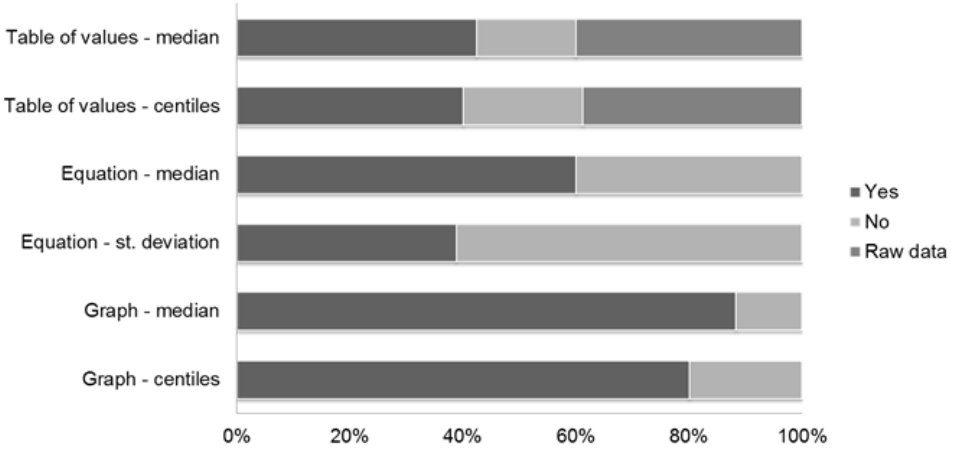


Figure 1.7: Use of presentation methods in the studies included in the fetal biometry reference charts systematic review

1.5 Review of newborn size reference charts

The third systematic review investigated newborn reference size charts. In 1963, Lubchenco *et al.* [59] used data on birthweight according to GA of infants born in a hospital in Colorado, USA to construct percentile charts for estimating intrauterine growth. They pointed out the importance of fetal growth and its relationship to both the immediate well-being and long-term outcome of the newborn. In a major step in our understanding of the relationship between neonatal size and birth outcomes, Battaglia and Lubchenco [60] related these percentile charts to the risk of neonatal mortality in 1967. Since then, many more charts have been published

based on cross-sectional measures at birth. These charts are used to assess the size of newborns and classify them as small, appropriate, or large for their GA. Classification is done using cut-off points derived from the reference charts.

Studies that generate newborn size reference charts use a wide range of study populations, methods for estimating GA, methods for obtaining anthropometric measurements, and methodological and statistical strategies [17, 61]. New charts continue to emerge [62] with continued controversies [63]. To the best of my knowledge, these studies had not been critically appraised in a systematic manner. In the third INTERGROWTH-21st systematic review, the methodological quality of studies that aimed to create neonatal anthropometric charts was reviewed to determine the features of these charts that can affect clinical decision-making and international comparisons.

One hundred and five studies were identified [16]. The studies were scored, using a set of 29 criteria, as having a low or high risk of bias based on the study design (17 items), statistical methods (8 items), and reporting methods (4 items) used. The majority (61%) of the studies were hospital-based, 39 (37%) were population-based, and the source of the data was unclear in the remaining 3 (3%) studies. The distribution of the included studies by region was 31% from Europe, 24% from Asia, 20% from North America, 14% from Latin America and the Caribbean, 7% from Africa, and 4% from Oceania. Most (95%) of the studies reported birthweight for GA, but neonatal length and head circumference (HC) were only reported in 40% and 37% of the studies, respectively. The minimum GA considered in the charts was 20–23 weeks in 30% of the studies, 24–27 weeks in 34% of the studies, and 28 weeks or higher in 36% of the studies.

Only 20 of the 105 studies (19%) had a total quality score above 50%, with only one above 75% [64]. Birthweight according to GA was compared between the six studies with the highest quality scores. For example, the largest difference

between these six charts at the 10^{th} centile was 343 g for girls and 352 g for boys at 35 weeks of gestation.

The medians (IQR, range) of the total quality scores for each of the three domains are summarised in Figure 1.8. The median quality score for study design was 30% (IQR 17.6, range 5.9%-76.5%), for statistical methods was 51% (IQR 25, range 0.0%-87.5%), and for results reporting was 52% (IQR 25, range 0%-100%). The three scores were summed to give a total quality score. The median total quality score for the 105 included studies was 39% (IQR 17.2, range 17.2%-79.3%). The

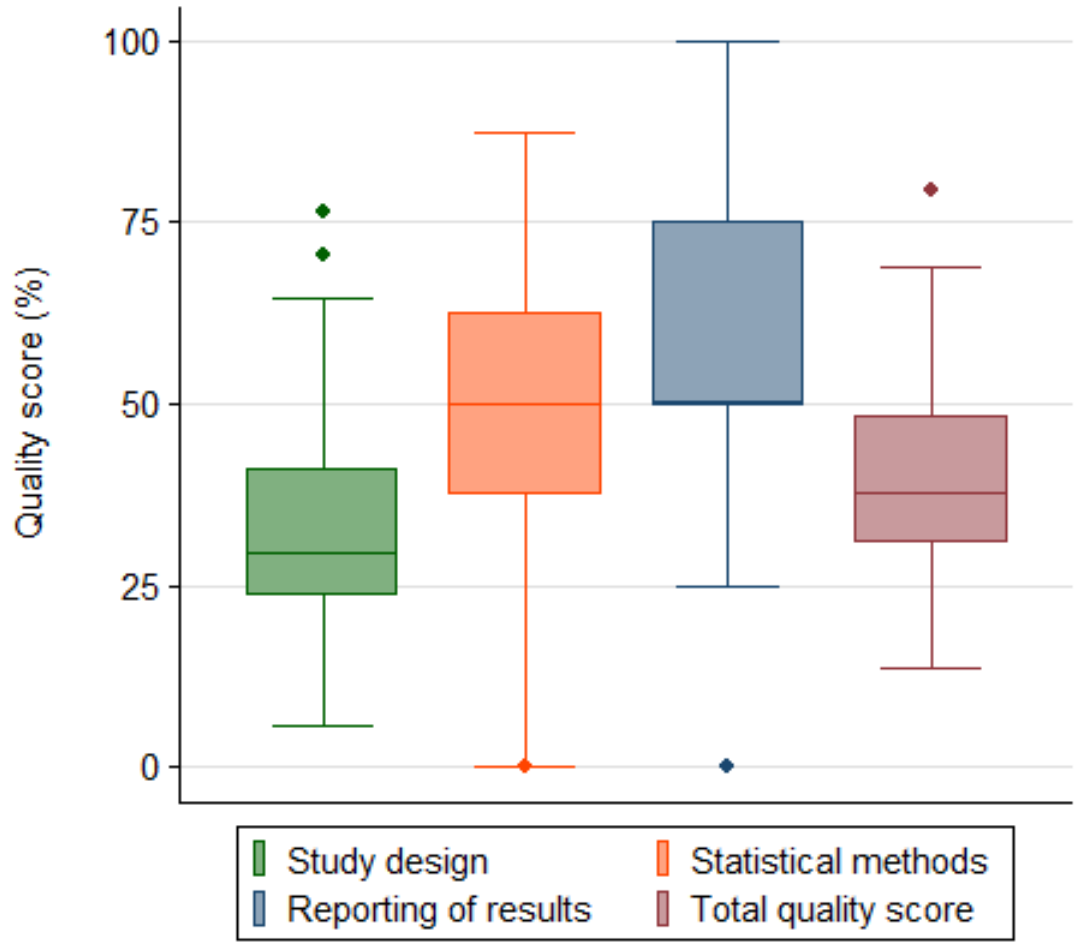


Figure 1.8: Median (interquartile range; range) of the quality score for each set of criteria studied: study design, statistical methods, reporting of results, and total quality score.

strategy used to estimate a reliable GA was reported in <30% of the studies. Only 21% of the studies used both LMP and ultrasound assessment to determine GA and only 27% of the studies excluded newborns with an unreliable GA.

The anthropometric measures (the main outcome measure) had severe methodological problems. The use of standardised instruments and the time at which the measures were taken after birth were described in <40% of the studies. Measurement techniques, protocols, and operator training were reported in only 17% of the studies. The use of at least two operators for performing measurements, equipment calibration, and operator standardisation were reported in <10% of studies. However, most of the studies reported the definition of their target population and the number of neonates at each GA.

Two-thirds of the studies used what are regarded as appropriate statistical models to create their reference charts. Smoothed centiles and separate charts for boys and girls were reported in about 80% of the studies. Seventy-one percent of the studies clearly stated whether GA was expressed as completed weeks, to the nearest week, or weeks and days. Seventy-six percent of the studies reported at least the 10^{th}, 50^{th}, and 90^{th} centiles, or parameters allowing their computation. However, only 18% of the studies presented the charts with z-scores or in a format that allowed these scores to be computed. The most common shortcomings were observed in items related to anthropometric evaluation, GA estimation, follow-up duration, the reporting of postnatal care and morbidities, the assessment of outliers, covariates, and chart presentation [16].

1.6 Discussion

The results of the three systematic reviews highlight the great variability in the many reference charts currently in use for dating pregnancy and comparing fetus and newborn size to the general population to identify those at risk. It has been

assumed that much of this variability is due to differences in populations in different geographical areas. However, the three systematic reviews summarised in this chapter demonstrated considerable heterogeneity in the design of the studies that create these charts. Consensus in methodology is essential to appraise population differences in fetal measurements.

Accurate GA estimation is a fundamental prerequisite for creating fetal and newborn size charts. There is robust evidence that ultrasound alone or corroborated with LMP to determine GA is more reliable than LMP alone [65]. It has also been argued that ultrasonography alone may be marginally superior to LMP confirmed by ultrasound [66], although such differences are small. Regardless of the combination of methods used, the fetal measurement used for dating pregnancy and the GA window during which it is applied should be clearly specified.

The issue of how best to select samples in research studies that aim to create reference equations of fetal size, especially those involving second and third trimester biometry remain unresolved. Some authors have proposed using as unselected as possible a sample to best represent the underlying population [67]. However, a number of pathological conditions (environmental, medical, or obstetric) may be prevalent in such a sample, which is likely to affect the reference equations derived. Pathological processes such as smoking [68], hypertension and pre-eclampsia [69], maternal disease, abnormal karyotype and congenital anomalies [70], preterm delivery [71], and stillbirth [72] are well known to affect fetal size later in pregnancy. There is also now evidence to suggest that early fetal growth restriction can be evident as early as the first trimester [73].

It is reasonable to choose a reference equation for use in a clinical service from a publication with the lowest risk of methodological bias. The INTERGROWTH-21st systematic review of pregnancy dating only identified 4 studies that satisfied more than 18 of the 29 defined quality criteria. Figure 1.2 shows that using

any of these four charts leads to very small differences in GA estimation, when compared with the remaining charts.

The availability of multiple references, mostly of poor quality, to evaluate a simple but clinically important measure, namely size at birth, is unusual in medicine. This underlying conceptual issue is the result of the widely held belief that fetal or newborn growth differs across regions, ethnic groups, and socioeconomic levels, so requires population-specific charts for its evaluation. Fetuses and newborns are currently judged depending on the region or even health institution where they are evaluated, rather than on their clinical needs. Comparisons across populations are almost impossible as cut-off points differ between studies. This concept has been challenged by evidence demonstrating similarity in the genetic make-up of different non-isolated populations worldwide [74, 75, 76] and by recent comparisons between populations of early and late fetal [12], infant, and child growth [19].

The three systematic reviews were based on a similar approach and criteria for evaluating methodological quality. The reviews revealed that the centile values of charts designed in different studies differ a great deal from one another. The charts were based on different populations and were created with different sample selection, methodology, and statistical modelling methods [16, 17]. Anthropometric measures for such studies should be evaluated using standardised instruments, measurement protocols, and trained and standardised staff [77]. However, most of the studies in the three reviews did not score well on these items. Many of the studies obtained routinely collected measurements without using calibrated equipment or simply collected data from medical records or birth certificates. Some of the studies included in the reviews may have simply failed to report on some of the aspects and items used in the review to evaluate and judge methodological quality. Therefore, these results should be interpreted with caution. Some of the items considered in the checklist are more important than others. For example,

failure to provide a statement about sample size considerations that were undertaken when designing a study may not necessarily lead to bias, whereas sample selection is crucial and is highly likely to result in bias if not considered carefully.

How a study is reported and presented is a key element when evaluating its quality and the robustness of its findings. The inclusion of the quality criterion of reporting the number of fetuses or neonates at each GA indicates the sample size distribution across GA's and the quality of the data collected. An adequate sample size at each GA is essential for precise estimates, especially at extreme centiles where variability is greatest, such as the 3^{rd} and 97^{th} centiles. The application of these reference charts, which are already in clinical use, directly affects the identification of at-risk fetuses and newborns in possible need of treatment and nutritional strategies.

The criteria used to measure methodological quality in the three systematic reviews form a methodological quality assessment checklist. This checklist is not intended to commend or discard studies. It should be used as a consensus guideline to improve consistency in reporting, help design new studies, and as a guide for evaluating similar studies for future research in early pregnancy biometry and human growth studies. This framework provides a sensible way to compare the methodological rigour of studies and an objective, quantitative assessment of study methodology. It can be used to improve consistency in fetal growth research and highlight limitations that should be avoided in future research.

2

Design and methodological considerations for the construction of human fetal and neonatal size and growth charts

2.1 Introduction

A size or growth chart is a plot of a series of centile curves that illustrate the distribution of a selected body measurement in relation to age. Such charts allow an individual to be placed in the context of like individuals. Charts of measurements are useful for assessing humans at all stages: fetuses, neonates, children, and adults. This thesis focuses on fetal growth up to birth. Fetal growth charts are primarily used to compare the size of a fetus with reference data when GA is known at a specified time [24], to estimate GA from fetal size (CRL, biparietal diameter (BPD), and HC are commonly used) [43, 21, 78], and to assess a fetus's rate of growth between two time points (velocity charts) [79, 80].

The most famous record of human growth is of the height of one boy, measured nearly every 6 months from birth to 18 years. It was made during the years 1759–1777 by Count Philibert Gueneau de Montbeillard using his son and later published by Buffon in a supplement to the Histoire Naturelle [30]. Francis Galton (1822-1911) was the first to demonstrate that the Laplace-Gauss distribution, or the 'normal distribution', could be applied to human growth, for example, growth in height from birth to adulthood [81]. From this finding, he coined the use of percentile scores for comparing measurements with the normal distribution [82]. A first application of this approach was in growth in height, which is normally distributed from birth to adulthood. Henry Pickering Bowditch (1840-1911) was the first to apply this concept by constructing percentile charts for growth in height and weight of Boston children. He described the method as 'Galton's percentile grades' [83].

Growth charts are intended to aid in making clinical judgements. They are used as a screening tool in the identification or classification of newborn size as small, appropriate or large for a specified GA at birth, based on cut-off points on a specified reference chart [60, 59]. Many research applications use charts to describe the average pattern. However, reference charts are used to identify subjects whose measurements fall in extreme centiles (for example below the 3^{rd}, 5^{rd}, or 10^{rd} centiles or above the 90^{th}, 95^{th}, or 97^{th} centiles), as they are the basis on which most clinical decisions are made. These extreme centiles are likely to be indicative of growth restriction or other clinical complications affecting growth. Most studies developing reference growth centiles are based on cross-sectional data collected at one time point during pregnancy. A limiting aspect of using cross-sectional data is that growth trajectories are not monitored over time.

In most fields of medicine that require the identification of what is regarded as 'normal', internationally accepted classifications, cut-offs, or standards are applied. These global cut-offs are chosen as a result of evidence of adverse outcomes in

relation to the target measurement, such as in the definition of hypertension, anaemia, and diabetes. However, no such standard values apply to reference charts, as illustrated by the INTERGROWTH-21st systematic reviews of published charts of fetal biometry, pregnancy dating, and neonatal size [16, 28, 27, 84]. There is certainly no explicit link between cut-offs and the risk of bad outcomes, perhaps because physical size varies greatly across populations and such outcomes are rare. The systematic reviews revealed wide variations between the centile values reported by the included studies. These variations were a result of considerable methodological heterogeneity: the charts were based on different populations and created with different sample selection, methodology, and statistical modelling methods [16, 17].

For example, the INTERGROWTH-21st review of pregnancy dating charts identified 29 studies with the main aim of constructing charts to predict GA from CRL [27]. The studies' inclusion criteria, exclusion criteria, and maternal demographic characteristics generated the highest potential for bias. None of the studies had a systematic method for checking the quality of their ultrasound measurements. The four studies with the lowest risk of bias had the smallest variations in their GA estimations. This result was particularly evident when comparing extreme centiles such as the 3^{rd}, 5^{th}, 95^{th}, and 97^{th} centile distributions for estimated GA values, which are usually the focus for clinical decisions [27].

The INTERGROWTH-21st review of the methodology used in ultrasound studies aiming to create charts of fetal size identified 83 studies. Again, the highest potential for bias was found in the inclusion and exclusion criteria, ultrasound quality control measures, and sample size calculations. Only six of the studies indicated their sample size calculations [28]. INTERGROWTH-21st review of newborn charts included 105 studies. Shortcomings were most often observed in items on anthropometric evaluation, GA estimation, follow-up duration, the reporting of postnatal care and morbidities, the assessment of outliers, covariates, and chart presentation [16].

Many of these charts are in clinical use today and directly affect the identification of at-risk newborns that require treatment and nutritional strategies.

Reference charts are globally perhaps the most widely used charts, yet few methodologists and statisticians are working to improve the methodology in this area. The shortcomings of existing studies, as revealed by the results of the INTERGROWTH-21^{st} systematic reviews, are the motivation for discussing key issues in study design that underpin how to plan a quality study. I draw heavily on our experiences from the INTERGROWTH-21^{st} Project, a multicentre, multi-ethnic, population-based study conducted in eight geographic areas in Brazil, China, India, Italy, Kenya, Oman, the UK, and the USA [11]. The primary aims of the project were to produce international standards for pregnancy dating [21], fetal growth [24], and newborn size for GA [25], and to follow the postnatal growth of preterm infants [26].

The key methodological considerations for the good design of growth studies include, but are not limited to, the use of an appropriate study design that provides a robust answer to the intended question (for example, choosing between a longitudinal or cross-sectional design), distinguishing between size and growth, inclusion criteria, sample size, how pregnancy is dated, measurement procedures (for example, whether or not to take replicate measurements), quality control, and whether and how data from multiple geographical sites is combined if multiple recruitment centres are used. Each of these issues will be discussed in the following chapters. First I will clarify the important distinction between charts constructed using the descriptive and prescriptive approaches.

2.1.1 Descriptive versus prescriptive approaches

Growth charts are constructed using either a descriptive or prescriptive approach. A descriptive approach results in the construction of what is commonly referred to as **reference centiles or reference charts** whereas a prescriptive approach is used to

construct **standard centiles or standard charts**. The term *prescriptive approach* is used in the scientific literature to describe the process of producing biological norms or a desirable target to be achieved or aspired to at individual and population levels. Prescriptive standards show how growth should be independent of time and place [17]. For human growth, they are usually based on selected populations considered to be of optimal health, for example with adequate nutritional status and at low risk of abnormal growth. In contrast, the *descriptive approach* is commonly used to produce a reference chart that describes the anthropometry of a given population at a particular time and place, such as a hospital, region, or country. Descriptive reference charts are usually based on an unselected group of women with minimal exclusion criteria on risk factors for optimal health. Although they are used more widely, descriptive charts are only relevant to the source population. Different populations will differ in many aspects, such as rates of smoking during pregnancy, malaria, gestational diabetes, which can all affect optimal growth. In principle, following the descriptive approach requires separate reference charts for each population of interest.

The INTERGROWTH-21st review of 105 studies aimed at creating newborn size reference charts revealed that authors only stated whether their aim was to construct a prescriptive standard or a descriptive reference in half of the papers included in the review. This is problematic because the two chart types have different sensitivities and specificities for detecting growth disturbances. However, even when the studies did state that their aim was to create a standard, half did not actually produce one because their study population was incorrectly chosen. For example, two studies that aimed to create prescriptive standards did not exclude pregnancies with risk factors or conditions known to affect fetal growth or neonatal size [17, 85].

Studies that aim to develop standards use known risk factors for suboptimal growth as exclusion criteria. For example, the INTERGROWTH-21st Project had

several exclusion criteria related to obstetric history, gynaecological factors, socio-demographic features (e.g., age, BMI, and smoking), clinical factors (e.g., blood pressure and sexually transmitted infections), and current pregnancy (e.g., whether the pregnancy was a singleton with accurate pregnancy dating) [11]. A further consideration in developing a standard is an assessment of the data quality and the representativeness of the study population. A prescriptive approach is necessary for international relevance, as it results in charts that are independent of place and time.

Truely prescriptive charts are rare, perhaps because of the distinct differences observed in different populations, leading to the popular belief in differences in human growth *in utero*. Whilst these observable differences are generally due to epigenetics, the little evidence available has shown that only small variations in growth are explained by genetics [75]. Until recently, it was generally accepted that observed differences in preterm, fetal and neonatal growth were largely due to biological differences between different regions and ethnicities, resulting in a need for population-specific charts. This concept has recently been challenged by evidence demonstrating similarities in the genetic make-up of different non-isolated populations worldwide [74, 76], and more specifically by recent comparisons finding similarities in linear early and late fetal growth, newborn size at birth [12], and linear child growth [19] in diverse populations. However, the concept of similarity in growth is not new. In 1974, Habicht sought to understand the effect of ethnic differences on achieving growth potential. He compared weight and height data from preschool children of different ethnic backgrounds that were presumably well nourished. He found relatively small differences of 3% for height and about 6% in birthweight between preschool children of different ethnicities but comparable socioeconomic status and nutrition. In contrast, larger differences of 12% in height and 30% in weight were observed between these children and those, often of similar ethnic and geographical backgrounds, who lived in poor urban and rural regions [86].

There is no universal agreement on what constitutes optimal growth, or how it should be monitored [17]. The WHO recommends that an international human growth standard chart be based on longitudinal studies of selected populations with a low prevalence of maternal and fetal complications, where fetal or anthropometric measures are collected prospectively [77]. In 2006, the WHO adopted a prescriptive approach in its MGRS to produce international child growth standards from birth to 5 years [18]. These standards are now widely used in over 140 countries [87]. They were constructed using data from children born in six different regions of the world, born to mothers considered to be 'healthy', and living in environments believed to support what WHO researchers view as the optimal growth of children. The children were fed according to accepted international nutritional standards (including breast feeding), and their mothers were adequately nourished and avoided known adverse factors such as tobacco exposure. The study demonstrated that children born in different regions of the world can and should grow equally well. It showed that sex and ethnic origin are minor determinants of growth compared with adequate nutrition, environment, and health [18]. A similar approach was used in the INTERGROWTH-21st Project for the development of preterm, fetal and newborn standards [24, 25].

2.2 Study design

The study design is of fundamental importance for any research study. An appropriate study design is dependent on the question or hypothesis that the study should answer. For example, as ultrasound scans are expensive, having multiple scans may not be cost-effective. The timing of these measurements is therefore important. Establishing the optimal design based on the number of scans and timing of the scans that maximises the amount of information that can be deduced would be helpful for planning and conduct of fetal growth studies. The

choice of an optimal design is a trade off between the amount of extra information provided by each extra scan in relation to cost implications, time, and labour. The establishment of an optimal design is an area that requires further research.

There are many design challenges for studies that aim to construct growth charts from fetal and neonatal measurements. The INTERGROWTH-21st systematic review of the methodology in observational studies of fetal size found that the majority of the included studies had a cross-sectional design rather than a longitudinal design. In the review of fetal size charts, 50% of the studies were found to have a high risk of bias based on the study design [28]. As an example, in the review, a cross-sectional design was considered to be at low risk of bias if it was clearly specified that only one examination per fetus was included in the analysis. Similarly, a longitudinal design was considered to be at low risk of bias if ultrasound data were analysed using a method that took into account their serial nature.

2.3 Cross-sectional, longitudinal, or mixed designs

The way measurements are collected should be informed by the question being addressed. For example, size at a specific time (such as birth) can be obtained using cross-sectional data, but velocity requires longitudinal data. Most studies that aim to develop reference growth centiles are based on cross-sectional data collected at one time point during pregnancy. A limiting aspect of using cross-sectional data is that growth trajectories are not monitored over time.

Assessments of size over time are the most common type of analysis for fetal measurements and can be done using either a cross-sectional or longitudinal design. The nature of the data should inform the analysis methodology used. Analysis of longitudinal data should address correlated measurements from the same individual and, if relevant, the use of repeated measurements taken on a single visit. For example, in the FGLS component of the INTERGROWTH-21st

Project, ultrasound measurements of fetal size were taken in triplicate at each visit to minimise measurement error. This longitudinal study collected data in a three-level hierarchy: measurements within visits within participants. The data analysis should therefore be tailored to the hierarchical structure of the data and account for the correlations of the measurements within a subject and the variability within and between individuals at a given site. A multi-level regression analysis can deal appropriately with such data structures.

Multicentre studies can introduce additional complexities by using a mixed design, in which some participants are studied longitudinally and others cross-sectionally. A mixed design can be useful for studying growth intensively in periods of rapid growth using a longitudinal design and studying growth less intensively in periods of slow growth using a cross-sectional design. This may be an efficient, cost-effective approach. A good example is the WHO-MGRS, which combined a longitudinal study design from birth to 24 months with a cross-sectional study of children aged 18 to 71 months [88]. A mixed design is also likely to arise when routine data collected from individuals requiring close monitoring who are seen more than once. The use of routine clinical measurements is not recommended, however, because of the high likelihood of inaccurate measurements and potential for bias compared with data collected specifically for research purposes.

2.4 Size and growth

In principle, size relates to measurements at a specific time, whereas growth relates to a change in size over time. In practice, the term growth is widely used for both types of data and is thus sometimes used inappropriately [89, 90, 91]. Centile charts showing fetal size at different GA depict the average attained size. Charts derived from a single measurement of each fetus or neonate at a specified time point depict size. These size charts should not be confused with the dynamic

process of growth. True growth charts are derived from a series of anthropometric measurements made of each fetus or neonate at multiple time points [73, 67, 92, 93]. Strictly speaking, only charts derived from longitudinal studies that incorporate more than one measurement per individual should be called growth charts. However, longitudinal studies can be used to produce both size and growth charts. Centiles derived from cross-sectional or longitudinal data will tend to closely agree with one another as they only differ in analyses that account for the non-independence of the longitudinally obtained observations. Size and growth refer to different information, are used in different clinical applications, and have different interpretations.

2.5 Who to include

The choice of an appropriate sample and target population is of great importance for making comparisons and for inferring to the general population. The target population is the population to which the chart will apply. It is defined by the study's inclusion criteria, for example, geographical area, ethnic group, and parity. Charts produced from a subgroup of a population, such as women who are obese, would be inappropriate for making inferences about the general population as they are not representative and lack external validity. The INTERGROWTH-21st systematic review of the methodology used in ultrasound studies aimed at creating charts of fetal size showed that about half of the 83 included studies had a prospective design, in which ultrasound examinations were performed for research purposes only [28]. However, it is now common in clinical practise to routinely collect ultrasound information in computerised databases. Reference centiles for fetal size, for example, are often constructed from routinely collected data. Although retrospective analysis of such databases is a practical solution for generating a large sample size, the resulting sample will be mainly descriptive and cannot be used to assess or represent optimal growth. The alternative is a prospective study, in which the participant

recruitment and collection of clinical, demographic, and ultrasound data is purposely and solely carried out with the objective of creating size charts and not as part of routine care provision. Data that are collected prospectively, and specifically for the purpose of developing reference centiles, are therefore recommended [67, 92].

2.6 Sample size

Sample size affects centile precision, so must be considered when planning studies intending to develop such instruments to ensure adequate coverage of the population and any planned subgroup analysis [88, 94]. There is very limited literature on what to consider when determining the sample size of fetal growth studies [95, 96, 97]. In 1995, a WHO Expert Committee recommended a rule of thumb for growth studies of, that a sample of at least 200 individuals overall or, if relevant, for each subgroup for which separate charts will be produced [77]. Calculating sample size is not straightforward as it depends on factors such as the study design (longitudinal, cross-sectional, or mixed), number of repeated measurements per individual, existence of replicate measurements, and practicality (cost, time, and manpower) [94]. These factors must be considered when choosing the sample size without compromising the power of the study. For example, it may be necessary to take into account the maximum number of women who can be given an ultrasound scan each day at a particular facility. The INTERGROWTH-21st systematic review of the methodology used in published ultrasound studies for developing size or pregnancy dating charts found that only 6 of 83 published ultrasound growth or size charts included their sample size calculations in their description of their methodology [28, 27].

Sample size calculations can be based on either parametric or nonparametric methods. Nonparametric methods do not make any distributional assumptions and can be implemented using simulation and bootstrap techniques [97, 98, 99]. Regression-based methods for sample size can also be evaluated by either non-

parametric or parametric approaches, depending on the distribution of the covariate [100, 101]. Methods based on regression-based limits are commonly used in clinical chemistry studies involving normal reference ranges [102]. The same methods can be applied in fetal and neonatal growth studies [103].

Sample size calculations for growth charts based on longitudinal data are complex [95, 104]. The standard errors of centiles are overestimated in longitudinal studies as they ignore the existence of a series of measurements from each fetus [104, 96, 105]. In general, longitudinal studies are preferable as they are more efficient and have greater power than cross-sectional studies. Royston (1995) defined this efficiency as the design factor, D, which is the number of fetuses in a cross-sectional study that would give the same precision as one fetus in a longitudinal study. He used a simulation study of ultrasound-based BPD and compared the variance of a centile in longitudinal and 'equivalent' cross-sectional designs. He calculated the design factor (effect) to be $\sim$2.3 [104]. A longitudinal study requires approximately half the sample size of a cross-sectional study to estimate a given centile with the same precision.

The accuracy of estimated centiles is inherently variable. Extreme centiles exhibit large imprecision because there are few observations at the extreme ends of the distribution, while the median has the greatest precision. In this thesis, I will demonstrate two approaches (precision and accuracy of a single centile and regression based methods for sample size calculation) for calculating sample size for creating reference centiles for normally distributed data, in this case, CRL. Similar approaches can be applied to non-normal data, such as birthweight after transformation. Other approaches also exist, for example simulation and bootstrapping, as was been demonstrated by Harris *et.al.* [98], Linnet [99], and Jennen-Steinmetz [97].

2.7 Precision and accuracy of a single centile

Data that are conditionally normally distributed, for example fetal size measurements tend to be close to a normal distribution at each GA and thus can be summarised using the mean, μ, and SD, σ [105]. Any required centile can be estimated from the mean and SD as $\mu + z_\alpha \sigma$, where z_α is the normal equivalent deviate (z score) corresponding to that centile. For normally distributed unreplicated data, such as CRL, the standard error of the p^{th} centile is obtained from the standard formula for the variance of a centile of normal distribution:

$$SE_p = SD\sqrt{\frac{1 + \frac{1}{2}z_p^2}{n}} \text{ [106]} \tag{2.1}$$

where SE is the standard error, SD is the standard deviation of the measurement (which will increase with GA), z_p is the value of the standard normal distribution corresponding to the p^{th} centile, and n is the sample size. For example, for the 2.5^{th} or 97.5^{th} centile, $z_p = \pm 1.96$, giving SE = 0.08 SD for a sample size of 500 and 0.03 SD for a sample of 4,000.

2.8 Regression-based reference limits

A reference range (also known as a reference interval or the normal range) is the range of values obtainable for a physiological measurement. It forms a basis for comparison or a frame of reference for a physician or other health professional to interpret a set of test results for a particular patient. References can be obtained from the general population and therefore represent on average what is expected in a specified population. In some cases, reference ranges are obtained from populations deemed to be of optimal health and thus represent the expected norm in the absence

of infection or disease. For example, pregnancy reference ranges for hormone levels, blood tests, and urine tests are formulated using healthy subjects.

If we take a range of observations from a population (the difference between the two most extreme values) and continue to sample the population, we will continue to find observations outside that range, and the range will continue to grow. A reference interval is therefore used to refer to a range between two quantiles. The reference interval usually used is the normal range and runs from the 2.5^{th} centile (*the lower reference limit*) to the 97.5^{th} centile (*the upper reference limit*). It is also called the 95% reference range or 95% reference interval. This normal range excludes 5% and includes 95% of measurements from apparently healthy individuals. Values outside a reference range are not necessarily pathologic and are not necessarily abnormal in any sense other than statistically. Confusion here can arise between the use of 'normal' in medicine and of 'normal distribution' in statistics. Nonetheless, values outside the reference range are indicators of probable pathology [107].

Regression-based reference ranges for estimating sample size were first proposed by Royston [96] and were extended by Bellera and Hanley [95]. Regression analysis can be used to obtain reference limits that account for factors such as age, gender, and parity with corresponding confidence intervals (CI) [108, 109, 110, 111]. There are currently no recommended strategies for estimating sample size when constructing regression-based reference ranges and CI. In clinical chemistry, analytical variability is usually accounted for when developing and establishing references [112]. Analytical variability refers to factors likely to influence the experimental design, which includes the laboratory, day the test was taken, analyst, instrument, etc. In 1987, Linnet [102] proposed that the analytical variation due to measurement error should be less than the biological variation and that the sample size should be large enough to make the width of the 90% CI for the reference limit smaller than the width of the 95% reference range. The number

of individuals required to achieve a particular ratio between the two widths can be calculated, assuming a normal distribution. For example, define R as the ratio of the width of the $100(1-\alpha)\%$ CI for the reference limit to the width of the $100(1-\beta)\%$ reference range. When R = 0.1, the estimated sample size is 206, as demonstrated below. This approach indicates how large a sample size is needed to estimate centiles with adequate precision [98].

Example: Let the ratio between the two widths, R, be 0.1, z_p the standard normal deviate for the p^{th} centile, and α and β the type I and II errors, respectively.

1. The 95% reference range is obtained by the relationship $\mu + z_\alpha\sigma$, where z = 1.96 and μ and σ are estimated from the data. The width is 3.92σ.

2. The sampling SD of $\mu + z_\alpha\sigma$ is estimated:

$$\sigma\sqrt{\frac{1+\frac{1}{2}z_p^2}{n}}. \tag{2.2}$$

3. Considering the 95% reference range and substituting z = 1.96 results in

$$2.921 \times \left(\frac{\sigma^2}{n}\right). \tag{2.3}$$

4. The width of the 90% CI for the 95% reference range is thus

$$(2 \times 1.645) \times \left(1.709\frac{\sigma}{\sqrt{n}}\right) = 5.623\left(\frac{\sigma}{\sqrt{n}}\right). \tag{2.4}$$

5. The ratio between the two widths is

$$R = \frac{5.623\left(\frac{\sigma}{\sqrt{n}}\right)}{3.92\sigma} = \frac{1.434}{\sqrt{n}}. \tag{2.5}$$

6. For R = 0.1,

$$n = \left(\frac{1.434}{0.1}\right)^2, \tag{2.6}$$

resulting in a sample size of 206. Bellera and Hanley [95] extended Linnet's approach to accommodate different sampling strategies. They assumed uniform or Gaussian GA distribution and calculated the sample size required to achieve a given degree of precision. They derived sample size formulas for estimating the 95% reference limit with a 95% CI for R = 0.1 (relative margin of error), as proposed by Linnet. For example, we can calculate a 95% $CI(z_{(1-\alpha/2)} = 1.96)$ for the 95^{th} centile of the CRL reference limit ($z_p^2 = 1.645$) with a relative margin of error of $\sim$ 10% (R = 0.1). Assuming the range of gestation is approximately 4SD, the minimum required sample size is given by:

$$n \geq z_{1-\alpha/2}^2 \left(5 + \frac{\frac{1}{2}z_p^2}{z_{1-\beta/2}^2} \times R^2\right) \tag{2.7}$$

For normally distributed data, sample size estimates for the mean or median can be estimated using the simplified formula:

$$n \geq z_{1-\alpha/2}^2 \left(1 + \frac{\frac{1}{2}z_p^2}{z_{1-\beta/2}^2} \times R^2\right) \text{ [95, 96]}. \tag{2.8}$$

The estimated sample size is 4,288 for adequate precision at the extreme centiles where variability is greatest (worst case scenario) and 5,088 at the median (best case scenario). For the same power of a study and effect size, the total sample size required at the median will be higher than that required at the extreme centiles because of the differences in variability. Variability is much less at the median than at the extremes and therefore a larger sample size will be required to detect the

same effect size (Table 2.1).

Margin of error (R)	Worst case scenario (range of gestation $\cong 4SD$)	Best case scenario (Median gestational age)
10%	4,288	5,088
15%	1,904	2,264
20%	1,064	1,272

Table 2.1: Estimated minimum sample size required using the specified margin of error (ratio of the width of a 90% CI for a reference limit to the width of the 95% reference range)

The primary considerations of sample size in the INTERGROWTH-21st Project were based on requiring (a) the sample size to be large enough to yield precise estimates of extreme centiles (e.g., the 3rd and 97th centiles), which requires a definition of 'precise', for which there is no standard approach; (b) sufficient power to explore ethnic-specific (i.e., site-specific) growth in the FGLS component, in the event that ethnic differences did emerge from the data in the main growth indicators; and (c) the FGLS component to yield an adequate number of newborns for inclusion in the Preterm Postnatal Follow-up Study (PPFS). Although statistical considerations were important, certain logistical issues were also critical. For example, one key consideration was the number of women who could be scanned at a centre in a week. This practical issue was relevant as a bespoke ultrasound machine model produced for the project was used at every centre, and each centre was provided with one machine.

The target total sample size of 4,000 from all of the eight sites combined was larger than most previous studies. This sample size was adequate for producing reliable curves and exploring variability between countries. It was estimated that fewer than 5% of the women would be lost to follow-up and that about 10% of women would be excluded from the development of the fetal growth standards due to developing pregnancy complications severe enough to affect fetal growth, as identified in the protocol [79]. Considering the sample size for the FGLS component in relation to

the precision and accuracy of a single centile as first proposed by Royston [113] results in a precision of 0.08 SD and 0.03 SD at the 2.5^{th} or 97.5^{th} centile for a sample size of 500 and 4,000 fetuses, respectively. The sample size calculation in the INTERGROWTH-21^{st} Project was based on a cross-sectional design. Royston suggested that a longitudinal design has equivalent precision to a cross-sectional design with a sample size $\backsim$ 2.3 times larger than it. Based on that value, the longitudinal component of the FGLS with 4,000 fetuses would have equivalent precision to a cross-sectional study of over 9,000 fetuses [22]. Therefore, the sample size for FGLS was more than sufficient for constructing of fetal charts with great precision, even for extreme centiles such as the 3^{rd} and 97^{th} centiles.

2.9 Quality control

Meticulous standardisation, continuous surveillance, and ongoing monitoring of adherence to measurement protocols during data collection are essential to ensure consistency and to minimise systematic error. Such efforts can lead to early detection and pre-warning, and can flag deteriorating standards [114, 115, 116, 117, 118]. Quality control has recently been shown to be an important process in monitoring performance and competence in areas such as medicine [119, 120]. One of the most important aspects of any research involving measurements is the quality of the data. Great effort is often expended on the design of a study, but little thought may be given to how the method of data collection will affect the quality of the measurements. For example, many factors can introduce variability in ultrasound data, such as multiple sonographers and scan machines, a lack of standardised procedures when data are collected by more than one person, a lack of specific training, and failure to monitor ultrasound image quality.

It has previously been argued that reference studies should be performed by a single operator to improve the repeatability of the data by avoiding inter-observer error.

However, as ultrasound scans in most clinical services are performed by multiple operators, inter-observer variability is inevitable and should not be ignored. It is reasonable for reference studies to take inter-observer variability into account when using multiple operators and to take quality assurance steps to improve the quality and consistency of measurements. Such steps include saving and independently reviewing scan images, and measuring intra- and inter-observer variability. A formal exercise to standardise the contributing ultrasonographers should be conducted, as this improves the reproducibility of fetal biometry [121].

The INTERGROWTH-21st systematic review of the methodology used in ultrasound studies aimed at creating charts of fetal size revealed that $\sim$ 20% of studies did not report which ultrasound equipment was used, fewer than half described the measurement techniques used, none of the studies reported an ultrasound quality control measure, and >90% did not incorporate blinding when taking measurements [28].

The precision with which measurements are taken depends on the equipment used. In ultrasound studies, for example, advances in technology have improved the magnification of ultrasound machines, leading to better measurements. In anthropometry, weighing equipment must have an acceptable precision level. For example, the MGRS and INTERGROWTH-21st studies used SECA 376 scales that were precise to 5 g for weight measurements <7.5 kg and were precise to 10 g for measurements between 7.5 kg and 20 kg. The scales were calibrated at least twice every week.

Precision can be evaluated by assessing the intra- and inter-observer variation of the measurements collected. Measurement accuracy can be enhanced by an explicit measurement protocol, training and standardisation of the staff involved in taking measurements to avoid mistakes due to repetitiveness, assessment of image quality in the case of ultrasound images, standardisation of study protocols, checking for digit

preference, and ongoing monitoring of adherence to the measurement protocol during data collection. In the MGRS, checks were carried out when measuring skinfolds using a skinfold calliper that read to 0.2 mm units to ensure that measurements did not end in odd decimal values (e.g., 0.1 mm). An analysis of digit preference for one site with nine observers found that one observer tended to overestimate measurements, shown by a disproportionate frequency of the digit 0 (8.4%) versus the digit 2 (34.4%) [88]. Monitoring these elements and rectifying problems helps to ensure consistency and minimise systematic error. Numerical methods of monitoring measurement processes include Bland-Altman plots [122], technical errors of measurement [123], and cumulative summation charts [124, 125, 126].

In the FGLS component of the INTERGROWTH-21st Project, serial fetal growth scans were conducted every 5 $\pm$ 1 weeks from recruitment at 9^{+0}˘13^{+6} weeks of gestation until, but not beyond, 42^{+0} weeks of gestation. The health institutions participating in the project were diverse and used different pathways and protocols for scanning pregnant women in their routine clinical practice. For the data collected to be comparable within and between the study sites, all ultrasound measurements had to be performed in a standardised manner. Strict ultrasound fetal anthropometric measurement protocols were used so that data of the highest quality would be obtained from all of the centres. The data collected at each site could therefore be compared and potentially combined into a single dataset to generate growth standards.

Standardisation and quality control of measurements were necessary to ensure that the ultrasonographers measured all of the fetal biometric dimensions in an identical fashion. To oversee this process, the project had an Ultrasound Quality Unit (USQU) team, whose mandate was (a) development of standard operating procedures, (b) initial standardisation (involving training, assessment, and certification) of the ultrasonographers, (c) site specific standardisation exercises,

(d) quality control monitoring of routine replicate measurements by re-measurement and quality assessment of a random 10% sample of images, (e) analysis and reporting of ultrasound data quality, and (f) identification of retraining needs. The USQU was also responsible for ensuring adherence to the protocol and ongoing quality control assessment. This included a pilot reproducibility study, site visits, quality assessment of ultrasound images, assessment of collected data, and evaluation and repetition of ultrasound measurements.

The USQU team met every month to produce data quality statistics, which included plots of intraobserver reliability of individual ultrasonographers, cumulative summation charts of intra-observer reliability, and site comparisons, including comparing site-specific bias with the rest of the study. Areas of potential concern were highlighted and, if appropriate, site visits and retraining were arranged [127, 128]. These components were incorporated due to their importance in ensuring data quality. A review by Biau *et al.* showed the wide applicability of quality control methods in specialties such as surgery, endoscopy, and anaesthesia [129] using a learning curve analysis to test quality [126, 130, 131]. A detailed data quality control process that ensures good quality data is the cornerstone of any high-quality study.

2.10 Routinely collected data versus research data

Medical records have been used as a source of data for research since 1917, when Codman began using patient information cards to track long-term health outcomes [132]. Great advancements in technology have recently contributed to a shift from paper based records to electronic medical records. Sweden and Denmark are among the earliest examples of nations that have made the transition to electronic record-keeping systems that save time, money, and lives. Such comprehensive databases are goldmines of clinical information and have the potential to improve clinical practice, permit real-time learning, and create a large evidence base for clinical care [133, 134].

However, concerns have been raised about the utility of patient records for research purposes. Medical records are intended for patient care. Data is not recorded in a systematic manner, as would be the case with research [135]. The reliability of retrospectively collected data is thus a concern. Other difficulties include illegibility, incomplete records, and a lack of standardised documentation for ease of classification and comparability between similar settings [136, 137, 138]. If not well addressed, these concerns can bias study results and limit external validity. Put simply, routine data are not fit for purpose if the aim is to develop standards.

Recall bias remains a source of error in both prospective and retrospective studies, as even cooperative patients may not correctly recount a full history of events, behaviour, or practices. Several studies have shown that patient recall bias can vary with age, medical condition, and the type of service offered [139, 140, 141]. When recalling pregnancy history, patients are more likely to accurately remember number of pregnancies, miscarriages, birth weight, and type of delivery than medication and imaging [142, 143]. There has also been disagreement over the cause of the discrepancies, with evidence both for and against socioeconomic status, ethnicity, and neighbourhood type as contributing factors [144, 145]. Medical data must be accurate, reliable, and routinely recorded. Retrospective studies using such data would allow the investigation of rare diseases, accommodate assessment of conditions with long latency periods, and could function as pilot studies to identify weaknesses and improve study design for further prospective research [137]. Reliable data will improve healthcare systems worldwide, including those in the developing world. Following a surge in freely available internet-based technologies, shared data will have an important impact on worldwide public health.

Villar *et al.* [146] showed that certain obstetrical information retrieved from medical records can be reliable. In an inter-rater agreement study of antenatal and neonatal variables collected in a large teaching obstetric unit, information routinely collected

by hospital staff was compared with that collected by a specifically trained physician and social worker. They observed excellent agreement for some variables such as maternal and newborn anthropometric measures, and previous birthweight, but poor agreement for variables such as indicators of physical activity, work during pregnancy, and blood pressure measures. They hypothesised that the poor agreement for some of the variables was due to problems in how the questions were phrased, patients recall, interviewer bias, and data abstraction. They recommended that epidemiological studies using routine data should include a reliability component, and proper standardisation of personnel and instruments. They should also include validity data and examples of the questions used in any published reports [146].

2.11 Statistical methodology

When constructing reference or normal ranges, we want centiles that change smoothly with gestation and provide a good fit to the raw data, and a statistical model that is as simple as is necessary [67, 92, 22]. In preparation for MGRS analysis, the WHO conducted an extensive literature review of existing methods for the construction of growth curves [147]. When choosing the statistical analyses for constructing reference charts, (a) the statistical methods used should be clearly identified and described; (b) an assessment of whether the normality assumption is reasonable, as is usually the case for fetal data conditional on GA, should be conducted; (c) both the mean and SD should be modelled as a function of GA in a way that accounts for the increasing variability with gestation that is typical in growth data; (d) the modelling should provide smooth centile curves; and (e) a goodness-of-fit assessment with graphical evaluation of the superimposed centiles should be conducted to compare the predictive model to the raw data.

A review of neonatal size charts found that most of the 105 included studies described the statistical methods used, although more than half did not satisfy all of the above

criteria [16]. In Chapters 4 and 5, I discuss in detail the statistical methodology that can be applied to fetal and neonatal data with examples based on the analyses conducted as part of the INTERGROWTH-21st Project.

2.12 Handling of repeated anthropometric and ultrasound measures

In clinical practice, measurements are only made once because of clinical work load and thus most reference centiles are constructed using single measurements. However, in some research settings, including studies of growth, it is common to take duplicate or triplicate measurements. Repeated measurements are taken so that the mean, a better estimate of true value, can be used; to ensure data quality; and to allow the expected within- and between-variation among sonographers and anthropometrists to be calculated. Using the average of two or more measurements gives a more reliable estimate for an individual. Although using averages to develop centile charts tends to underestimate the actual variability for single measurements, the effect is minimised by using highly experienced trained staff and strict protocols to take highly reproducible measurements.

It is expected that averaging measurements reduces measurement variability, resulting in tighter (or narrower) centiles than if single measurements are used. For example, in the FGLS, ultrasound measurements were taken in triplicate and anthropometric measurements in duplicate at each patient visit. Statistical methods can be used to account for this reduced variability by applying a small correction to the observed variability of all of the repeated measures, as previously suggested by Bland and Altman [122]. However, this correction was developed in a different context and the effect on growth data is as yet unknown and requires exploration. Other methods, such as multi-level models, handle this issue differently: rather than taking the average of measurements, the data are treated in a hierarchical manner.

For example, in the FGLS, ultrasound data were measured three times on three separately obtained ultrasound images of each structure in a blinded fashion. The data structure was therefore a hierarchy with three levels: triplicate measurements taken at each visit, measured on three separately obtained ultrasound images of each structure in a blinded fashion (level 1), repeated ultrasound measurements per subject at designated visits during pregnancy (level 2), and subject measurements from each of the eight sites (level 3).

2.13 Handling data from multiple sites

Most studies of fetal and neonatal growth are done in a single centre. The need for a large sample size and greater generalisability may lead to a multicentre design, which brings additional challenges. Assessing how appropriate it is to pool data from multiple sites is challenging, as a judgment of the similarities in the fetal growth and newborn size patterns across the populations must be made. Subjects within the same site tend to be more similar to each other than to subjects from other sites. The combinability of studies in a meta-analysis is usually judged qualitatively based on the similarity of the studies, i.e. similarity of the participants, interventions, and outcome variables. It is also standard practice to quantify the statistical heterogeneity of the results [148, 149, 150, 151], although this is more likely to influence the type of analysis than whether the studies can be combined.

As multi-centre studies are rare in human growth studies, the combinability problem is not common. However, both the MGRS and INTERGROWTH-21st Project were faced with this problem and focussed on the differences between the results at each site. Although many statistical methods focus on differences between groups, there are no standard statistical rules to evaluate these differences. Statistical significance is not appropriate for judging combinability, as even unimportant differences can be statistically significant in very large samples. As the INTERGROWTH-21st Project

aimed to develop international standards, only a small amount of heterogeneity could be tolerated in the data. The differences between the sites had to be quantified and evaluated [12]. There was the added problem of needing to judge whether the data from the eight sites could be combined and quantifying the differences between the sites across time, which was meassured with GA. Changes in variability with GA also had to be taken into account. The goal of the analysis was thus to check whether the sites had similar fitted centile curves by GA, rather than similar raw data.

The goal is not to test a defined hypothesis. Instead, we are interested in whether there is an acceptable level of disagreement between each site when compared with all of the sites. It is more credible, and thus recommended, to set in advance the criterion for judging whether the differences between the centile curves from each site are acceptable before conducting the analysis.

The INTERGROWTH-21st Project used the same criteria as the MGRS: it was decided before the analysis that a difference of 0.5 SD or greater between the centile curves from a site and all of the sites at any GA would indicate that the data from that site were too different to be pooled [22, 19]. If data from all eight sites met this criterion from all of the analyses, all of the data would be pooled and used to construct international standards. It is recommended that multi-centre studies should quantify and evaluate the differences between their sites using prespecified criteria, as was done in the INTERGROWTH-21st Project.

2.14 Reporting and presenting results

Altman *et al.* [67] and Royston *et al.* [93] discussed and recommended appropriate ways of reporting and presenting growth study results. A table of included observations should be reported that shows how many individuals or women were recruited in each GA window (e.g., each week of pregnancy) with the mean, SD, and associated sample size for each measurement at each completed GA. Another

table should show selected fitted centile values (e.g., the 10^{th}, 50^{th}, and 90^{th}) and regression equations for both the mean and SD that enable the calculation of any desired centiles and z-scores. An adequate sample size at each GA is essential for improved precision estimates, especially at extreme percentiles (e.g., the 3^{rd} and 97^{th} percentiles) where variability is greatest. Ioannou *et al.* [28] found variable reporting of results in fetal growth studies. The vast majority of the publications included in their review reported charts of median and centile curves. Most of the included studies (82%) reported tables with median values and 78% included selected centiles. Sixty-percent of the included studies reported equations for the mean or median, but only 39% reported equations for the SD, either as a fixed number or as a function of GA [27].

A goodness-of-fit assessment, with graphical evaluation of the superimposed centiles, is essential for comparing a predictive model to the observed data. A smooth change in the mean superimposed onto the raw data should be reported to allow model assessment. The fitted centiles alone without superimposing raw measurement data do not allow judgements of model fit to be easily made. The INTERGROWTH-21^{st} review of 105 studies found that a quarter of the studies only reported raw centiles instead of smoothed centiles [16, 28]. In other cases, the regression model used to smooth the centiles was poorly described or not appropriate for the purpose. About 25% of the studies included in the review did not report any centiles and z-scores could only be computed for 16% of the studies. Z-scores (or parameters that allow them to be computed) and equations of the mean and SD should be provided to enable easy comparisons of studies.

2.15 Summary

Altman and Chitty [67] discussed some of the considerations for the design and methodology of studies of fetal size with the aim of improving the quality of future

studies. Royston and Altman [93] discussed longitudinal studies of fetal size for similar reasons. Ioannou *et al.* [28] reported a positive correlation between quality scores and year of publication [28], showing that the methodological quality of fetal size these studies is steadily improving thanks to efforts such as these. However, the INTERGROWTH-21st systematic reviews of pregnancy dating and fetal and newborn charts showed that many studies of fetal size are still conducted poorly [16, 28, 27]. This chapter discussed some of the key issues that should be considered when designing fetal size studies and built on similar work by Altman and Chitty [67] and Royston and Altman [93].

More work is needed on the calculation of the sample size required for fetal growth studies. In particular, aspects of longitudinal studies have not yet been considered, such as the effect of correlations between measurements of the same individual at different ages, the number of replicates per measurement, the timing of measurements (in general more measurements are needed in periods of more rapid growth to accurately capture the pattern of growth), and the number of observations per individual.

In this chapter, I have reiterated some of the concepts previously identified as important for growth studies, focusing on considerations and concepts related to study design. I have provided more details and practical examples based on our experiences from the INTERGROWTH-21st Project. Chapters 4 and 5 discuss statistical methodology and analyses for cross-sectional studies and longitudinal studies.

3

Assessing the combinability of linear growth data

3.1 Background

The combinability of data and its usefulness in comparing different populations is a complex concept that is difficult to assess precisely [152]. The combinability problem needs a solution, given the rise in collaborative research efforts to create international standards, guidelines and programmes. This reflects the development of an intellectual atmosphere in which researchers are increasingly looking for internationally comparable datasets to lend greater weight and significance to their research. There is also a growing need for data for comparisons and for monitoring trends across individual countries and groups of countries.

Chapter 2 briefly outlined design and methodological considerations when constructing human growth charts. Most studies of fetal and neonatal growth are conducted in a single centre. However, the need for a large sample size and greater generalisability

may lead to the use of a multicentre design, which brings additional challenges. For example, the INTERGROWTH-21st Project used a multi-centre design to develop international growth standards for preterm babies, fetuses, and newborns. As excessive heterogeneity is incompatible with the concept of internationally relevant standards, an evaluation or quantification of site differences was required. Ideally, all of the data from the eight study sites would be used to construct a single global standard for each measurement, as this would be the strongest basis for the construction of growth curves for international clinical applications. The similarity of the fetal growth and newborn size patterns in each population had to first be judged to determine whether pooling was appropriate. Subjects within the same site tend to be more similar to each other than to subjects from other sites. The data from the different centres thus had to be similar enough to be combined. As most human growth studies use a single centre, this problem is uncommon in the field and there is no universally recognised method for judging the acceptable amount of heterogeneity in growth data from several sites [12].

Statistical methods only exist for examining heterogeneity and combining results from several studies in meta-analysis. Meta-analysis is a two-stage process involving the calculation of an appropriate summary statistic for each of a set of studies followed by the combination of these statistics into a weighted average. The weights are usually chosen to reflect the amount of information that each trial or study contains. Methods are available for combining odds ratios, risk ratios and risk differences for binary data, and hazard ratios for time-to-event data. Continuous data can be combined as differences in means or as standardised differences in means when a mixture of measurement scales has been used.

An important step is the thoughtful consideration of whether it is appropriate to combine all or some of the studies in a meta-analysis to yield an overall summary statistic. In trials, for example, consistency of trial results with a common effect

across a variety of circumstances provides important, powerful corroboration of the generalisation of the treatment effect, so that a greater degree of certainty can be placed on its application to wider clinical practice [153]. Statistical investigation of the degree of variation between individual study results, which is known as heterogeneity, can often contribute to making decisions regarding the combinability of the results [154]. The combinability of studies in a meta-analysis context is usually judged qualitatively by the similarity of the studies, using, for example, the similarity of the participants, interventions, and outcome variables.

The INTERGROWTH-21st Project used a prescriptive approach to select women recruited in all the eight study sites so as to evaluate the growth potential of the fetuses and newborns. The study recruited women at low risk of fetal growth disturbances according to a predefined set of criteria [24] and the socio-economic and demographic characteristics of the underlying populations, in settings with diverse ethnic backgrounds. This strategy allowed fair comparisons across populations where the health and nutritional needs of mothers were met and adequate, standardised antenatal care was provided. The prescriptive approach used by INTERGROWTH-21st meant that the eight sites were largely similar, so a qualitative evaluation of how similar the eight sites were, like that commonly applied in meta-analyses, was not relevant.

In meta-analyses, it is also standard practice to quantify the statistical heterogeneity of the results [155, 149, 156, 157], although the results are more likely to influence the type of analysis used than whether studies can be combined. Statistical significance is not appropriate for judging combinability in the INTERGROWTH-21st context, as even unimportant differences can be statistically significant in large samples.

This chapter focuses on how to assess the combinability of linear growth data and how the differences between the eight sites were quantified. The data from all of the sites were compared using prespecified criteria. The cut-off for acceptability

was prespecified in advance of the analysis to enhance credibility. This enabled an assessment of whether the level of disagreement between the sites was low enough to allow pooling. The increase in variability with GA had to be taken into account. The analysis focused on the similarity between the fitted centile curves by GA at each site, rather than the similarity between the raw data from each site.

3.2 Methods

The INTERGROWTH-21st analytical strategy was conceptually similar to the strategy applied in the MGRS [158]. The two projects used identical methodology in geographically diverse urban areas in which mothers' health and nutritional needs were met, sanitation practices and the environment were judged not to constrain growth, and adequate, standardised antenatal care was provided. This methodology helped to reduce the variability that can typically be introduced by such factors.

3.2.1 Choosing measures for comparing populations

To decide whether the data from each site were combinable, fetal growth and newborn size from each population were explored by mapping skeletal growth as a continuous process from post-conception to birth. As in the MGRS, fat-independent measures of linear growth (i.e. CRL, FHC, and birth length (BL)) were chosen to compare the sites. These measures are recommended for comparing growth across ethnic or environmental conditions in different populations [19]. They are more resistant to environmental factors such as maternal nutrition, infections during pregnancy, or pregnancy-related complications (skewing) in response to "excessive nutrition" than weight or other fat-related indicators. The fetal HC (FHC) also allows for continuous evaluation with the same measurement in the postnatal period. Although linear measures of size can be affected by under nutrition or infection, these factors are unlikely to play an important role in a healthy population.

Furthermore, linear measures are approximately normally distributed (unlike fat-related indicators). They are more precise than fat-related measures especially for fetal ultrasound, and were also used to compare populations in the WHO MGRS to construct infant and child growth standards [19, 18, 159].

It was therefore decided *a priori* [11] that the similarities between fetal growth and newborn size would be compared using the fat-free mass indicators CRL $<14^{+0}$ weeks of gestation (Figure 3.1), FHC $\geq 14^{+0}$ weeks of gestation (Figure 3.2), and birth length (BL). CRL is generally accepted to be the best ultrasound marker in the first trimester. FHC is more resistant to intrauterine environmental insults than fat-related measures such as abdominal circumference. BL is the leading marker of linear growth and was the main measure used in the MGRS to compare populations [19]. HC at birth (BHC) was obtained as a complementary measure to match FHC because it is the only skeletal measure available from early pregnancy to childhood. Data were thus gathered on early fetal size, longitudinal fetal growth and newborn size for the complete FGLS cohort and on newborn size for the matching, low-risk, FGLS-like subpopulation from the eight study sites using standardised procedures.

It was decided before the analysis, that a difference of 0.5 SD or greater between the centile curves from one site and the pooled estimate using data from all of the sites at any GA, would indicate that the data from that site were too different to be pooled [22, 19]. If data from all of the eight sites were found to be combinable in all of the the analyses, all of the data would be pooled and used to construct international standards.

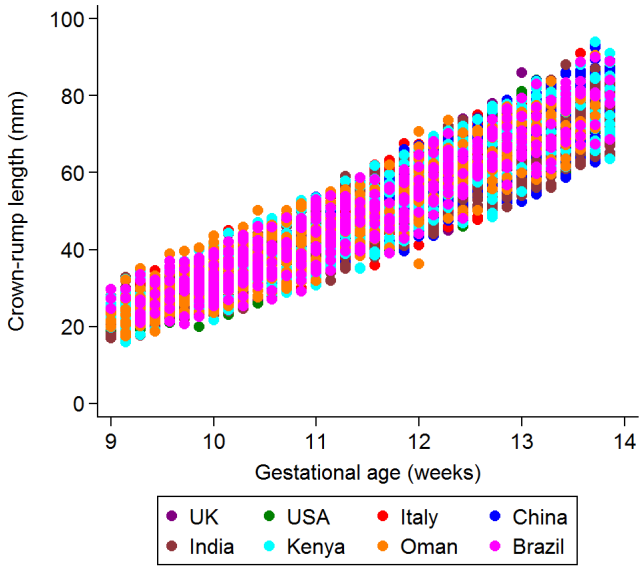


Figure 3.1: Scatter plot of crown-rump length according to gestational age (weeks), separated by study site

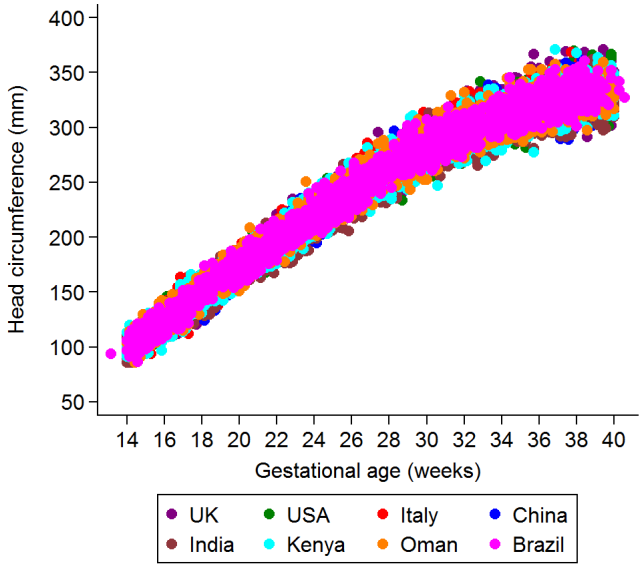


Figure 3.2: Scatter plot of head circumference according to gestational age (weeks), separated by study site

3.3 Analytical strategy

Data from all eight sites were carefully explored and evaluated to determine if they agreed with the prespecified criteria. The appropriateness of pooling the data from the sites for constructing standards was assessed by comparing the site means, SDs, and fitted centiles from each site with the corresponding values from the data from all of the sites combined. A difference of >0.5 SD between the values from an individual site and the pooled sample [19] was used as the trigger to consider whether to include the data from an individual site in the pooled data. The decision depended on the magnitude and nature of the difference between the data from that site and the pooled sample.

The proportion of the total variability that was explained by site differences was quantified. A sensitivity analysis was conducted to investigate the influence and robustness of the estimated smoothed centiles. The influence and robustness of each site's results was evaluated by excluding each site's data in turn from the pooled analysis at different GA and noting the degree to which the estimates and precision changed [19].

3.4 Data

The data were taken from the FGLS and NCSS of the INTERGROWTH-21st Project. Ultrasound was used to take fetal anthropometric measurements prospectively from 14^{+0} weeks until birth, in a cohort of women with optimal health and adequate nutritional status who were at low risk of intrauterine growth restriction. FHC obtained every 5 weeks ($\pm$ 1 week) from 14^{+0} to 42^{+0} weeks gestation was used, giving possible ranges of 14–18, 19–23, 24–28, 29–33, 34–38, and 39–43 weeks gestation. At each visit, FHC was measured three times on three separately obtained ultrasound images of each structure. Each image was measured once in a blinded fashion [24].

The FHC data structure was composed of three hierarchies: measurements within visits within participants. Level 1 is the three measurements taken at a visit, HC_1, HC_2 and HC_3. Level 2 is the repeated ultrasound measurements taken for each subject over multiple visits during pregnancy. Level 3 is the measurements taken for multiple subjects at each of the eight sites.

The FGLS enrolled 4,233 women who each visited between one and six times ($\sim$95% visited at least four times during pregnancy), giving 20,030 women visits. At each visit, three ultrasounds were collected, resulting in 59,973 FHC observations across the eight sites. One hundred and seventeen ultrasound measures were missing. The fetal ultrasound information was complemented with birthweight (BW), BL and BHC, collected within 12 hours of birth, in the NCSS study. These data were used to evaluate linear growth prospectively in a multi-ethnic population from <14 weeks of pregnancy until the immediate neonatal period.

3.5 Statistical methods and results

This section focuses on four methods that were selected for quantifying the combinability of the data from the eight sites. Three of the four methods were applied to child growth data by the WHO in the MGRS, but none had been assessed with fetal or newborn data before. As the four methods were complementary, the results from each method in support of or against pooling data from any individual site were expected to agree with one another.

3.5.1 Variance component analysis

Meta-analyses are often based on binary end-points and usually use cut-off points for continuous variables. Classic meta-analyses adjust for the effect of site as either a fixed effect or a random effect [160, 161, 162]. They aim to evaluate site effects and compare them against each other while estimating the overall variability of

the site effects. This is important as it informs the eventual choice of analyses and allows between-site relationships to be quantified.

An analysis of variance (ANOVA, variance component analysis) approach was used to quantify the amount of variability in the fetal and newborn measurements that was attributable to site identity. The percentage of variance in the cross-sectional measures (CRL, BHC, and BL) due to between-site variance was calculated [19]. The FHC was measured several times during pregnancy in the same subject. This longitudinal design introduced another level of complexity, which was dealt with by estimating the variance between individuals in the same site (within-site variance). A multilevel random effects regression model was applied for the cross-sectional and repeated measures as appropriate, adjusting for GA. GA was treated as a fixed effect, whereas site and individual identity were treated as random effects.

Variance component	Crown-rump length (N = 4,265)		Fetal head circumference (N = 4,237)		NCSS (FGLS-like) newborn length [a] (N = 20,166)	
	Estimate (SE)	Proportion	Estimate (SE)	Proportion	Estimate (SE)	Proportion
Variance between sites	0.65 (0.38)	1.9%	5.15 (2.82)	2.6%	0.12 (0.06)	3.5%
Variance between individuals within a site	[b]	[b]	36.64 (1.54)	18.6%	[b]	[b]
Residual variance	33.64 (0.73)	98.1%	155.70 (1.73)	78.8%	3.34 (0.02)	96.5%

[a]The NCSS (FGLS-like) subpopulation is composed of the low-risk women in the total NCSS population. They were selected using the same eligibility criteria as in the FGLS population.

[b]Variance between individuals within the same site cannot be estimated because the measures were only collected once per subject.

[c]Adjusted by gestational age as a fixed effect.

Table 3.1: Variance component analysis for crown-rump length, fetal head circumference and newborn length [c] for pregnancies with a live singleton birth and no congenital malformation.

3.5.2 Meta-analytic assessment using regression analysis

Sauerbrei and Royston [157] proposed a new strategy for meta-analysis of continuous covariates in observational studies that can be applied with modifications to the FGLS and NCSS data. Their strategy is based on estimating the functional relationship between a continuous covariate and the outcome in a regression model while adjusting for confounding factors. The functional form for the continuous variable of interest in each site (i.e., GA and the biometric parameters of CRL, FHC, BL, and BHC here) is first estimated. The individual site functions are then combined by weighted averaging to obtain summary estimates of the functions. This approach can be implemented using fractional polynomials (FPs). A key advantage is that the variability in the individual sites is reflected and comparisons can be made between the sites.

CRL and FHC data from each site were modelled separately for each measure as a function of GA using the FP regression method. In contrast to Sauerbrei and Royston's proposed method, weightings were not applied as the study was designed to ensure an equal contribution from each of the eight sites. The fitted FP models for each site were compared by superimposing selected centiles. An overall FP model was also obtained using the data combined from all eight sites. This overall model was then fitted to each site's individual dataset. The model fits were superimposed and compared. Judgements were made through visual inspection of the growth or size patterns across the sites.

Country	Crown-rump length		Fetal head circumference	
	FP powers for the mean	FP powers for the SD	FP powers for the mean	FP powers for the SD
Brazil	3	1	2, 2	1
China	1	1	2, 3	1
India	-2, -2	-2	2, 2	1
Kenya	2	1	2, 2	1
Oman	2	1	2, 2	1
UK	1	1	2, 3	1
USA	1	1	2, 2	1
Italy	1	1	2, 3	1

Fractional polynomials especially FP2 and FP3 based on different powers can sometimes result in very similar overall fit of the data. For example, differences in general fit between an FP2 with powers ($p_1 = 2$ and $p_2 = 2$) and FP2 with powers (($p_1 = 2$ and $p_2 = 3$) can be indistinguishable. Decisions on the best model fit need not be based on statistical significance testing alone as other evaluations based on general fit, a trade-off between a simple versus more complex model, and diagnostics should be considered.

Table 3.2: Summary of the fitted fractional polynomial (FP) powers for the mean and standard deviation (SD) for each sites' crown-rump length and fetal head circumference.

Fractional polynomials offer great flexibility in allowing non-integer powers, logarithms, and repetition of powers [163, 94]. FPs are defined by power terms restricted to a predefined set of integer and non-integer values, p (-2, -1, -0.5, 0, 0.5, 1, 2, 3). The power 0 denotes natural logarithmic transformation, so that x^0 equals to $log_e(x)$ rather than $x^0 = 1$. In the case of repeat powers, the second term is multiplied by $log_e(x)$. The best power transformation, x^p, is chosen from the set of powers, p, with software. An automated algorithm for selecting these powers has already been implemented in the statistical software programs STATA and R. The degree of an FP model is defined as the number of powers, p, of the explanatory variable. For instance, a first-degree FP (FP1) model with power p will be of the form $Y = b_0 + b_1 x^{p_1}$ and an FP2 of the form $Y = b_0 + b_1 x^{p_1} + b_2 x^{p_2}$.

For example, the best FHC model for the mean using data from Brazil was an FP2 ($p_1 = 2$ and $p_2 = 2$) and will be of the form $Y = b_0 + b_1 x^2 + b_2 x^2 log_e(x)$. Similarly for China, the FHC model for the mean will be of the form $Y = b_0 + b_1 x^2 + b_2 x^3$.

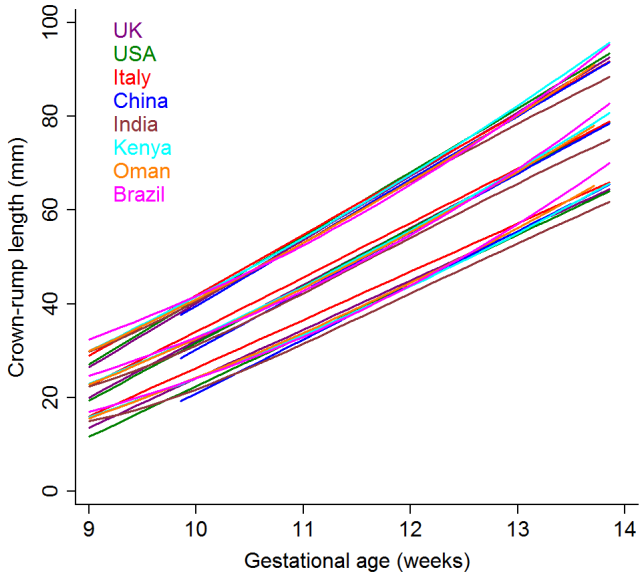


Figure 3.3: Fitting a separate fractional polynomial model to each site's crown-rump length (CRL) data.

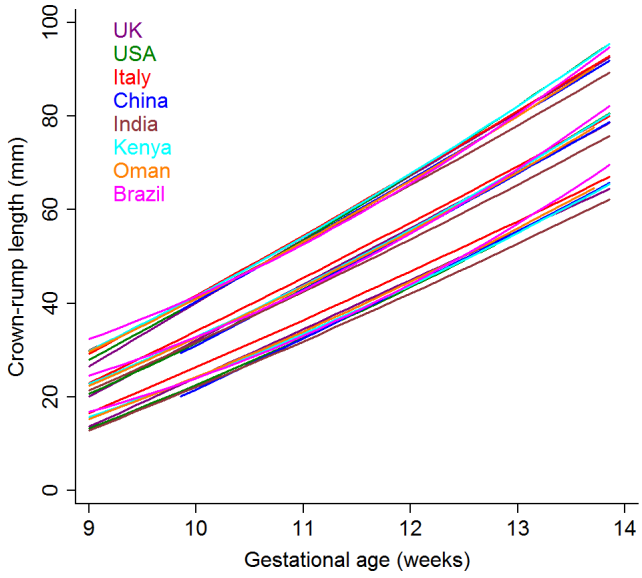


Figure 3.4: Fitting the overall fractional polynomial model using data from all of the sites (mean = 1, 2 and standard deviation = 1) to each site's crown-rump length (CRL) data.

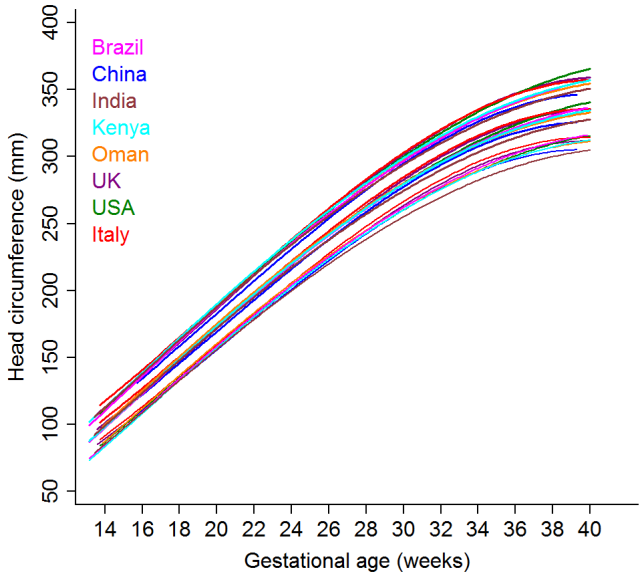


Figure 3.5: Fitting a separate FP model to each site's data.

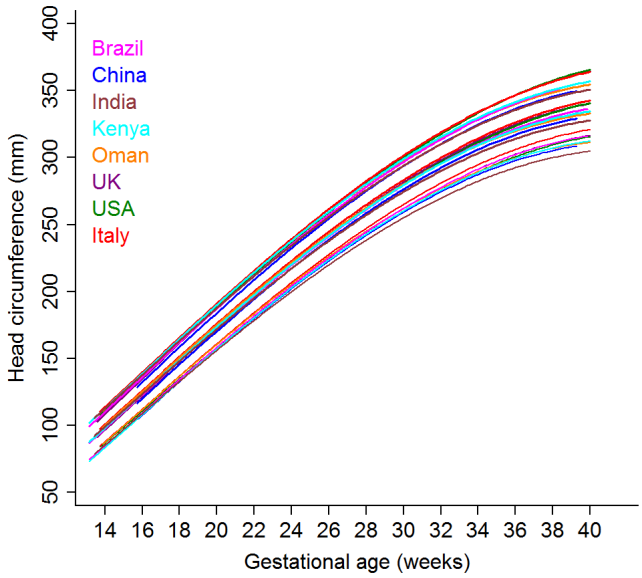


Figure 3.6: Fitting the overall FP model using data from all of the sites (Mean = 2, 2 and SD = 1) to each site's data.

3.5.3 Standardised site difference

The standardised site difference (SSD) is a statistical measure of great biological value. It allows the direct comparison of different biometric parameters and is commonly used to compare growth in different populations. SSD was used to compare the fetal and newborn anthropometric measures CRL, FHC, BHC, and BL collected from the eight sites across GA. An SSD of <0.5 was prespecified in the INTERGROWTH-21st Project protocol as adequate for combining data from all of the sites to create international standards [22].

The SSD was computed for each site and for each measure by comparing the crude site means for CRL, HC, BHC, and BL at different GA windows to the crude pooled mean (i.e. the mean calculated from the pooled eight-site data set) at the corresponding GA windows, expressed as a function of the crude pooled SD (Approach 1). For example, the GA windows in which FHC measures were taken were 14–19, 20–24, 25–29, 30–34, 35–39, and 40–43 weeks. The differences were calculated as the difference between the mean from a specified site and the mean of the pooled sites. Each difference was expressed as a proportion of the pooled SD at each corresponding GA to give the SSD score [19], which is similar to a z-score.

To account for increasing variance with GA, an adjusted FHC measure was calculated from the expected mean FHC measurement obtained from fitting an FP regression model, adjusted at the midpoint of each GA interval. This approach assumed uniform growth during each 5-week interval. An adjusted SSD was thus calculated based on the adjusted FHC measurements compared with the pooled adjusted FHC values and pooled adjusted SD at the corresponding GA categories (Approach 2).

Approach 1	1. For each site and GA category, 14–19, 20–24, 25–29, 30–34, and 35–39 weeks, calculate the crude observed mean FHC (mm) and crude observed SD 2. Calculate a crude SSD measure for each site according to its GA category: **Crude SSD = (crude site mean of observed CRL/FHC (mm) – crude all sites mean of observed CRL/FHC (mm)) / crude all sites SD of observed CRL/FHC (mm)**
Approach 2	1. Fit an FP regression model for the mean and SD relating FHC to GA category using the pooled data set 2. For each observed CRL/FHC measurement (CRL_{obs}/FHC_{obs}), obtain the expected mean CRL/FHC, $\text{CRL}_{exp}/\text{FHC}_{exp}$, and SD, SD_{exp}, from the equations of the mean and SD from the best-fit FP model according to GA and calculate a z-score: **z-score = (CRL_{obs} / FHC_{obs} – CRL_{exp} / FHC_{exp}) / SD_{exp}** 3. For each site and GA category, calculate an SSD measure for each site according to GA group: **SSD = (site mean z-scores – all sites mean z-scores) / all sites SD of z-scores** 4. Based on the z scores, calculate an adjusted FHC measurement (adj. HC_{obs}) at the median GA for classes 14–19, 20–24, 25–29, 30–34, and 35–39 weeks: **Adjusted $\text{CRL}_{obs}/\text{FHC}_{obs}$ = (z-score*SD_{exp}) + $\text{CRL}_{exp}/\text{FHC}_{exp}$** **at the median GA for each GA category** 5. For each site and gestational age category, calculate the mean adjusted HC and the corresponding SD of the adjusted HC. 6. Calculate an SSD measure for each site according to each GA category **Adjusted SSD = (site mean adjusted CRL/FHC (mm) – all sites mean adjusted CRL/FHC (mm)) / all sites SD of adjusted CRL/FHC (mm)**

Table 3.3: Summary of the crude and adjusted approaches used to calculate a standardised site difference (SSD). CRL, crown-rump length; *exp*, expected; FHC, fetal head circumference; GA, gestational age; HC, head circumference; *obs*, observed; SD, standard deviation.

		Approach 1					Approach 2		
GA (weeks)	Country	Sample (n)	GA (days)	CRL (mm)	CRL (mm)	Crude standardised site difference	CRL (mm)	CRL (mm)	Adjusted standardised site difference
			Median	Mean	SD	SSD	Mean	SD	SSD
9^{+0} to 9^{+6}	Brazil	66	67	29.49	4.46	0.43	27.86	4.26	0.26
	China	1	69	35.50	.	1.67	30.89	.	1.04
	India	109	66	25.98	4.73	-0.30	26.02	3.95	-0.20
	Kenya	116	66	27.22	4.64	-0.04	27.11	3.74	0.07
	Oman	147	67	28.39	4.99	0.20	27.08	3.91	0.06
	UK	25	66	25.01	3.54	-0.50	25.42	2.74	-0.36
	USA	34	66	25.56	4.85	-0.39	25.45	3.36	-0.35
	Italy	43	66	27.98	4.23	0.11	27.47	4.12	0.16
	All	541	66	27.43	4.84	0.00	26.83	3.91	0.00
10^{+0} to 10^{+6}	Brazil	108	73	35.94	5.45	-0.05	36.38	4.84	-0.02
	China	22	75	38.10	5.80	0.30	35.97	4.82	-0.10
	India	92	73	35.17	6.49	-0.17	35.45	5.34	-0.20
	Kenya	110	73	37.17	6.44	0.15	37.29	5.19	0.16
	Oman	169	72	36.23	5.94	0.00	36.74	5.14	0.05
	UK	19	73	36.49	6.62	0.04	36.14	5.94	-0.06
	USA	27	73	35.43	6.78	-0.13	35.83	5.37	-0.12
	Italy	22	72.5	36.44	5.89	0.03	36.64	6.20	0.03
	All	569	73	36.24	6.11	0.00	36.47	5.20	0.00
11^{+0} to 11^{+6}	Brazil	90	79	47.49	6.27	-0.28	49.60	5.33	-0.16
	China	89	81	49.07	6.76	-0.04	49.78	5.65	-0.13
	India	140	80	46.96	6.59	-0.36	48.92	6.06	-0.28
	Kenya	119	80	48.17	7.03	-0.18	49.43	6.01	-0.19
	Oman	106	80	48.07	5.32	-0.19	50.27	4.78	-0.04
	UK	105	82	51.46	6.76	0.32	50.61	6.14	0.02
	USA	54	81	48.87	8.06	-0.07	49.94	6.44	-0.10
	Italy	259	81	51.71	5.57	0.36	52.53	4.60	0.36
	All	962	81	49.36	6.61	0.00	50.49	5.62	0.00

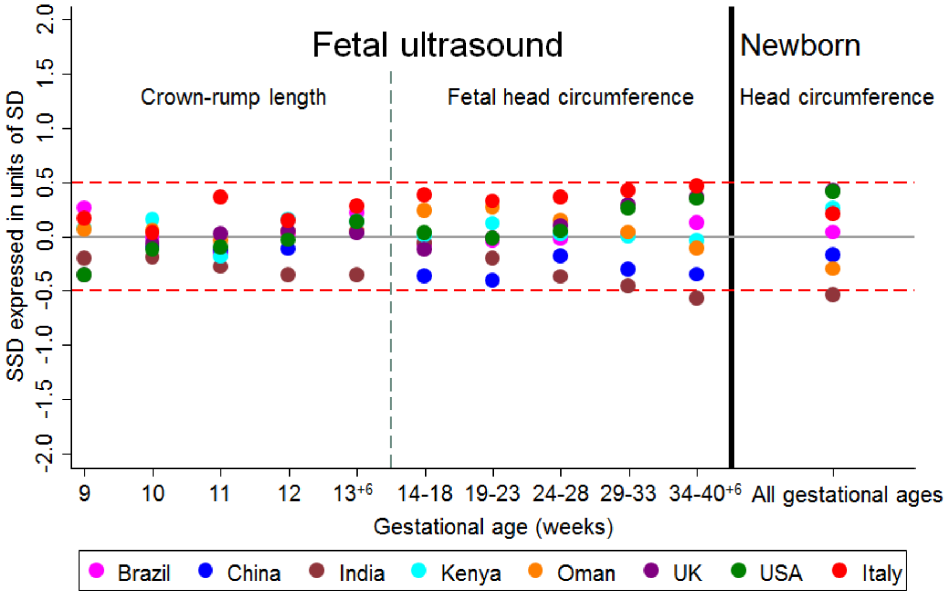
		Approach 1					Approach 2		
GA (weeks)	Country	Sample (n)	GA (days)	CRL (mm)	CRL (mm)	Crude standardised site difference	CRL (mm)	CRL (mm)	Adjusted standardised site difference
			Median	Mean	SD	SSD	Mean	SD	SSD
12^{+0} to 12^{+6}	Brazil	88	87	61.24	6.72	0.07	61.42	5.35	0.03
	China	204	87	60.64	7.12	-0.02	60.54	6.17	-0.12
	India	122	87	59.45	7.20	-0.19	59.01	7.11	-0.36
	Kenya	154	87	62.00	7.10	0.18	62.22	6.30	0.15
	Oman	80	87	61.38	7.64	0.09	61.55	6.37	0.05
	UK	358	86	60.71	6.81	-0.01	61.56	5.97	0.05
	USA	98	87	60.62	7.63	-0.02	61.04	6.30	-0.04
	Italy	164	86	60.35	6.67	-0.06	62.17	6.20	0.15
	All	1268	87	60.76	7.04	0.00	61.26	6.25	0.00
13^{+0} to 13^{+6}	Brazil	60	93	73.48	7.64	0.13	74.37	6.52	0.22
	China	293	94	72.86	7.50	0.05	73.20	6.57	0.04
	India	162	94	71.02	6.88	-0.21	70.52	6.13	-0.36
	Kenya	100	93	73.37	7.19	0.12	73.82	6.93	0.13
	Oman	54	94	72.37	6.27	-0.02	73.23	6.23	0.05
	UK	123	93	71.87	7.22	-0.09	73.13	7.02	0.03
	USA	52	94	73.68	7.40	0.16	73.84	7.15	0.14
	Italy	20	92	73.90	8.11	0.19	74.78	7.11	0.28
	All	864	94	72.52	7.28	0.00	72.92	6.71	0.00

Table 3.4: Sample sizes, means, and standard deviations (SD) for crown-rump length (CRL) (mm) for each site and for the pooled eight-site data set. GA, gestational age.

		Approach 1					Approach 2		
GA (weeks)	Country	Sample (n)	GA (days)	FHC (mm)	FHC (mm)	Crude standardised site difference	FHC (mm)	FHC (mm)	Adjusted standardised site difference
			Median	Mean	SD	SSD	Mean	SD	SSD
14^{+0} to 18^{+6}	Brazil	390	115	128.67	17.75	-0.32	135.71	6.82	-0.08
	China	552	125	142.02	11.46	0.48	133.56	6.72	-0.36
	India	563	118	132.19	18.17	-0.11	135.82	7.97	-0.06
	Kenya	546	119	132.72	19.58	-0.07	136.35	8.13	0.01
	Oman	536	112	128.44	16.85	-0.33	138.08	7.05	0.24
	UK	584	122	138.21	14.74	0.25	135.40	7.23	-0.12
	USA	294	116	131.48	17.22	-0.15	136.54	7.12	0.03
	Italy	510	117	135.02	12.96	0.06	139.18	7.30	0.38
	All	3975	119	133.97	16.78	0.00	136.30	7.52	0.00
19^{+0} to 23^{+6}	Brazil	393	149	187.96	16.88	-0.28	194.14	7.72	-0.05
	China	582	161	202.54	13.80	0.58	191.11	7.46	-0.41
	India	571	151	189.14	17.29	-0.21	192.84	8.71	-0.20
	Kenya	577	152	191.94	19.27	-0.04	195.52	8.73	0.12
	Oman	550	146	187.62	16.42	-0.30	196.77	7.99	0.27
	UK	591	156	195.21	16.39	0.15	194.40	8.03	-0.01
	USA	293	152	191.85	17.14	-0.05	194.38	8.07	-0.02
	Italy	509	149	192.88	14.20	0.01	197.24	8.31	0.33
	All	4066	153	192.68	17.13	0.00	194.52	8.38	0.00
24^{+0} to 28^{+6}	Brazil	378	183	243.84	16.52	-0.21	248.53	8.87	-0.02
	China	545	196	256.94	15.91	0.53	246.96	9.00	-0.19
	India	584	186	242.65	17.29	-0.28	245.14	9.83	-0.38
	Kenya	578	188	247.06	19.40	-0.03	248.96	10.20	0.02
	Oman	553	181	242.56	16.80	-0.29	250.13	8.94	0.14
	UK	597	188	249.34	17.40	0.10	249.70	9.41	0.10
	USA	259	186	247.33	18.19	-0.02	249.20	9.66	0.05
	Italy	472	185	250.54	14.55	0.17	252.25	9.35	0.36
	All	3966	187	247.61	17.69	0.00	248.76	9.65	0.00

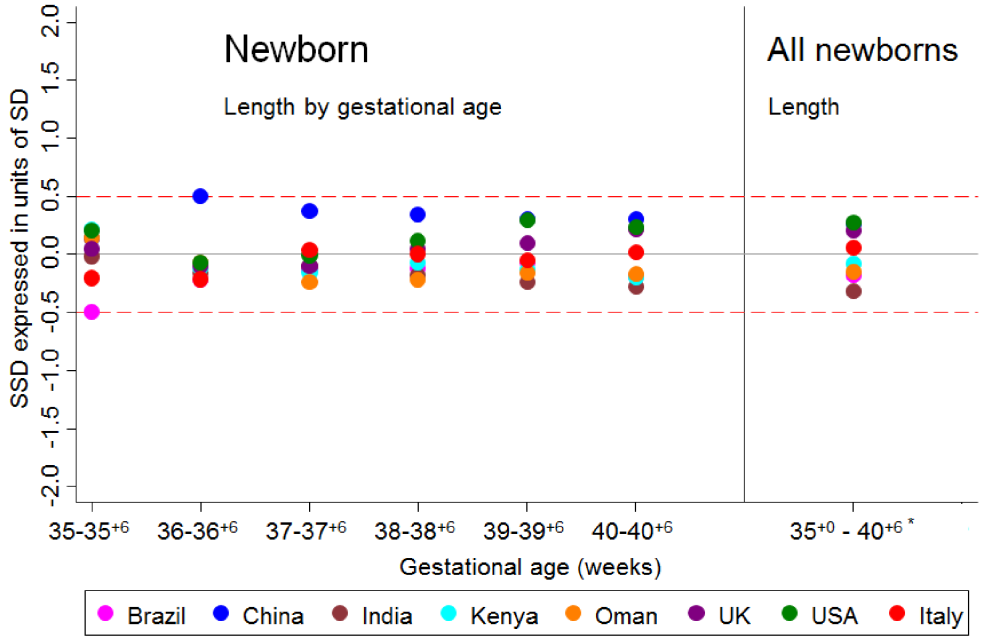
		Approach 1					Approach 2		
GA (weeks)	Country	Sample (n)	GA (days)	FHC (mm)	FHC (mm)	Crude standardised site difference	FHC (mm)	FHC (mm)	Adjusted standardised site difference
			Median	Mean	SD	SSD	Mean	SD	SSD
29^{+0} to 33^{+6}	Brazil	380	218	290.57	13.71	-0.14	294.37	9.50	0.02
	China	660	232	296.77	14.58	0.27	290.88	9.77	-0.31
	India	543	221	287.15	14.52	-0.36	289.23	10.85	-0.46
	Kenya	569	222	291.92	15.96	-0.05	294.20	10.69	0.00
	Oman	545	214	287.62	14.38	-0.33	294.60	10.15	0.04
	UK	590	221	295.32	14.87	0.18	297.31	10.70	0.29
	USA	265	221	294.91	16.06	0.15	296.97	11.06	0.26
	Italy	441	221	297.23	13.37	0.30	298.77	9.86	0.43
	All	3993	222	292.65	15.16	0.00	294.20	10.76	0.00
34^{+0} to 40^{+0}	Brazil	408	255	323.35	12.17	0.05	325.64	10.32	0.12
	China	579	264	322.58	11.35	-0.01	320.07	10.64	-0.35
	India	509	255	315.14	13.02	-0.55	317.42	11.41	-0.58
	Kenya	604	258	322.48	14.04	-0.02	323.71	11.40	-0.04
	Oman	601	249	318.56	12.75	-0.30	322.94	10.57	-0.11
	UK	712	258	327.08	14.02	0.32	328.44	12.04	0.36
	USA	255	257	326.58	14.47	0.29	328.27	11.77	0.34
	Italy	424	256	327.68	11.95	0.37	329.71	10.47	0.46
	All	4092	258	322.69	13.61	0.00	324.22	11.80	0.00

Table 3.5: Sample sizes, means, and standard deviations (SD) for head circumference (HC) (mm) for each site and for the pooled eight-site data set. GA, gestational age.



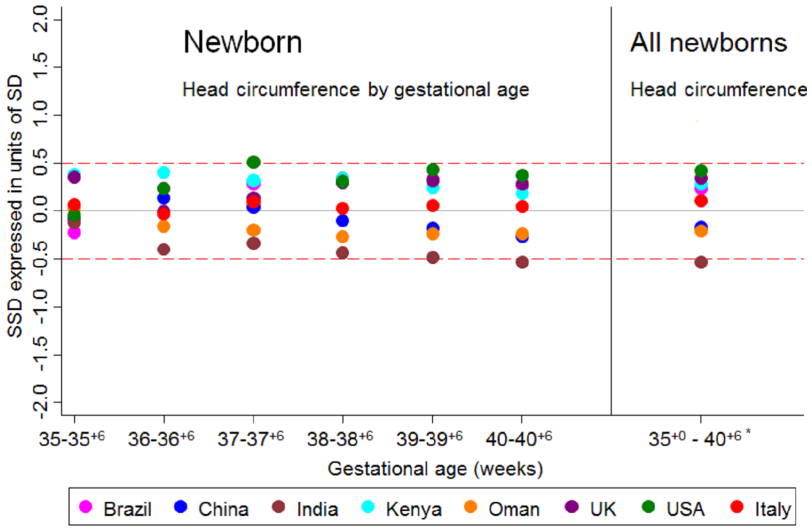
.

Figure 3.7: Standardised site difference (SSD) for crown-rump length (CRL) (N = 4,265), fetal head circumference (FHC) (N = 4,237) and head circumference at birth (BHC) (4,217) of the Fetal Growth Longitudinal Study (FGLS). SSD = (site mean FHC/CRL/BHC – all sites' mean FHC/CRL/BHC at each gestational age interval) / all sites' standard deviation (SD) of FHC/CRL/BHC at each GA interval. SSD was adjusted at the median gestational age for all sites at each gestational age interval. SSD at birth are mean values by site across all GAs, 25^{+0} to 43^{+0} weeks. The dashed red horizontal line shows 0.5 SD [11]



.

Figure 3.8: Standardised site difference (SSD) of newborn length (N = 20,166). The NCSS (FGLS-like) subpopulation represents the low-risk women in the total Newborn Cross-sectional Study (NCSS) population, selected using the same eligibility criteria as in the Fetal Growth Longitudinal Study (FGLS).
SSD = (site mean newborn length – all sites' mean newborn length at each GA interval) / all sites' standard deviation (SD) of newborn length at each GA interval. SSD was adjusted at the median GA for all sites at each GA interval. SSD at birth are mean values by site across all GAs 35^{+0} to 40^{+6} weeks. The dashed red horizontal line shows 0.5 SD [11]



.

Figure 3.9: Standardised site difference (SSD) of newborn head circumference (N = 20,046). The NCSS (FGLS-like) subpopulation represents the low-risk women in the total Newborn Cross-sectional Study (NCSS) population, selected using the same eligibility criteria as in the Fetal Growth Longitudinal Study (FGLS).
SSD = (site mean newborn head circumference – all sites' mean newborn head circumference at each GA interval) / all sites' standard deviation (SD) of newborn head circumference at each GA interval. SSD was adjusted at the median GA for all sites at each gestational age interval. SSD at birth are mean values by site across all GAs $35^{+0}4$ to 40^{+6} weeks calculated as above. The dashed red horizontal line shows 0.5 SD [11]

3.5.4 Sensitivity analysis

A sensitivity analysis was performed to evaluate the effect of excluding the data from a single site, in comparison with including the data from all eight sites. It showed whether excluding data from a particular site affected the overall centile fit based on the pooled data from all eight sites. The sensitivity analysis was based on the results from the FP regression modelling and was also expressed as SSDs. The effect of excluding each site's data on the pooled mean was assessed using selected percentiles (3^{rd}, 50^{th} and 97^{th} centiles) relative to the pooled mean for each of the GA categories and expressed as a function of the pooled SD. The graphs for each measure were compared visually, supported by calculations of SSD for each centile within the GA windows. The effect of site heterogeneity was assessed. An acceptable effect was a difference of < 0.50 SSD. If the difference was greater than this cut-off, further investigation was carried out to establish whether there was a consistent pattern across different GA windows and other fetal biometry for that site. If the difference was not consistent across different GA windows and other fetal biometry, the difference was considered acceptable and no action was taken.

Strategy	1. For each GA category (for example, for CRL, 9 to <10, 10 to <11, 11 to <12, 12 to <13 and 13 to <14 weeks), for each measure, calculate the observed mean and SD of the measurements pooled from all eight sites 2. For each measure and GA category, calculate the observed mean and SD of the measurements pooled from seven sites, excluding one site at a time 3. Calculate the SSD for each GA category when pooling the data from seven sites, excluding one site at a time: **SSD = (Observed mean CRL/FHC (mm) when excluding one site at a time - All sites mean observed CRL/FHC (mm)) / All sites SD of observed CRL/FHC (mm)**

Table 3.6: Summary of the approach used to calculate a standardised site difference (SSD) for sensitivity analysis. CRL, crown-rump length; FHC, fetal head circumference GA, gestational age.

GA (weeks)	Country	Sample	C50	SD	C50 SSD	C3	C3 SSD	C97	C97 SSD
9^{+0} to 9^{+6}	Pooled	541	27.43	4.84	0.00	19.30	0.00	36.53	0.00
	Excluding Brazil	475	27.16	4.81	-0.06	19.21	-0.02	36.19	-0.07
	Excluding China	540	27.40	4.81	-0.01	19.30	0.00	36.55	0.00
	Excluding India	432	27.80	4.77	0.08	19.65	0.07	36.69	0.03
	Excluding Kenya	425	27.47	4.87	0.01	19.72	0.09	37.00	0.10
	Excluding Oman	394	27.04	4.70	-0.08	19.24	-0.01	35.93	-0.12
	Excluding UK	516	27.54	4.84	0.02	19.31	0.00	36.98	0.09
	Excluding USA	507	27.52	4.81	0.02	19.57	0.06	37.00	0.10
	Excluding Italy	498	27.38	4.87	-0.01	19.25	-0.01	37.00	0.10
10^{+0} to 10^{+6}	Pooled	569	36.24	6.11	0.00	25.00	0.00	48.00	0.00
	Excluding Brazil	461	36.28	6.28	0.01	25.00	0.00	48.13	0.02
	Excluding China	547	36.24	6.15	0.00	25.00	0.00	48.02	0.00
	Excluding India	477	36.55	6.05	0.05	25.40	0.07	48.13	0.02
	Excluding Kenya	459	36.10	6.05	-0.02	25.00	0.00	47.18	-0.13
	Excluding Oman	400	36.35	6.23	0.02	25.00	0.00	48.00	0.00
	Excluding UK	550	36.31	6.13	0.01	25.00	0.00	48.00	0.00
	Excluding USA	542	36.37	6.09	0.02	25.00	0.00	48.02	0.00
	Excluding Italy	547	36.32	6.15	0.01	25.00	0.00	48.01	0.00
11^{+0} to 11^{+6}	Pooled	962	49.36	6.61	0.00	35.97	0.00	61.08	0.00
	Excluding Brazil	872	49.52	6.62	0.02	36.00	0.00	61.55	0.07
	Excluding China	873	49.34	6.61	0.00	36.00	0.00	61.49	0.06
	Excluding India	822	49.73	6.55	0.06	36.00	0.00	61.55	0.07
	Excluding Kenya	843	49.48	6.55	0.02	36.05	0.01	61.47	0.06
	Excluding Oman	856	49.47	6.76	0.02	36.00	0.00	61.55	0.07
	Excluding UK	857	49.04	6.56	-0.05	35.67	-0.05	60.30	-0.12
	Excluding USA	908	49.32	6.53	-0.01	36.00	0.00	60.80	-0.04
	Excluding Italy	703	48.47	6.77	-0.13	35.30	-0.10	60.73	-0.05

GA (weeks)	Country	Sample	C50	SD	C50 SSD	C3	C3 SSD	C97	C97 SSD
12^{+0} to 12^{+6}	Pooled	1268	60.76	7.04	0.00	47.70	0.00	74.42	0.00
	Excluding Brazil	1180	60.67	7.11	-0.01	47.46	-0.03	74.23	-0.03
	Excluding China	1064	60.77	7.07	0.00	48.00	0.04	74.09	-0.05
	Excluding India	1146	60.89	7.05	0.02	47.88	0.03	74.70	0.04
	Excluding Kenya	1114	60.56	7.05	-0.03	47.63	-0.01	74.20	-0.03
	Excluding Oman	1188	60.70	7.03	-0.01	47.70	0.00	74.24	-0.03
	Excluding UK	910	60.76	7.19	0.00	47.00	-0.10	74.70	0.04
	Excluding USA	1170	60.79	7.00	0.00	48.00	0.04	74.10	-0.05
	Excluding Italy	1104	60.82	7.12	0.01	47.42	-0.04	74.69	0.04
13^{+0} to 13^{+6}	Pooled	864	72.52	7.28	0.00	58.00	0.00	86.24	0.00
	Excluding Brazil	804	72.47	7.28	-0.01	57.62	-0.05	86.11	-0.02
	Excluding China	571	72.35	7.18	-0.02	58.00	0.00	86.00	-0.03
	Excluding India	702	72.90	7.35	0.05	58.70	0.10	86.63	0.05
	Excluding Kenya	764	72.42	7.31	-0.01	57.40	-0.08	86.00	-0.03
	Excluding Oman	810	72.54	7.36	0.00	57.59	-0.06	86.34	0.01
	Excluding UK	741	72.65	7.30	0.02	58.00	0.00	86.45	0.03
	Excluding USA	812	72.44	7.27	-0.01	57.82	-0.02	85.99	-0.03
	Excluding Italy	844	72.50	7.27	0.00	58.00	0.00	86.10	-0.02

Table 3.7: Means (C50), standard deviations (SD), 3^{rd} centiles (C3) and 97^{th} centiles (C97) for crown-rump length (CRL) (mm) for the pooled data set and for the pooled data set with one site's data excluded. GA, gestational age; SSD, standardised site difference.

GA (weeks)	Country	Sample	C50	SD	C50 SSD	C3	C3 SSD	C97	C97 SSD
14^{+0} to 18^{+6}	Pooled	3975	133.97	16.78	0.00	101.85	0.00	162.34	0.00
	Excluding Brazil	3585	134.55	16.57	0.03	101.92	0.00	162.39	0.00
	Excluding China	3423	132.67	17.14	-0.08	100.97	-0.05	162.38	0.00
	Excluding India	3412	134.27	16.52	0.02	102.41	0.03	162.36	0.00
	Excluding Kenya	3429	134.17	16.28	0.01	102.69	0.05	162.00	-0.02
	Excluding Oman	3439	134.83	16.61	0.05	101.67	-0.01	162.34	0.00
	Excluding UK	3391	133.24	17.00	-0.04	101.47	-0.02	162.33	0.00
	Excluding USA	3681	134.17	16.73	0.01	101.88	0.00	162.34	0.00
	Excluding Italy	3465	133.97	16.78	0.00	101.85	0.00	162.34	0.00
19^{+0} to 23^{+6}	Pooled	4066	192.68	17.13	0.00	159.75	0.00	222.40	0.00
	Excluding Brazil	3673	193.19	17.08	0.03	159.76	0.00	222.79	0.02
	Excluding China	3484	191.04	17.07	-0.10	159.75	0.00	222.58	0.01
	Excluding India	3495	193.26	17.03	0.03	160.56	0.05	222.70	0.02
	Excluding Kenya	3489	192.81	16.74	0.01	160.78	0.06	222.56	0.01
	Excluding Oman	3516	193.48	17.10	0.05	159.87	0.01	222.39	0.00
	Excluding UK	3475	192.25	17.21	-0.03	159.50	-0.01	222.40	0.00
	Excluding USA	3773	192.75	17.12	0.00	159.75	0.00	222.40	0.00
	Excluding Italy	3557	192.66	17.51	0.00	159.38	-0.02	222.41	0.00
24^{+0} to 28^{+6}	Pooled	3966	247.61	17.69	0.00	213.49	0.00	278.04	0.00
	Excluding Brazil	3588	248.01	17.76	0.02	213.47	0.00	278.62	0.03
	Excluding China	3421	246.12	17.51	-0.08	213.36	-0.01	277.82	-0.01
	Excluding India	3382	248.47	17.62	0.05	214.73	0.07	278.88	0.05
	Excluding Kenya	3388	247.70	17.38	0.01	214.19	0.04	277.58	-0.03
	Excluding Oman	3413	248.43	17.70	0.05	213.47	0.00	278.75	0.04
	Excluding UK	3369	247.30	17.73	-0.02	213.42	0.00	278.58	0.03
	Excluding USA	3707	247.63	17.66	0.00	213.33	-0.01	277.90	-0.01
	Excluding Italy	3494	247.21	18.04	-0.02	213.01	-0.03	277.90	-0.01

GA (weeks)	Country	Sample	C50	SD	C50 SSD	C3	C3 SSD	C97	C97 SSD
29^{+0} to 33^{+6}	Pooled	3993	292.65	15.16	0.00	264.95	0.00	319.95	0.00
	Excluding Brazil	3613	292.86	15.30	0.01	264.85	-0.01	320.44	0.03
	Excluding China	3333	291.83	15.15	-0.05	264.88	0.00	320.38	0.03
	Excluding India	3450	293.51	15.08	0.06	265.56	0.04	320.51	0.04
	Excluding Kenya	3424	292.77	15.03	0.01	265.13	0.01	319.73	-0.01
	Excluding Oman	3448	293.44	15.13	0.05	265.47	0.03	320.53	0.04
	Excluding UK	3403	292.18	15.17	-0.03	264.53	-0.03	319.33	-0.04
	Excluding USA	3728	292.49	15.09	-0.01	264.88	0.00	319.35	-0.04
	Excluding Italy	3552	292.08	15.28	-0.04	264.60	-0.02	319.59	-0.02
34^{+0} to 40^{+0}	Pooled	4092	322.69	13.61	0.00	297.05	0.00	348.67	0.00
	Excluding Brazil	3684	322.61	13.76	-0.01	296.72	-0.02	348.78	0.01
	Excluding China	3513	322.71	13.95	0.00	296.56	-0.04	349.34	0.05
	Excluding India	3583	322.76	13.36	0.01	298.28	0.09	349.28	0.04
	Excluding Kenya	3488	322.72	13.54	0.00	297.25	0.01	348.65	0.00
	Excluding Oman	3491	323.40	13.63	0.05	297.93	0.06	349.33	0.05
	Excluding UK	3380	321.76	13.35	-0.07	296.53	-0.04	347.29	-0.10
	Excluding USA	3837	322.43	13.52	-0.02	297.04	0.00	348.41	-0.02
	Excluding Italy	3668	322.11	13.68	-0.04	296.62	-0.03	348.58	-0.01

Table 3.8: Sample sizes, means (C50), standard deviations (SD), 3^{rd} centiles (C3) and 97^{th} centiles (C97) for HC (mm) for the pooled dataset and for the pooled data set with one site's data excluded. GA, gestational age; SSD, standardised site difference

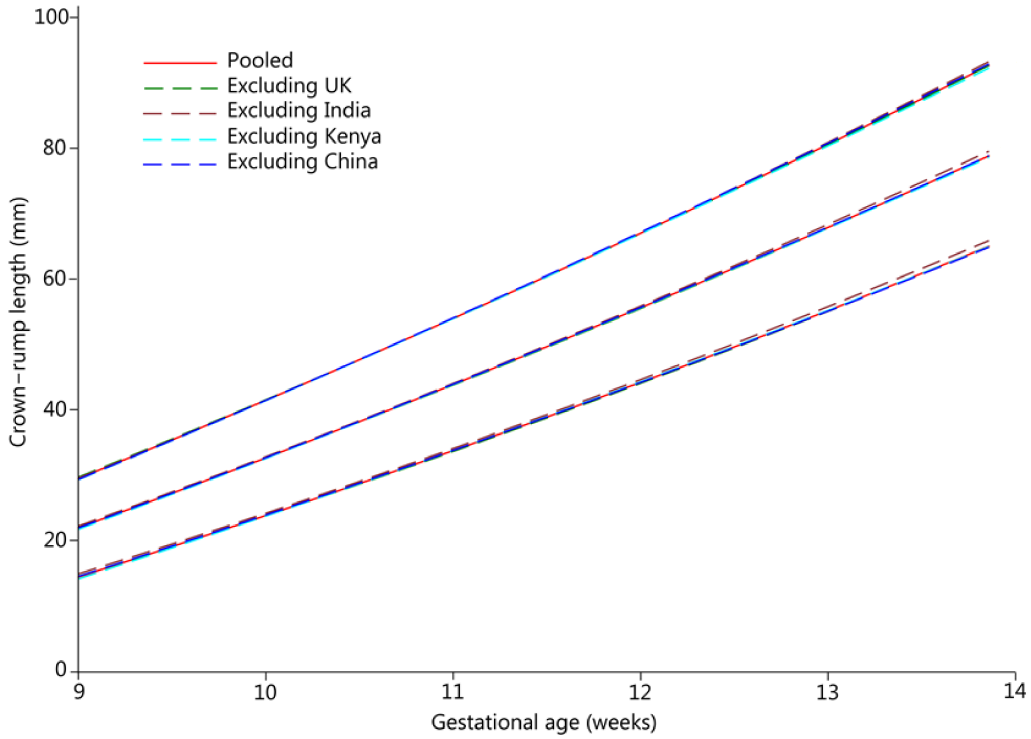


Figure 3.10: Crown-rump length (CRL) at the 3^{rd}, 50^{th} and 97^{th} centiles estimated using fractional polynomial regression models [30] for the pooled Fetal Growth Longitudinal Study population (N = 4,265, solid line) and the effects of excluding, one at a time, the samples from the UK, India, Kenya, and China, whose general populations are usually believed to have very different sizes from one another. The plot demonstrates that removing data from these four different populations had no effect or only a minimal impact on the results from the remaining pooled sample. Exclusion of the other countries did not have an effect on the centiles either as shown in Tables 3.7 and 3.8 (plot not shown).

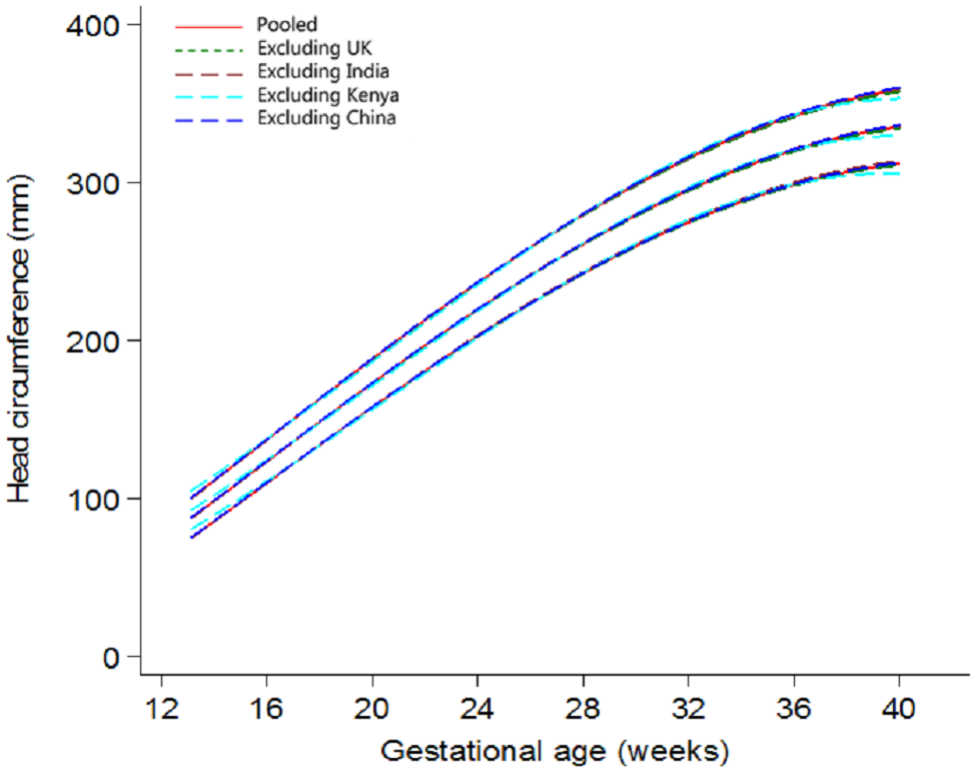


Figure 3.11: Fetal head circumference at the 3^{rd}, 50^{th}, and 97^{th} percentiles estimated using fractional polynomial regression models [30] for the pooled Fetal Growth Longitudinal Study population (N = 4,237, solid line) and the effects of excluding, one at a time, the samples from the UK, India, Kenya, and China, whose general populations are usually believed to have very different sizes from one another. The plot demonstrates that removing data from these four different populations had no effect or only a minimal impact on the results from the remaining pooled sample. Exclusion of the other countries did not have an effect on the centiles either as shown in Tables 3.7 and 3.8 (plot not shown).

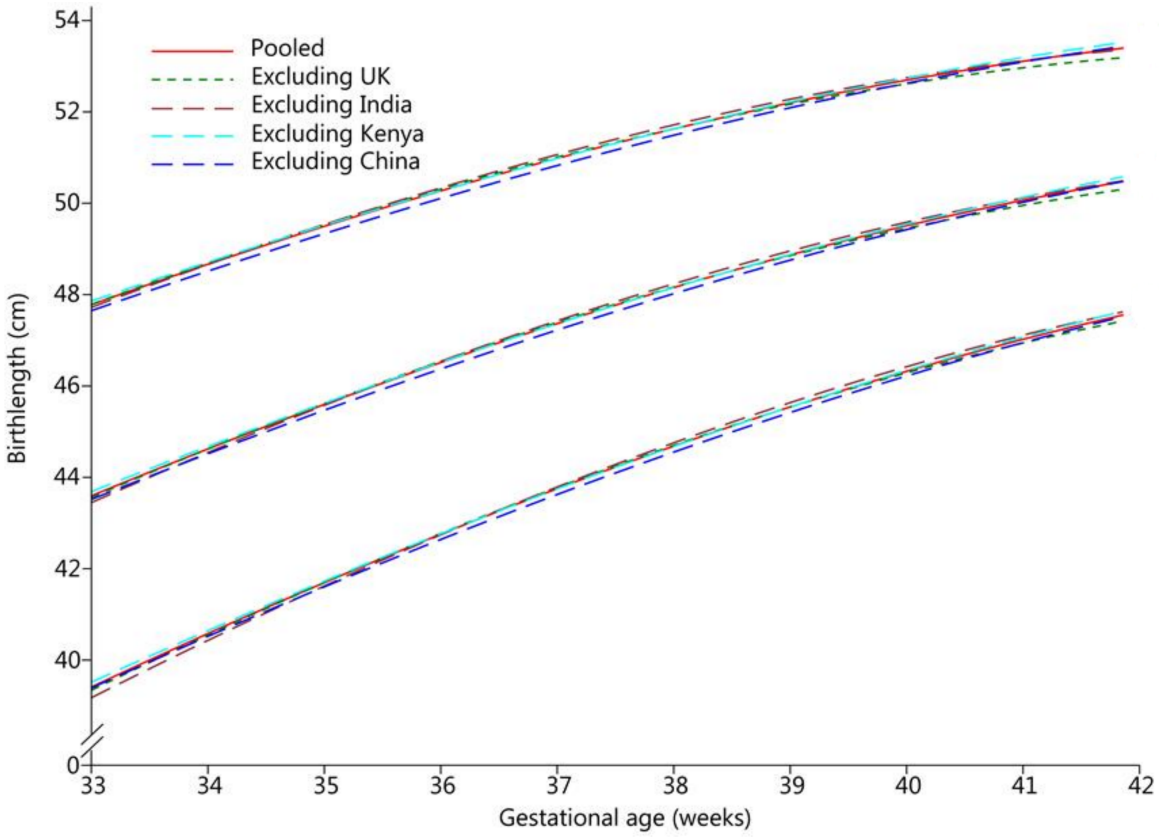


Figure 3.12: Birth length at the 3^{rd}, 50^{th}, and 97^{th} centiles estimated with fractional polynomial regression models for the pooled FGLS-like subpopulation, which represents the low-risk proportion of the total Newborn Cross-sectional Study population (N = 20,166) selected with the same eligibility criteria as the Fetal Growth Longitudinal Study (solid line) and for the pooled sample and the effects of excluding, one at a time, the samples from the UK, India, Kenya, and China, whose general populations are usually believed to have very different sizes from one another. The plot demonstrates that removing data from these four different populations had no effect or only a minimal impact on the results from the remaining pooled sample. Exclusion of the other countries did not have an effect on the centiles either as shown in Tables 3.7 and 3.8 (plot not shown).

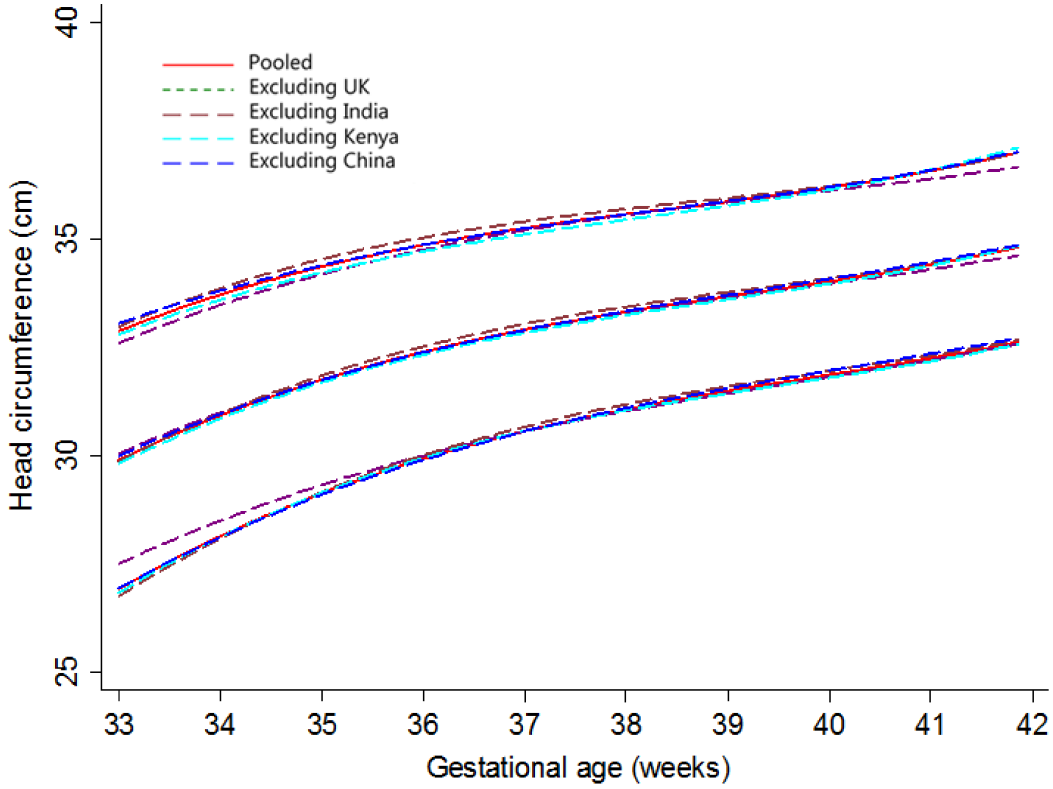


Figure 3.13: Birth head circumference at the 3^{rd}, 50^{th}, and 97^{th} centiles estimated with fractional polynomial regression models for the pooled FGLS-like subpopulation, which represents the low-risk proportion of the total Newborn Cross-sectional Study population (N = 20,046) selected with the same eligibility criteria as the Fetal Growth Longitudinal Study (solid line), and for the pooled sample and the effects of excluding, one at a time, the samples from the UK, India, Kenya, and China, whose general populations are usually believed to have very different sizes from one another. The plot demonstrates that removing data from these four different populations had no effect or only a minimal impact on the results from the remaining pooled sample. Exclusion of the other countries did not have an effect on the centiles either as shown in Tables 3.7 and 3.8 (plot not shown).

3.6 Summary of results

The variance component analysis showed that the percentage between-sites variance of the total variance was 1.9% for CRL, 2.6% for FHC, and 3.5% for BL. The within-site variance for FHC, the measure repeated during pregnancy, was seven times higher than the between-sites variance, at 18.6% (Table 3.1). The data from each site was also modelled by fitting FP regression models. Although the regression models for each site had different regression coefficients, they were of similar functional forms. For CRL, four of the eight models had FP powers = 1 for the mean and SD = 1, two had FP powers = 2 for the mean and SD = 1, one had FP power = 3 for the mean and SD = 1, and one had FP powers = -2, -2 for the mean and SD = -2 (Table 3.2 and Figure 3.3). The overall FP model for CRL using the pooled data had FP powers = 1, 2 for the mean and SD = 1 (Figure 3.4). For HC, five of the eight models had FP powers = 2, 3 for the mean and SD = 1, and the remaining three had FP powers = 2, 2 for the mean and SD = 1. (Table 3.2 and Figure 3.5). These FP powers are known to be so similar that their respective models' fit can be visually indistinguishable [163]. The overall FP model using the pooled data had FP powers = 2, 2 for the mean and SD = 1 (Figure 3.6).

The SSD analysis (Table 3.3) showed that the pooled SD for CRL ranged from 3.91 mm at 9^{+6} weeks to 6.71 mm at 13^{+6} weeks of gestation (Table 3.4). For FHC, the pooled SD ranged from 7.52 mm at 18^{+6} weeks to 11.80 mm at 40^{+0} weeks (Table 3.5). Ten GA windows, from 9^{+0} to 40^{+0} weeks of gestation, were checked, representing 80 comparisons. Of these, 79 comparisons had values <0.5 of the pooled SD. CRL ranged from -0.36 to +0.36, and FHC ranged from -0.58 (the only value outside the defined range) to +0.46 (Figure 3.7). CRL variability was constant over time, but there was some evidence of a wider range of SSD values for FHC with GA, mostly reflecting a smaller pooled SD for FHC rather

than larger differences between sites (Figure 3.7).

Linear size at birth was evaluated across the sites. The pooled SD ranged from 2.3 cm (at 35^{+0} to 35^{+6} weeks of gestation) to 1.7 cm (at 40^{+0} to 40^{+6} weeks of gestation). All of the SSD values for the GA ranges for which an adequate sample size was available were within the prespecified -0.5 to +0.5 intervals (Figure 3.8). Figure 3.8 also shows the average SSD per site across all GAs. The SSD values for BL across sites and GAs were all <0.5 of the pooled SD ranging from -0.33 in India to +0.26 in China. An SSD analysis of the average SSD measures across GA by site was also conducted for BHC for the same fetuses in the FGLS cohort to complement the FHC measures (Figure 3.9). On average, the BHC measures across the eight sites were consistent with the fetal measures, ranging from -0.55 to 0.42 (Figure 3.9).

Separate sensitivity analyses were conducted for CRL and FHC to evaluate the effect on the centiles of the remaining pooled sample of removing a single site's data (Table 3.6). No substantive effects on the remaining pooled sample's 3^{rd}, 50^{th}, and 97^{th} centiles were observed for any of the two primary measures after the data from any of the eight sites were removed from the pooled data (Tables 3.7 and 3.8). Figures 3.10–3.13 show examples of the sensitivity analyses on the centile curves derived for CRL, FHC, BL, and BHC, respectively, showing the effects of excluding the samples from China, India, Kenya, or the UK. The people in the general populations of these four countries are believed to be very different in size. Removing fetuses and newborns from any of these four populations had no or only a minimal effect on the results from the remaining pooled sample.

In summary, the eight sites were compared using SSD analysis during early pregnancy using CRL (40 comparisons), late pregnancy using FHC (40 comparisons), and at birth using BL (48 comparisons). Of these 128 comparisons, only one was marginally greater than the >0.5 cut-off. Across GAs and the sites, the at-birth measure BL had SSD values well below the <0.5 threshold on average. The results of the

SSD, variance components, and sensitivity analyses showed that the eight study populations were sufficiently similar in terms of skeletal size, based on our predefined criteria, for the data to be pooled to estimate centiles for the pooled population. The meta-analytic assessment using FP regression showed great similarity in model fit and functional form for each site's data and therefore supported and complemented the results obtained from the other three methods.

3.7 Discussion

There is increasing evidence that the proportion of human genetic variation due to differences between populations is modest and that the genetic differences between individuals from the same population are much greater than those between individuals from different populations. For example, epidemiological and clinical studies have consistently demonstrated similar growth patterns across ethnic groups in infants and children from relatively affluent, well-nourished backgrounds [164, 165].

Habicht *et al.* [86] first proposed the now well-established idea that infant and child growth are more influenced by health, socioeconomic status, and environmental conditions than by ethnic differences. The strongest scientific evidence in support of this idea to date was provided by the MGRS of healthy, breastfed children with minimal environmental, health and nutritional constraints on growth from six populations in Brazil, Ghana, India, Norway, Oman, and the USA [19, 166]. The study demonstrated striking similarities in the linear growth of the children from the six sites [19]. The data could therefore be pooled to construct a single international standard from birth to 5 years of age, which has since been adopted by more than 140 countries worldwide [18, 167].

The conclusions of two INTERGROWTH-21st systematic recent systematic reviews strongly supported the need for international standards to assess growth patterns in the prenatal and neonatal periods [28, 27]. Creating these standards using the

INTERGROWTH-21st data first required deciding whether the data from the eight centres could be combined. Although it was clearly desirable that this international study yield a single set of growth standards, between-country differences may have been too large to allow such a unified approach. However, there is no standard methodology available to determine combinability. There is also no consensus on how best to compare sets of centiles across a range of GAs.

Four complementary methods were used to compare the data from the different countries: variance component analysis, FP regression, SSD analysis, and sensitivity analysis. The analyses were conducted under the assumption that the comparisons should be based on the distance between centiles and that statistical significance was not relevant to the problem. The criteria used to determine whether the discrepancies between countries were consistently too large for combinability were largely informed by the criteria used by the MGRS in their similar work. The exact nature of the variation between centres was quantified based on statistical considerations but the level of acceptable variability was set using statistical justifications, biological plausibility and clinical judgement. This level had to balance the advantages of a single standard with the consequences of pooling heterogeneous data. Pooling heterogeneous data would result in misclassification of actual growth patterns or potential of specific sites. If true differences did exist for a specific site(s), they would not be detected, resulting in a missed opportunity to investigate further what could otherwise explain those differences. Misclassifications could have serious clinical implications for the affected populations or sites.

Variance component analysis was used to demonstrate that only 1.9-–3.5% of the total variability in fetal skeletal growth and newborn length could be attributed to between-site differences. This was remarkably similar to the 3% variability reported by both the MGRS for infant length [19] and Habicht *et al.* [86] for child height. SSD analysis (equivalent conceptually to a z-score) conducted in 16 GA windows

from 9^{+0} weeks of gestation to birth for CRL, FHC,and BL (128 comparisons) indicated that only one study site mean deviated marginally (0.58 SD) from the corresponding pooled mean of all of the sites at the corresponding GA. A sensitivity analysis that excluded one site at a time from the pooled centiles across GAs for the four markers of skeletal growth and size showed minimal or no effect on the means or the 3^{rd} and 97^{th} centiles of the remaining pooled sample.

The sensitivity analysis results are reassuring because excluding data from single sites had negligible effect on the estimated pooled extreme centiles. There was some variability between the populations, mostly at the extremes of GA in some parameters. This variability may have arisen because of true inter-ethnic differences [19], unstable estimates due to the small sample sizes in some GA windows, or simply differences in protocol implementation, despite best efforts to standardise rigorously across the study sites. However, the variability between the sites represented only 3% of the total variance in skeletal growth, whereas the variability between individuals within a site was seven times higher (Table 3.1). This finding addresses the a *priori* question: whether the variability in the three primary size measures was greater between populations than within populations [11]. This is of tremendous practical importance because the results support the notion that the data were similar enough to be pooled together for the construction of a single international standard. Of potentially greater biological significance, the principal skeletal growth measure for comparing newborns across populations, BL, was very similar at all of the sites.

Further details regarding the baseline characteristics, environmental characteristics and working conditions at each site, supplementation during pregnancy, pregnancy and perinatal events, and indicators of mortality and morbidity for each site have been published elsewhere [12, 168]. The results showed that the eight selected sites were similar at baseline despite their wide geographical diversity, demonstrating adherence to the recruitment protocol and confirming that the

populations considered were at low risk of fetal growth impairment and adverse maternal and perinatal outcomes. The data from the INTERGROWTH-21st Project were therefore pooled and used to construct international standards for fetal growth [24], newborn size [169], and preterm postnatal growth [26], and pregnancy dating chart [21]. The clinical implication for these analyses is that the global standards generated using the INTERGROWTH-21st data will provide a global set of international fetal and newborn standards to allow growth to be monitored from post-conception to childhood.

In summary, the results presented in this chapter have demonstrated that fetal skeletal growth and newborn linear size were strikingly similar in the eight geographically diverse populations studied when mothers' environmental, health and nutritional conditions were met. This conclusion supports the pooling of the data for the construction of international standards. The results are in remarkable agreement with those of the MGRS, and suggest that the differences in fetal growth and newborn size reported in the literature are more likely due to environmental and socioeconomic differences than genetic variation, as has been demonstrated for infants and children. It was important to demonstrate that the data from the eight sites were similar enough to be pooled together to construct a single unified global standard, which was done using four complementary statistical methods. This analysis was a prerequisite for the analyses presented in Chapters 4–6. All of the analyses that follow were performed using the pooled data from all eight sites.

4

Statistical methodology for cross-sectional studies of human growth: Using the INTERGROWTH-21st Project as a case study

4.1 Introduction

The way measurements are collected is important. In fetal growth, for example, when measurements are taken is key as more measurements are generally needed in periods of most rapid growth to accurately capture the pattern of growth. The way measurements are taken also informs the choice of analysis. For example, analysing repeated measures data requires accounting for the non-independence assumption that underpins most statistical methods. Fetal and newborn size charts are mostly based on cross-sectional data [28]: each woman or newborn is only measured

once. This chapter does not provide an exhaustive review of all of the statistical approaches that can be applied to analyses of cross-sectional data. Instead, it aims to give a brief overview of common methodology used for deriving charts of fetal size based on a cross-sectional design. I demonstrate and compare four commonly used statistical methodologies for constructing GA-related size charts from cross-sectional data using the mean and SD [92], LMS method [170], LMST, and LMSP methods [171, 172]. These approaches are compared by evaluating model fits using goodness-of-fit statistics and diagnostic plots. The INTERGROWTH-21st birthweight data for boys and girls from the NCSS cohort is used as an example.

4.2 Aims and considerations

The aim of the chapter is to illustrate alternative modelling approaches. The modelling approaches chosen needed to: (a) develop smooth centiles that offer a good representation of the raw data; (b) model the data precisely, especially the outer centiles (e.g., the 3^{rd} and 97^{th} centiles) where variability is greatest; (c) produce non-crossing centiles; (d) allow estimates of z-scores and centiles to be obtained; (e) apply continuous age smoothing, not age binning; and (f) offer flexibility to account for both skewness and kurtosis when necessary. The methods and modelling approaches are illustrated with newborn weight data from the INTERGROWTH-21st project. To develop centiles that change smoothly with GA using simple statistical models without compromising model fit, I identified modelling approaches that could account for increasing variability with GA, which is a phenomenon often observed in growth data [92]. I evaluated goodness of fit both visually and formally using statistical tests. The analysis did not follow the conventional statistical significance testing approach as no hypotheses were tested; instead the goal was estimation [22].

4.3 Data and methods

4.3.1 Data

The data considered here are from the newborns of women who met strict individual eligibility criteria for a population at low risk of fetal growth impairment from the NCSS component of the INTERGROWTH-21st Project [25]. Anthropometric measurements were taken at birth using an electronic scale (Seca, Germany) to measure birthweight, a specifically designed Harpenden infantometer (Chasmors Ltd, UK) to measure BL and a metallic non-extendable tape (Chasmors Ltd, UK) to measure BHC [173]. The NCSS enrolled 59,137 pregnant women at the eight study sites, of whom 20,486 (34.6% of the NCSS population) met the individual clinical and demographic eligibility criteria for inclusion in the newborn standards, had a reliable ultrasound estimate of GA and delivered a live singleton without a congenital malformation. These newborns constitute the NCSS prescriptive sub-population. To obtain precise estimates at each GA, a threshold of at least 50 observations for each GA was desired to construct the standards. This criterion resulted in 33 weeks as the lower limit and 112 babies were excluded. During data cleaning, we excluded a further 72 babies because they were regarded as implausible within each study site's distribution or were more than 5 SD of the all sites GA-specific mean. The final sample of 20,302 babies was therefore used to construct the standards and is the basis upon which the analyses reported here are based.

4.4 Methodology background

The choice of statistical methods was informed by recommendations by Altman and Chitty, recent literature reviews [92, 105, 67, 174, 175, 147], and the INTERGROWTH-21st systematic review of the methodology used in previous ultrasound studies that created fetal size references [28]. A normality assumption

is the basis of most statistical methods and is thus commonly applied in data analysis. In the current context, the issue is whether the measurements of fetuses or newborns are normally distributed at a specific GA. This assumption underpins the mean and SD method.

When the normality assumption is violated, logarithmic transformation is commonly used due to its desirable mathematical properties of back-transformation to original values [176], ease of fit and variance stabilisation [96]. Logarithmic transformation can be extended to shifted logarithmic transformation of the form $log(y + k)$, although it is rarely used in practice. This simply involves adding or subtracting a constant number, k, based on whether the distribution of the dependent variable is negatively or positively skewed. To obtain estimates in the original scale, the final model is first back-transformed using antilog, then the constant, k, is subtracted. The normal, log-normal, and shifted log-normal distributions are all two-parameter distributions defined by the mean (μ) and variance (σ^2). The normal distribution is denoted by NO(μ, σ^2).

New approaches for fetal and neonatal size reference construction extend these two-parameter models to three- and four-parameter models by exploring more flexible distributions that offer a good representation of the data. In addition to μ and σ, they use shape parameters such as nu (υ) for measuring skewness and tau (τ) for measuring kurtosis. For example, the power exponential distribution [177] is a three-parameter distribution, PE(μ,σ,τ), and is suitable for data with higher kurtosis (leptokurtic) and lower kurtosis (platykurtic) than the normal distribution. The t-family distribution is a three-parameter distribution, TF(μ,σ,τ), and is suitable for modelling leptokurtic data. The Box-Cox Cole and Green distribution (BCCG), sometimes called the Box-Cox normal (BCN) distribution [163], is a three-parameter distribution, BCCG(μ,σ,υ), and is suitable for modelling skewed data. It is defined by a transformed random variable, Z, which is assumed to follow

a normal distribution. It is suitable for positively or negatively skewed data.

The Box-Cox t (BCT) distribution proposed by by Rigby and Stasinopoulos [171] is a four-parameter distribution (μ,σ,υ,τ) denoted by BCT(μ,σ,υ,τ) and is defined through a transformed random variable, Z, that is assumed to follow a truncated t-distribution with degrees of freedom $\tau > 0$, treated as a continuous parameter. The Box Cox Power Exponential (BCPE) model (Rigby and Stasinopoulos, 2004) [172] is a four-parameter distribution (μ,σ,υ,τ) denoted by BCPE(μ,σ,υ,τ) and is defined through a transformed random variable, Z, that is assumed to follow a standard power exponential distribution with power parameter, $\tau > 0$, treated as a continuous parameter. The BCPE distribution is flexible. It simplifies to the normal distribution when $\upsilon = 1$ and $\tau = 2$, and simplifies to the BCCG when $\upsilon \neq 1$ and $\tau = 2$ (the LMS method distribution). The parameters that define a given distribution are the median (μ), coefficient of variation (σ), Box-Cox transformation power (υ), and a parameter related to kurtosis(τ).

4.5 Analytical approaches

There are several different methods available for the construction of age-related centiles, each of them with advantages and limitations [147]. A single method is unlikely to be able to overcome all data modelling challenges associated with these data. Although some methods will apply for most situations, inevitably a number of features unique to different methods will be desirable. In January 2003, the WHO convened a group of statisticians and child growth experts to review available methods for constructing age-related centiles and develop a strategy for assessing their strengths and weaknesses. The group reviewed 30 methods for attained growth curves and agreed on four methods: the mean and SD method (using FPs), LMS method, LMST method, and LMSP method [147]. The INTERGROWTH-21st systematic reviews of the methodological quality of studies designed to create fetal

and neonatal anthropometric charts revealed that these four methods are the most common and widely used for constructing reference charts [28, 16]. Based on the aims and considerations of the preferred modelling methods coupled with findings from the WHO review and our systematic review, I decided to focus on these methods for the analyses of the newborn data. Four approaches were evaluated: the mean and SD, LMS, LMST, and LMSP methods.

4.5.1 Mean and SD method

The mean and SD method is the most common parametric approach. It is based on the assumption that for each GA, the measurement of interest has a normal distribution with a mean and SD that vary smoothly with GA. A desired centile curve is calculated using:

$$C_\alpha = \mu + K \times SD, \tag{4.1}$$

where K is the normal equivalent deviate (z score) corresponding to a particular centile, e.g., K = 1.88 for the 97^{th} centile and -1.88 for the 3^{rd} centile, and μ and SD are the mean and standard deviation, respectively at the required GA for the reference population. The mean and SD method is based on least squares regression analysis, in which the mean and SD centile curves are modelled as polynomial functions of GA. The mean and SD method fits separate models for the mean and SD to account for the increasing variability with GA that is typical of fetal and newborn data. Either conventional polynomials or FPs can be used. The method requires the assumption of a normal distribution, as when using conventional polynomials.

I used Royston and Altman's [163] FPs due to their great flexibility in allowing non-integer powers, logarithms, and repetition of powers [163, 94], and because they have been shown to fit fetal data very well [92, 67, 163, 104, 178, 179, 36, 78]. FPs are defined by power terms restricted to a predefined set of integer and non-integer

values, p (-2, -1, -0.5, 0, 0.5, 1, 2, 3). The best power transformation, x^p, is chosen from the set of powers, p, with software. An automated algorithm for selecting these powers has already been implemented in the statistical software programs STATA and R. The term x^0 denotes *log(x)* rather than $x^0 = 1$. The degree of an FP model, m, is defined as the number of powers, p, of the explanatory variable. For instance, a first-degree FP (FP1) model with power p will be of the form Y $= b_0 + b_1 x^{p_1}$ and an FP2 of the form Y $= b_0 + b_1 x^{p_1} + b_2 x^{p_2}$. FP1, FP2, and FP3 models relating newborn anthropometric measures to GA were explored.

4.5.2 LMS method

Van't Hof, Wit, and Roode [180] first suggested a method to deal with non-normal anthropometry data. Using skewed skinfold data as an example, they suggested a power transform [181] at each age to remove skewness, making the data approximately normally distributed. The proposed method consists of seven steps, which allows the power transform to change smoothly with age and to vary from one age to another. Cole [182, 183] generalised this method using three parameters λ, μ, and σ, the initials of which are respectively L, M, and S, giving rise to the name LMS method. The LMS method assumes that the given power of a biometric or anthropometric trait at a given age follows a normal distribution and thus that the data can be summarised by three age-dependent functions. M(t) and S(t) represent the median and coefficient of variation (SD/median), respectively, of each biometric trait at each age. L(t) represents the value of the power needed to normalise the data at each age.

The three curves, L(t), M(t), and S(t) are estimated as cubic splines by non-linear regression and by maximising the penalised likelihood. Three smoothing parameters for the three curves are thus obtained (equivalent degrees of freedom (edf)) from each fitted curve as a function of the smoothing parameters. The edf of each L, M,

and S curve is a measure of complexity. For example, edf = 1 indicates a constant, edf = 2 refers to a straight line, edf = 3 is a quadratic curve, and edf $\geq$ 4 refers to more complex curve shapes. The choice of edf is somewhat subjective and is an indication of how well the data has been smoothed. Low and high edfs correspond to over- and under-smoothed curves, respectively. It is desirable to balance between model complexity (in terms of smoothing) and model fit to the raw data. The three curves together allow any centile to be calculated:

$$C_{100\alpha}(t) = M(t)(1 + L(t)S(t)Z_\alpha)^{(1/L(t))} \qquad L(t) \neq 0, \tag{4.2}$$

or

$$C_{100\alpha}(t) = M(t)exp[S(t)Z_\alpha] \qquad L(t) = 0, \tag{4.3}$$

where $C_{100\alpha}(t)$ is the expected value of a given centile 100α of a measured anthropometry at a given age, M(t) is the median, S(t) is the coefficient of variation, L(t) is the power transform and Z_α is the normal equivalent deviate of size α (SD score or z-score). SD scores (SDS) are recommended for making direct comparisons between different anthropometric measures and can also be used to compare different populations [170, 182]. The SDS values for an individual can similarly be obtained:

$$SDS = \frac{[y(t)/M(t)]^{L(t)} - 1}{L(t)S(t)} \qquad L(t) \neq 0, \tag{4.4}$$

or

$$SDS = \frac{log[(y(t)/M(t)]}{S(t)} \qquad L(t) = 0, \tag{4.5}$$

where y(t) is the measured anthropometry for the child at a given age t, M(t) is the median, S(t) is the coefficient of variation, and L(t) is the power transform at that age [170].

4.5.3 LMS extensions: the LMST and LMSP methods

The LMST method proposed by Rigby and Stasinopolous [171] is an extension of the LMS method that models both skewness and kurtosis larger than 3. It can be used to model excess kurtosis over the normal distribution (leptokurtic data) when the Box-Cox transformation fails to transform the data close to normality due to the presence of kurtosis.

The LMSP method proposed by Rigby and Stasinopolous [172] is also an extension of the LMS method that models both skewness and kurtosis. Unlike the LMST method, which can only model leptokurtic data, the LMSP method can model any type of kurtosis, i.e., leptokutosis, platykutosis, or mesokurtosis. The LMSP achieve this greater utility by using the more flexible BCPE distribution.

4.6 Model selection

Seven candidate models were created with the mean and SD method, one two-parameter model, two three-parameter models, and four four-parameter models. One candidate three-parameter model was created with the LMS method. Two candidate four-parameter models were created with the LMS extension methods, one LMST and one LMSP. The best model *within* each class of models was identified, i.e., the best two-, three-, and four-parameter models were each identified. Using this information, the single best model *across* the classes created by a particular approach (either the mean and SD method or the LMS method and its extensions) was identified in an add-up stepwise fashion, starting from the simplest class. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to compare the fit of the models within and across classes [184]. Model choice was not based on AIC and BIC alone as other criteria were also considered. For example, aesthetic appeal of the fitted models to the raw data was also considered. In

addition, simpler models that showed equally good model fits when compared to more complex models that resulted in small differences in AIC and BIC were preferred. Desirability of having models with the same functional form for the males and females data was also considered. Models that fitted well across the entire GA range were deemed to be better that models with a smaller AIC or BIC but showed inadequate fit or unexpected shifts in centiles especially at either ends of the distribution.

4.7 Diagnostics

Goodness of fit of the resultant model was assessed by comparing residuals (observed values minus fitted values) according to GA. Formal statistical testing such as the AIC, BIC, and Q-statistic, was also conducted and considered when deciding whether to select a more complex model. Overall model fit was visually evaluated using quantile-quantile plots of the residuals, which can reveal any departures from normality; plots of residual vs. fitted values; and the distribution of fitted z-scores across GAs. The worm plot introduced by van Buuren *et al.* [185] can either consist of a collection of detrended quantile-quantile plots, each of which applies to a GA group, or a single worm plot, representing the entire GA range. Residuals were calculated by subtracting each empirical quantile from its corresponding unit normal quantile. They were calculated according to GA to identify regions or intervals of GA within which the model did not fit the data adequately, in a process called model violation. This is a diagnostic tool for checking the residuals for different ranges non-overlapping ranges of the explanatory variable. A model that fits the data well should resemble a flat worm-like string. Any sudden changes in the shape and location of the worm represent regions where the data has been inadequately modelled.

Royston and Wright [186] proposed a goodness-of-fit test (Q-test) based on the distribution of fitted z-scores by testing the four moments $(\mu, \sigma, \upsilon, \tau)$ for normality across age using the modified D'Agostino [187] and Shapiro-Wilk tests. The

Q-test was used to find moments in the distributions that were poorly modelled. The Q-statistic from a particular age range was also used to indicate whether the corresponding moment was adequately modelled.

The Q-test combined with the worm plot patterns provided a robust assessment of each model's goodness-of-fit, especially in terms of evaluating local fit. Group-specific Q-test statistics resulting in absolute values of z1, z2, z3, or z4 that were larger than 2 were interpreted to indicate a misfit of mean variance, skewness, or kurtosis, respectively. The overall Q-test statistics combining all of the groups were based on a Chi-square distribution, which assumes that observations from different groups are independent.

4.8 Implementation

The models were fitted using the generalised additive models for location, scale and shape (GAMLSS) framework [188], available in the statistical software package R [189]. The GAMLSS model allows the distribution parameters μ, σ, ν and τ to be modelled as linear, non-linear parametric and non-parametric (smooth) functions of GA. The GAMLSS package provides a comprehensive framework with great flexibility and options for using different methodologies (e.g., the mean and SD method using FPs and the LMS method), distributions (e.g., the skew t-distribution type 3 and power exponential distribution), smoothing techniques (e.g., penalised splines and cubic splines), and diagnostics (e.g., worm plots and Q-statistics).

4.9 Results

Table 4.1 summarises the number of birthweight measurements according to GA for all of the newborns, divided into boys and girls. Scatter plots of the raw newborn size measurements of birthweight by GA for boys and girls are shown in Figure 4.1. The distributions of measurements were fairly similar for boys and

girls across GA, except at 33 weeks (Table 4.1, and Figure 4.1). The newborn data were close to being conditionally normal ('well-behaved') on GA and thus the different methods gave similar results, as Table 4.2 shows. Higher degree FPs, such as FP3, were not required. Boys and girls were analysed separately. Figure 4.2 summarises the analytical methods, associated distributions, smoothing techniques, and diagnostic tests used to evaluate model fit.

Table 4.2 shows the 20 models tested from the four methodological approaches and how well each model fitted the newborn data. The fitted centiles and corresponding goodness-of-fit plots for the 20 models are shown in Figures 4.3a–4.12b. These figures show similar plots for each model. Figure 4.3a (male birthweight) is described as a representative example.

The top left panel shows a simple FP fit of a two-parameter model, assuming a normal distribution (two powers for the mean and one for the SD) for male birthweight (mean and SD method). This plot is useful for demonstrating how the distribution of birthweight changes according to GA in two-week intervals. It is an informative way of assessing the distribution of the data by GA which can be useful for judging the modelling method choice. Patterns of non-normal data by GA can easily be detected and corrected for, as appropriate. In this plot, the data seem to be reasonably normally distributed by GA. Simple models based on the normality assumption should therefore perform well for these data.

The top middle panel shows the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to GA. The top middle panel plot provides a visual assessment of whether the smooth centiles offer a good representation of the raw data. In this plot, one can see a tendency for slight deviations of model fit for the 97^{th} centile to areas with no data at lower GA, < 35 weeks, (overestimation). Similarly, for higher GAs, > 41 weeks, predictions at the 97^{th} centile seem to have been underestimated.

The top right panel shows a worm plot. The worm plot is useful for identifying

regions or intervals of GA within which the model did not fit the data adequately. The vertical axis of the worm plot portrays the difference between each observation's location in the theoretical and empirical distributions. The two elliptic curves represent 'acceptance region' for a well-fitting model if the worm lies inside the two elliptic curves. The red curve in the plot is a penalised spline polynomial fitted to the points on the plot. The shape of the worm indicates how the data differ from the assumed underlying distribution and suggests useful modifications to the model. A flat worm indicates that the data follow the assumed distribution in that age group. Any sudden changes in the shape and location of the worm represent regions where the data were inadequately modelled. In this plot, the worm plot is flat for most of the middle age range, but changes shape and deviates from the expected zero line at lower and upper GAs. As some observations fall outside the two elliptic curves, the overall model appears not to have fitted well. This is a clear indication of inadequate model fit for male birthweight data at the lower and upper ranges of GA.

The bottom left panel shows a scatter plot of the residuals according to GA. The plot shows if and how the variability changes with gestation and is useful for checking whether variability of birthweight with GA (typical of newborn data) has been accounted for. The plot can be used to check for unexpected patterns and whether the expected proportion of values falls between or outside the expected z-scores (for example, 95% of values should lie within ± 2 z-scores, 2.5% below, and 2.5% above). In this plot, the variability seems reasonably constant according to GA except for earlier GA with a paucity of data.

The bottom middle plot shows a normal Q-Q plot of z-scores. It evaluates whether the residuals have a close-to-normal distribution. This is signified by a straight line cutting through the plot at 45° degrees. In this plot, similar to deductions made from the worm plot, there is deviation from a normal distribution at the bottom and top ends of the distribution representing the lower and upper ranges of GA.

The Q-statistic plot for specified GA ranges is shown in the bottom right panel. This plot is useful for testing normality and is based on the distribution of fitted z-scores by testing the four moments (μ,σ,υ, and τ) across GA. The Q-statistic from a particular GA range indicates whether the corresponding moment (μ,σ,υ,τ) was adequately modelled. The plot shows group-specific Q-test statistics resulting in absolute values of Q1, Q2, Q3, and Q4 (representing μ,σ,υ, and τ). Absolute values for each of the moments are represented by circles. The bigger the absolute values, the bigger the circle, and vice versa. Absolute values that are greater than 2 are shown in big red circles with a square inside the circle, while those less than 2 are shown in smaller blue circles. Any absolute values of Q1, Q2, Q3, and Q4 that are larger than 2 (big red circles) indicate a misfit of μ,σ,υ, and τ respectively: the residuals have a higher (or lower) μ,σ,υ, or τ than the null standard normal distribution. In this plot, the big red circles in Q3 (the third moment, skewness) and Q4 (the fourth moment, kurtosis) show that the data were skewed and kurtotic for these respective gestational ranges (38 to 41 weeks) and that these moments were not modelled adequately. The remaining two moments, mean (Q1) and variance (Q2), appear to have been modelled adequately, as they are depicted by very small blue circles. Similar observations can be made about the model fit for girls in Figure 4.3b.

For boys and girls, a three-parameter model fit assuming a power exponential model offered an improved fit over the two-parameter model, based on the Q-Q plot in Figures 4.4a and 4.4b. The Q-statistic plots show that skewness was still inadequately modelled. Fitting a three-parameter BCCG distribution, as expected, modelled the skewness appropriately, but not kurtosis (Figures 4.5a and 4.5b).

The four-parameter models, created with the BCT, skew exponential power type 3, skew t-distribution type 3, and BCPE distribution, offered significantly improved fits to the data. The BCT, skew exponential power type 3, and BCPE distributions gave very similar results. In Figures 4.6a–4.9b, the worm plots are flatter when

compared to the worm plots in Figures 4.4a–4.5b, and their Q-Q plots follow a straight line at 45° degrees on the expected line. The Q-statistic plots suggest a misfit for certain gestational ranges and the fitted 3^{rd} and 97^{th} centiles show a tendency to drift away from the data at lower GA (Figures 4.6a–4.9b).

Figures 4.10a and 4.10b show model fits for the LMS method. Fitting a BCCG distribution for male birthweight data failed to converge and therefore a normal distribution was fitted instead. The BCCG distribution was successfully fitted to the female birthweight data. The figures show that the model fit, as the fitted centiles are not smooth for lower GA (boys), the worm plots are not flat, the Q-Q plots show deviations from the expected line, and the Q-statistic plot show that skewness and kurtosis were inadequately modelled.

The LMST and LMSP methods greatly improved on the LMS method. However, in Figures 4.11a–4.12b, the model fits drift away from the data for lower GA, worm plots are not flat, and Q-statistic plots show inadequate fit for some moments at certain gestational ranges.

Gestational age (completed weeks)	Birthweight	
	Boys	Girls
	Number of observations	Number of observations
33	34	17
34	48	65
35	128	114
36	323	293
37	857	803
38	2,045	1,802
39	3,009	2,869
40	2,568	2,523
41	1,179	1,195
42	206	224
Total	10,397	9,905

Table 4.1: Number of birthweight measurements according to gestational age for boys and girls.

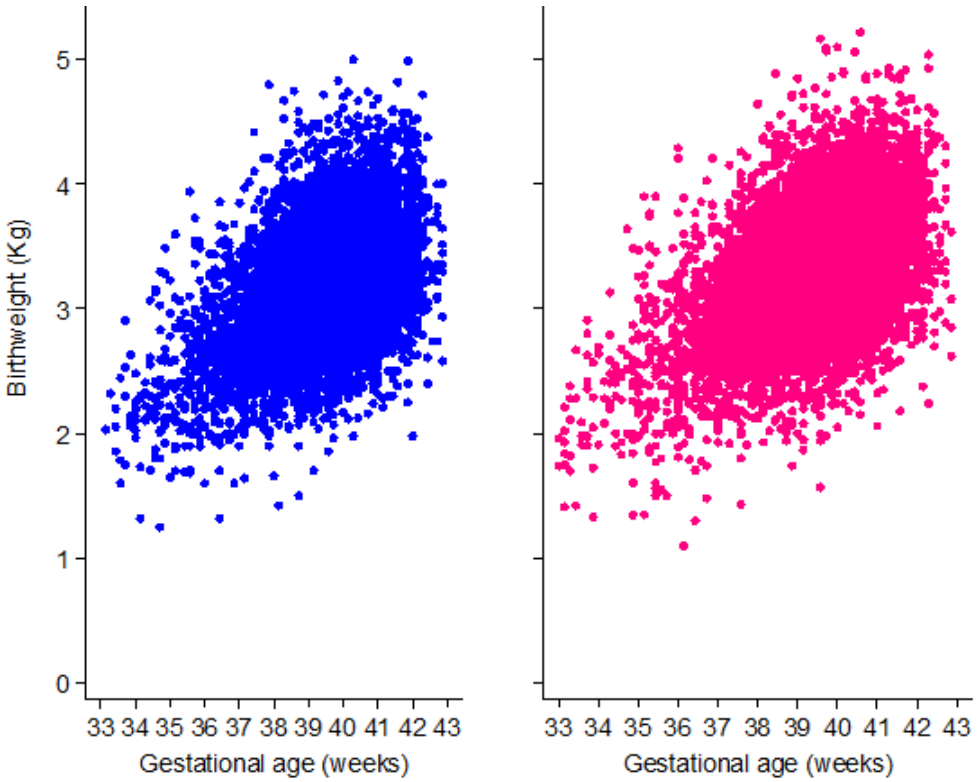


Figure 4.1: Scatter plot of birthweight measurements by gestational age for boys (left, blue) and girls (right, pink).

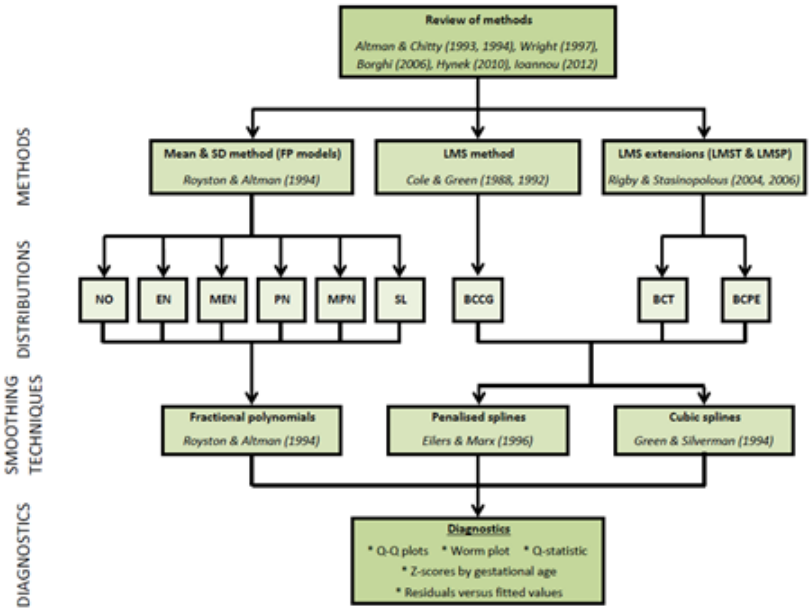


Figure 4.2: Summary of the methodological approaches tested.

BCCG, Box-Cox Cole and Green distribution; BCPE, Box-Cox Power Exponential distribution; BCT, Box-Cox t-distribution; EN, exponential normal distribution; MEN, modulus exponential normal distribution; MPN, modulus power normal distribution; SL, shifted (or three-parameter) lognormal distribution; NO, normal distribution; PN, power normal (or Box-Cox) distribution.

4.10 Results of the mean and standard deviation method

The mean and SD were modelled separately. Table 4.2 shows the seven models that were fitted and how well each model fit the newborn data using the mean and SD approach. The fitted centiles and corresponding goodness-of-fit plots for the seven models are shown in Figures 4.3a–4.12b. The best FP mean and SD model used an FP2 to model the mean and a linear FP1 to model the SD, skewness, or kurtosis as appropriate. For each model, the plots show (a) how the distribution of

birthweight changed according to GA, (b) the fitted 3^{rd}, 50^{th} and 97^{th} smoothed centiles across GA, (c) a worm plot, (d) a scatter plot of the residuals by GA, (e) normal Q-Q plots of the z scores and (f) the Q-statistic for specified GA ranges. As already described, the overall model fit was evaluated using these plots.

For boys and girls, the lowest AIC, BIC, and global deviance for the mean and SD models (M1 – M7) were those based on the skew exponential power type 3 (model M5) and skew t-distribution type 3 (model M6) distributions (Table 4.2). For boys, based on the AIC, BIC, and global deviances, the three-parameter model based on the power exponential distribution (model M2) offered a significantly better fit (global deviance = 11,475.9) than the two-parameter model based on a normal distribution (model M1, global deviance = 11,530.6).

For girls, the three-parameter models based on power exponential (model M2, global deviance = 10,106.6) and on the BCCG distribution (model M3, global deviance = 10,105.9) fit significantly better than the two-parameter model (model M1, global deviance = 10,167.5). Unsurprisingly, the four-parameter models performed better than the two- and three-parameter models. There were only small differences between the four-parameter models, with deviances ranging from 11,428 for the skew t-distribution type 3 model (model M6) to 11,444.6 for the BCPE model (model M7) (Table 4.2).

The skew t-distribution type 3 distribution provided the best fit for boys and girls, judging by the fitted smoothed centiles, which offered a good representation of the data); the completely flat worm plots; the observations all falling within the the two elliptic curves indicating the 'acceptable' regions; the Q-Q plots following a straight line at 45° degrees on the expected line; and the adequate modelling of most of the moments, outside certain gestational ranges.

Model choice was not only based on models with the lowest AIC and BIC values. Other considerations such as having one functional form (distribution) of the

FP model for birthweight data for boys and girls (though modelled separately) were considered.

Mean and SD method: Two-parameter models

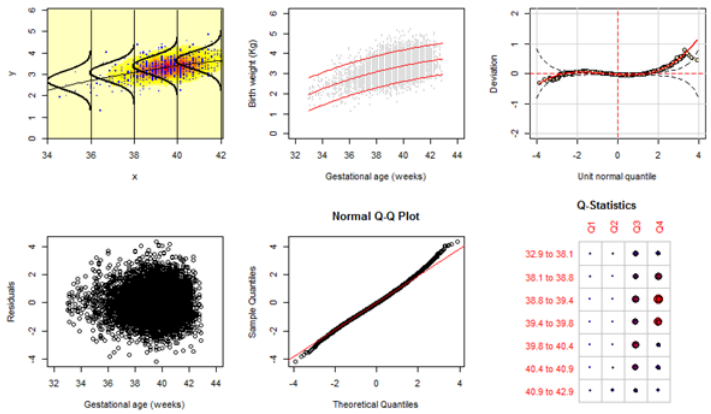


Figure 4.3a: The mean and SD method: Fractional polynomial fit of a two-parameter model assuming a normal distribution (two powers for the mean and one for the SD) for male birthweight (Model: M1_B, Table 4.2).

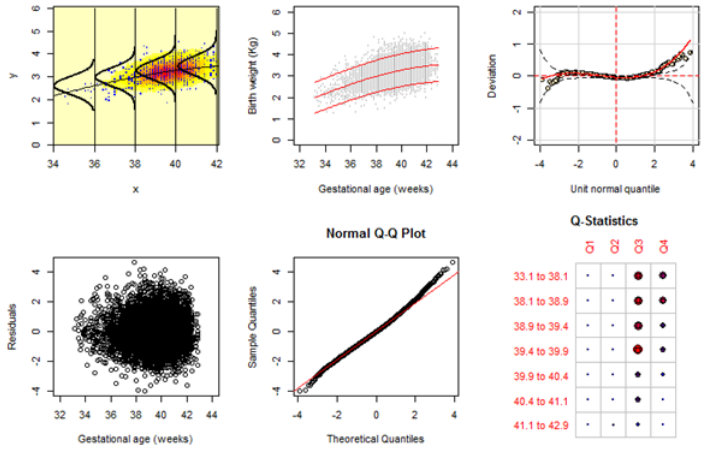


Figure 4.3b: The mean and SD method: Fractional polynomial fit of a two-parameter model assuming a normal distribution (two powers for the mean and one for the SD) for female birthweight (Model: M1_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel).

Mean and SD method: Three-parameter models

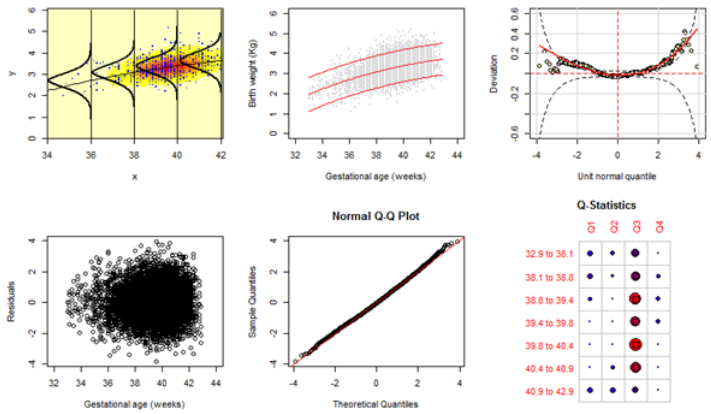


Figure 4.4a: The mean and SD method: Fractional polynomial fit of a three-parameter model assuming a power exponential distribution (two powers for the mean, one for the SD, and one for skewness) for male birthweight (Model: M2_B, Table 4.2).

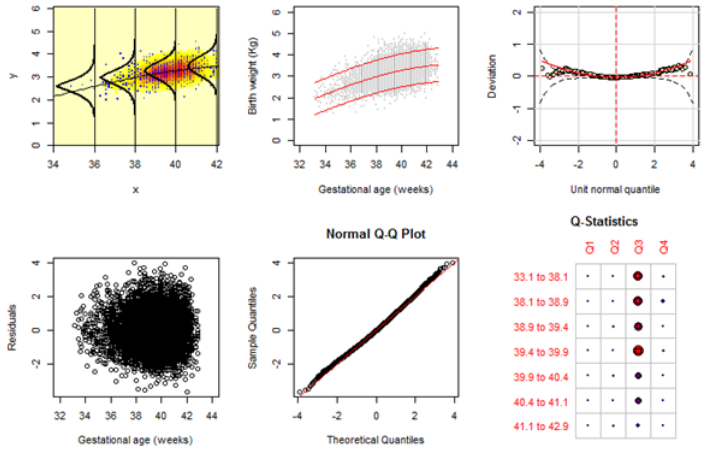


Figure 4.4b: The mean and SD method: Fractional polynomial fit of a three-parameter model assuming a power exponential distribution (two powers for the mean, one for the SD, and one for skewness) for female birthweight (Model: M2_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

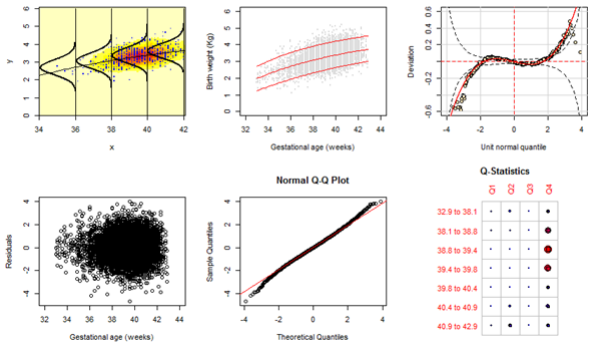


Figure 4.5a: The mean and SD method: Fractional polynomial fit of a three-parameter model assuming a Box-Cox Cole and Green distribution (two powers for the mean, one for the SD, and one for skewness) for male birthweight (Model: M3_B, Table 4.2).

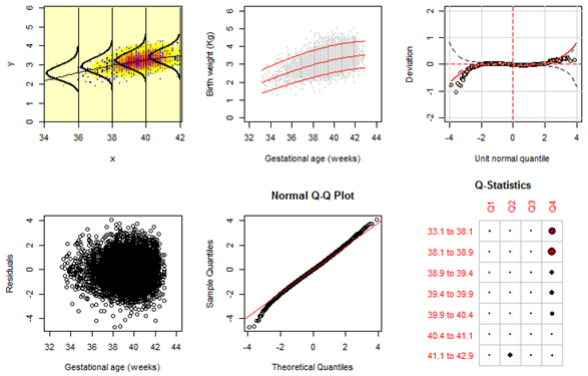


Figure 4.5b: The mean and SD method: Fractional polynomial fit of a three-parameter model assuming a Box-Cox Cole and Green distribution (two powers for the mean, one for the SD, and one for skewness) for female birthweight (Model: M3_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

Mean and SD method: Four-parameter models

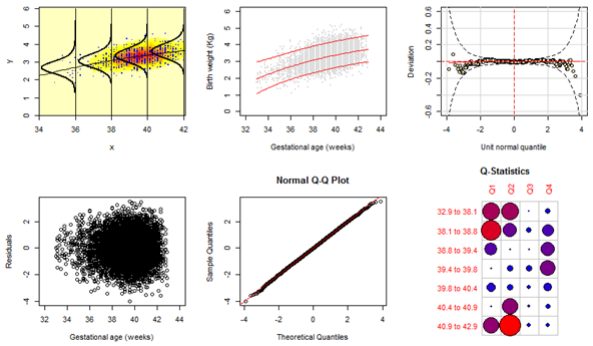


Figure 4.6a: The mean and SD method: Fractional polynomial fit of a four-parameter model assuming a Box-Cox t-distribution (BCT) (two powers for the mean, one for the SD, one for skewness and one for kurtosis) for male birthweight (Model: M4_B, Table 4.2).

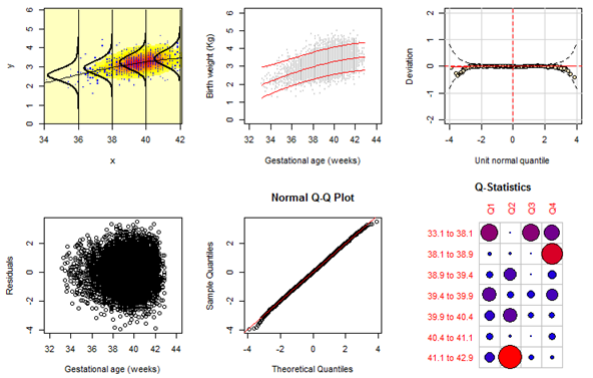


Figure 4.6b: The mean and SD method: Fractional polynomial fit of a four-parameter model assuming a Box-Cox t-distribution (two powers for the mean, one for the SD, one for skewness and one for kurtosis) for female birthweight (Model: M4_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

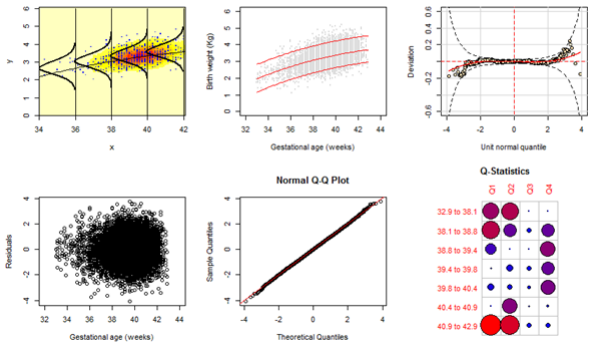


Figure 4.7a: The mean and SD method: Fractional polynomial fit of a four-parameter model assuming a skew exponential power type 3 (two powers for the mean, one for the SD, one for skewness and one for kurtosis) for male birthweight (Model: M5_B, Table 4.2).

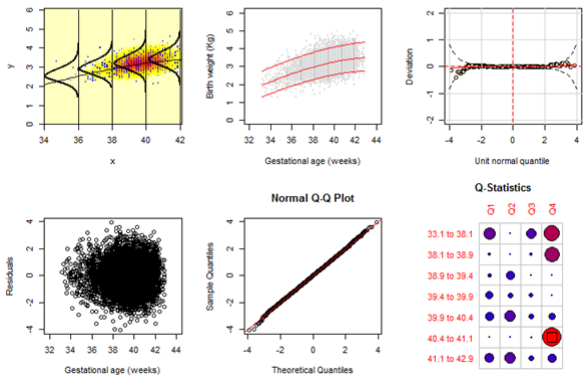


Figure 4.7b: The mean and SD method: Fractional polynomial fit of a four-parameter model assuming a skew exponential power type 3 (two powers for the mean, one for the SD, one for skewness and one for kurtosis) for female birthweight (Model: M5_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

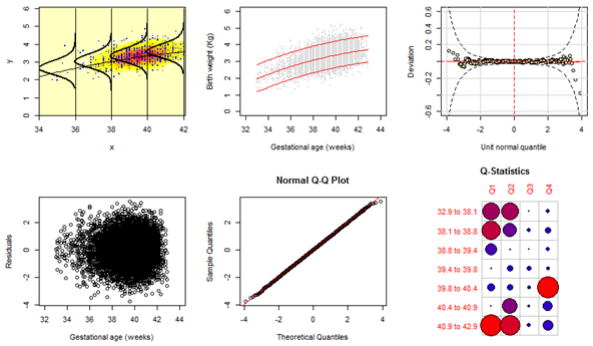


Figure 4.8a: The mean and SD method: Fractional polynomial fit of a four-parameter model assuming a skew t-distribution type 3 distribution (two powers for the mean, one for the SD, one for skewness and one for kurtosis) for male birthweight (Model: M6_B, Table 4.2).

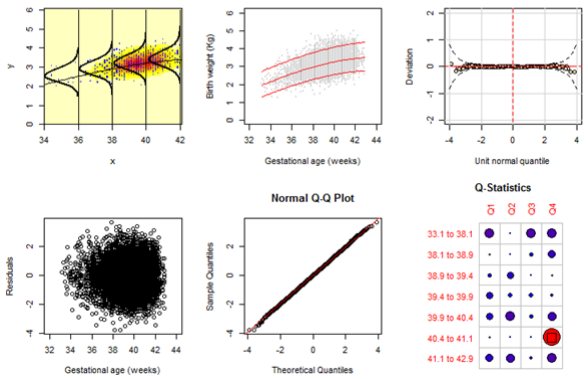


Figure 4.8b: The mean and SD method: Fractional polynomial fit of a four-parameter model assuming a skew t-distribution type 3 distribution (two powers for the mean, one for the SD, one for skewness and one for kurtosis) for female birthweight (Model: M6_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

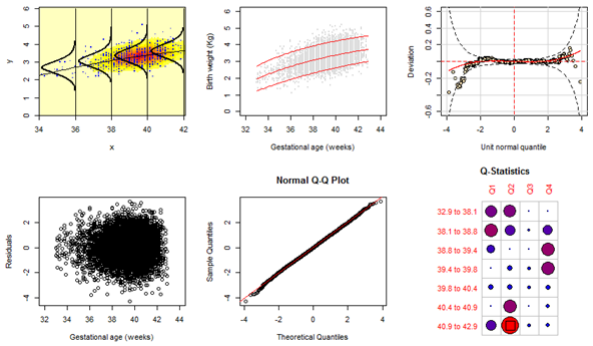


Figure 4.9a: The mean and SD method: Fractional polynomial fit of a four-parameter model assuming a Box-Cox power exponential distribution (two powers for the mean, one for the SD, one for skewness and one for kurtosis) for male birthweight (Model: M7_B, Table 4.2).

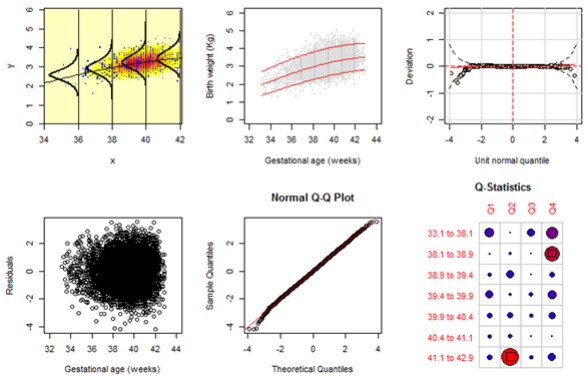


Figure 4.9b: The mean and SD method: Fractional polynomial fit of a four-parameter model assuming a Box-Cox power exponential distribution (two powers for the mean, one for the SD, one for skewness and one for kurtosis) for female birthweight (Model: M7_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

4.11 Results of the LMS, LMST, and LMSP methods

In the LMS method and its extensions, a Box-Cox power transformation of GA was required to remove skewness and normalise the data for the construction of the

centile curves. GA was therefore modelled as $GA^{3/2}$ rather than GA. After this transformation, skewness and kurtosis did not seem to vary by GA and were thus modelled as a constant. Of the three methods, the LMST method provided the best fit for boys (model M9) with a global deviance of 11,419.1, compared with 11,495.7 for the LMS method and 11,426.9 for the LMSP method. The LMSP method provided the best fit for girls (model M10), with a lower global deviance than the LMS and LMST methods. For boys, the LMS method did not converge under a BCCG distribution and thus a normal distribution was fitted instead (Table 4.2).

LMS method

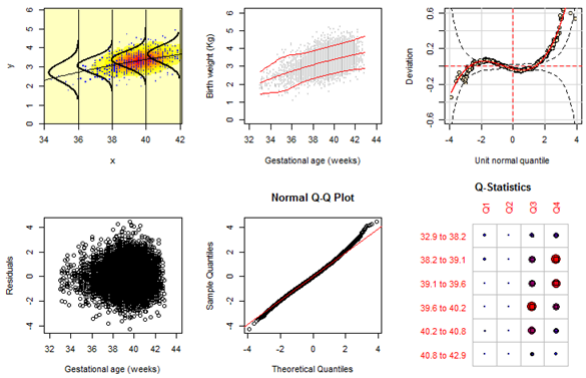


Figure 4.10a: The LMS method: Based on a two-parameter model assuming a normal distribution after the Box-Cox Cole and Green distribution failed. The model was fitted on log(n) where n = sample size for male birthweight (n = 10,320) with a transformation for gestational age = 1.5 (Model: M8_B, Table 4.2).

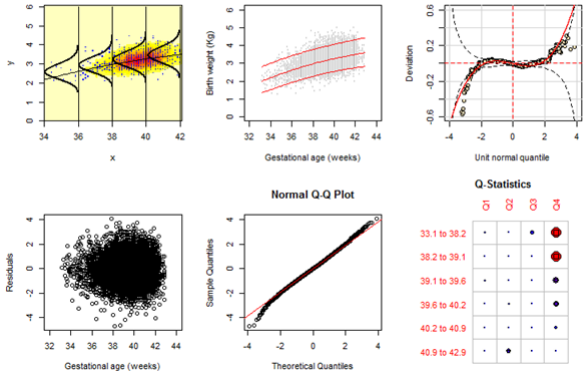


Figure 4.10b: The LMS method: Based on a three-parameter model assuming a Box-Cox Cole and Green distribution. The model was fitted on log(n) where n = sample size for female birthweight (n = 9,825) with a transformation for gestational age = 1.5 (Model: M8_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

LMST method

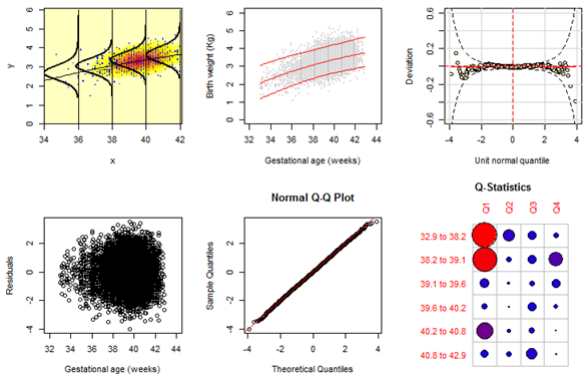


Figure 4.11a: The LMST method: Based on a four-parameter model assuming a Box-Cox-t distribution. The model was fitted on log(n) where n = sample size for male birthweight (n = 10,320) with a transformation for gestational age = 1.5 (Model: M9_B, Table 4.2).

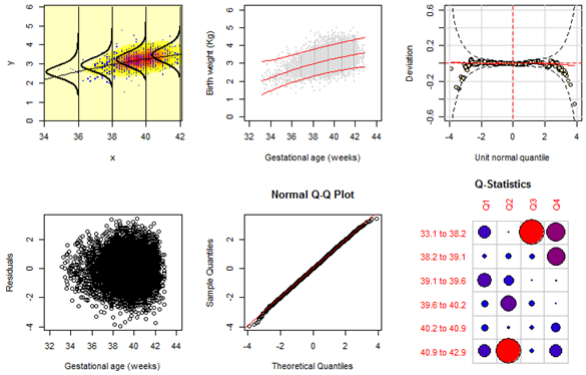


Figure 4.11b: The LMST method: Based on a four-parameter model assuming a Box-Cox-t distribution. The model was fitted on log(n) where n = sample size for female birthweight (n = 9,825) with a transformation for gestational age = 1.5 (Model: M9_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z-scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

LMSP method

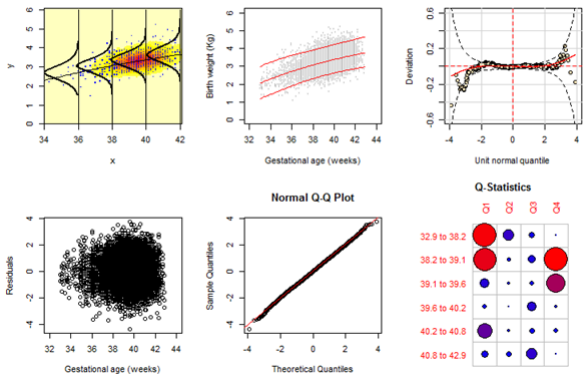


Figure 4.12a: The LMSP method: Based on a four-parameter model assuming a Box-Cox power exponential distribution. The model was fitted on log(n) where n = sample size for male birthweight (n = 10,320) with a transformation for gestational age = 1.5 (Model: M10_B, Table 4.2).

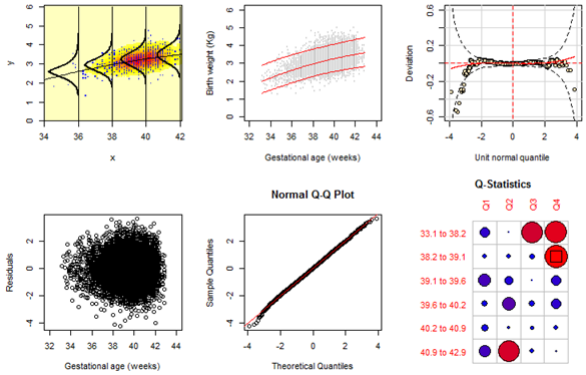


Figure 4.12b: The LMSP method: Based on a four-parameter model assuming a Box-Cox power exponential distribution. The model was fitted on log(n) where n = sample size for female birthweight (n = 9,825) with a transformation for gestational age = 1.5 (Model: M10_G, Table 4.2).

The plots show: (a) the distribution of birthweight according to gestational age (top left panel), (b) the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles according to gestational age (top middle panel), (c) a worm plot (top right panel), (d) a scatter plot of the residuals according to gestational age (bottom left panel), (e) normal Q-Q plots of the distribution of z scores (bottom middle panel), and (f) the Q-statistic for specified gestational age ranges (bottom right panel)

		Mean and standard deviation method (fractional polynomials)								LMS method	LMST method	LMSP method
		Parameters	2-parameters	3-parameters		4-parameters				3-parameters	4-parameters	
Measure	Sex	Distribution	NO	PE	BCCG	BCT	SEP3	ST3	BCPE	LMS(NO)	LMS(BCT)	LMS(BCPE)
Birthweight	Male	Model name	M1_B	M2_B	M3_B	M4_B	M5_B	M6_B	M7_B	M8_B	M9_B	M10_B
		Parameters	2	3	3	4	4	4	4			
		Mean	-1, -0.5	-0.5, 0	-2, -2	-2, -2	0.5, 1	0, 0.5	-2, -2	μ.df = 2.21	μ.df = 2.33	μ.df = 2.32
		SD	-2	-2	-2	-2	-2	-2	-2	σ.df = 6.67	σ.df = 3.04	σ.df = 3.10
		Nu	NA	3	-2	-2	3	3	-2	NA	ν.df = 2.00	ν.df = 2.00
		τ	NA	NA	NA	-2	NA	NA	3	NA	τ.df = 2.00	τ.df = 2.00
	Goodness of fit	Global deviance	11,530.6	11,475.9	11,500	11,432.1	11,433.7	11,428	11,444.6	11,495.7	11,419.1	11,426.9
		AIC	11,546.58	11,497.9	11,522	11,460.1	11,457.7	11,452	11,472.6	11,513.5	11,437.9	11,445.8
		BIC	11,604.6	11,577.7	11,601.7	11,561.6	11,544.6	11,539	11,574.1	11,577.9	11,505.8	11,514.1
	Female	Distribution	NO	PE	BCCG	BCT	SEP3	ST3	BCPE	LMS(BCCG)	LMS(BCT)	LMS(BCPE)
		Model name	M1_G	M2_G	M3_G	M4_G	M5_G	M6_G	M7_G	M8_G	M9_G	M10_G
		Parameters	2	3	3	4	4	4	4			
		Mean	3, 3	3, 3	3, 3	3, 3	3, 3	3, 3	3, 3	μ.df = 2.88	μ.df = 2.83	μ.df = 2.81
		SD	3	3	-2	-2	3	3	-2	σ.df = 2.81	σ.df = 2.71	σ.df = 2.98
		Nu	NA	3	3	3	-2	-2	3	ν.df = 2.00	ν.df = 2.00	ν.df = 2.00
		Tau	NA	NA	NA	3	NA	NA	3	NA	τ.df = 2.00	τ.df = 2.00
	Goodness of fit	Global deviance	10,167.5	10,106.6	10,105.9	10,048.5	10,051.7	10,050	10,050	10,098.4	10,049	10,043.3
		AIC	10,183.5	10,128.6	10,127.9	10,076.5	10,075.7	10,074	10,078	10,113.7	10,068.1	10,062.9
		BIC	10,241.2	10,207.8	10,207.1	10,177.4	10,162.1	10,161	10,178.8	10,169.1	10,136.7	10,133.4

Table 4.2: Summary of birthweight results for the mean and standard deviation, LMS, LMST and LMSP methods. AIC, Akaike information criterion; BCCG, Box-Cox Cole and Green distribution; BCT, Box-Cox t-distribution; BIC, Bayesian information criterion; df, degrees of freedom; NO, normal distribution; PE, power exponential distribution; SEP3, skew exponential power type 3 distribution; ST3, skew t-distribution type 3 distribution.

4.12 Discussion

This chapter has described the principal methodologies available for the construction of age-specific reference centiles. I have demonstrated their application using the recently published INTERGROWTH-21st newborn data for weight [25]. The choice of methodology is important as inaccurate centiles resulting from inferior methods can lead to incorrect judgements about fetal size development, resulting in sub-optimal clinical care [190]. Choosing the best model from among many is not trivial, especially when dealing with large datasets such as the INTERGROWTH-21st data (N = 20,302). Significance testing and goodness-of-fit statistics like the likelihood ratio test or the AIC are usually used to discriminate between models. However, these methods tend not to be useful when examining large datasets, as very small differences are statistically significant even if they are indistinguishable on actual centile plots. Model choice should not be based on statistical considerations alone, but also on the quality of the fit to the data and the aesthetic appearance of the model fit across the GA range. Reference centiles should ideally be produced that have the best fit to the data and change smoothly with GA using as simple a statistical model as possible that can easily be transformed into z-scores (SDS scores), to ensure comparability and usability.

I have explored a variety of methods and models for fitting reference centiles. In selecting the best model, considerations such as identifying a common distribution for boys and girls that best represents the birthweight data were taken into account. I preferred to use the same functional FP model form for boys and girls, even though their data were modelled separately. Based on these considerations and a following a thorough analysis of the diagnostic plots, I selected the skew t-distribution (type 3) [191] with four parameters (μ, σ, υ, and τ) as the most appropriate distribution for constructing birthweight curves for boys and girls.

I used FPs with two powers for the mean and one for the SD to obtain the fractional powers. These powers were incorporated in a GAMLSS framework to model skewness and kurtosis. The values for skewness and kurtosis were constant but non-zero, as they did not vary with GA.

Before starting this analysis, I had not anticipated that modelling the birthweight data would need a more complex distribution than the normal distribution. I believe the requirement for more complex distributions can be explained by our carefully selected population of healthy women. These women primarily had good pregnancy outcomes. This led to what I refer to as 'data heaping': very few deliveries were observed in early gestation (< 34 weeks) as most of the women had term deliveries (≥ 37 weeks). This data heaping posed challenges for the data modelling due to the non-uniform distribution of the data across GA. Having significantly more data points in late gestation affected the fit at the bottom end of the distribution. A more complex distribution that accounted for skewness and kurtosis was therefore required. Data heaping can also be overcome by selecting a sub-sample of observations to artificially construct a database with a balanced number of observations over the range of GA or weight measurements. However, this method discards and wastes data, which is not recommended given the time and cost associated with obtaining the data [192]. A weighting approach can also be used, in which more weight is assigned to the few observations at the bottom of the distribution and less weight to data points lying with the majority of the data.

The LMS method provided a suitable alternative for modelling this dataset due to its flexibility, ability to account for skewness, and closed formulation to the normal distribution based on the L, M, and S curves. However, it did not perform well with the sparse data near the end of the age range (called edge effects) and the Box-Cox transformation did not adjust for the presence of kurtosis, which were both characteristics of these data. Quantile regression techniques are often

used to model growth data and have been shown to perform as well as parametric methods [193, 194, 195, 196, 197]. These methods allow quantiles to be estimated as a smooth function of GA without making any distributional assumptions about the data. However, nonparametric methods were not considered here because they lack a simple closed formula that can be used to estimate any desired centile or z-scores for individuals. They therefore do not allow direct comparisons between groups and are not easy to use clinically.

In this chapter, I have demonstrated the application of the mean and SD, LMS, LMST, and LMSP methods using birthweight data from the INTERGROWTH-21st NCSS. The methodology and statistical considerations discussed here were also applied to the newborn BL and BHC data. These considerations and methodology were key to developing the recently published international newborn standards [25]. The mean and SD method was sufficient for modelling the fetal size data.

5

Statistical methodology for longitudinal studies of human growth: Using the INTERGROWTH-21st Project as a case study

5.1 Introduction

In chapter 4, I considered various analytical approaches for cross-sectional data. However, fetal and newborn size charts can also be constructed from longitudinal data. Analysis of longitudinal studies is complex and supported by a rather limited literature [67, 92]. Repeated measures data pose analytical challenges that require different analysis techniques to single measures data, because such data deviate from the independence of observations assumption that most classical statistical methods are based on. The effect of correlation between measurements of the

same subject at different ages, number of replicates per measurement, and the number of observations per subject need to be considered [198, 104]. A multi-level analysis accounts for the non-independence of observations by considering the hierarchical structure of the data and the correlation between measurements from the same fetus at different GAs.

The intention of this chapter is not to provide an exhaustive review of all of the statistical approaches that can be applied in the analysis of longitudinal data. Instead it aims to give a brief overview of common methodology used for deriving charts of fetal size based on a longitudinal design. In this chapter, I demonstrate and compare statistical methodologies for constructing GA-related size charts from longitudinal data (repeated measures data). FHC from the INTERGROWTH-21st Project is used as an example. I demonstrate the analysis using FP regression embedded in a multi-level framework. The effect of fitting various multi-level models is evaluated by fitting a two-level random intercept model, a two-level random intercept and slope model, a two-level random intercept and slope model fitted to a randomly selected single HC observation from a set of triplicate measurements for each woman at each visit, and a three-level random intercept and slope model.

I investigate and assess the effect of ignoring various design aspects of the repeated FHC data by (a) ignoring the multi-level structure of all FHC data by treating it like cross-sectional data, (b) randomly selecting a single HC measurement from a set of triplicate HC measurements taken at each visit for each woman to evaluate the value of repeated measurements, and (c) transforming the longitudinal FHC data to cross-sectional data (reducing the dataset) by selecting one HC measurement per woman at one time point at random, to evaluate the effect of reducing the data sample size. I compare the model fits from these three approaches using goodness-of-fit statistics and diagnostic plots.

5.2 Data

Ultrasound was used to take fetal anthropometric measurements prospectively from 14^{+0} weeks until birth in a cohort of women with optimal health and adequate nutritional status who were at low risk of intrauterine growth restriction. In this analysis, I only considered FHC obtained every 5 weeks ($\pm$ 1 week) from 14^{+0} to 42^{+0} weeks gestation. At each visit, HC was measured three times from three separately obtained ultrasound images in a blinded fashion, i.e., previous measurements of each structure were not available to the assessor through suppressed display [24]. The decision to take triplicate measurements was made at the design stage following advice from an expert ultrasonographer. True replicates were used as taking three measurements of the same image would show much less between-replicate information and was not considered to be worthwhile. The study recruited 4,233 women who each visited one to six times during pregnancy (95% visited at least four times), giving 20,030 women visits. With three ultrasounds at each visit, 59,973 FHC observations were made across the eight sites (117 ultrasound measures were missing).

The longitudinal design introduced another level of complexity. The data structure was composed of a three-level hierarchy, measurements within visits within participants. Level 1 is the triplicate measurements taken at each visit, HC_1, HC_2, and HC_3, the first, second and third measurements, respectively. Level 2 is the repeated ultrasound measurements taken for each woman over multiple visits during pregnancy. Level 3 is the measurements taken from women in a particular site (country). The data analysis must consider the hierarchical structure of the data and the correlations of the measurements within a subject and between subjects at a given site. Table 5.1 summarises the number of follow-up visits, the number of women across all sites who made only X visits with an FHC measurement, and the

total number of women who visited at least X times with an FHC measurement, covering GA from 14 to 40 weeks. Scatter plots of the raw FHC data by GA for all of the sites combined are shown in Figure 5.1.

Number of follow up visits (X)	Number of women who visited only X times visits in total	%	Number of women who visited at least X times visits	%
1	39	0.9	4,233	100
2	55	1.3	4,194	99.1
3	203	4.8	4,139	97.8
4	810	19.1	3,936	93.0
5	2,724	64.4	3,126	73.8
6	402	9.5	402	9.5
Total	4,233			

Table 5.1: Summary of the number of women at each visit at which fetal head circumference was measured.

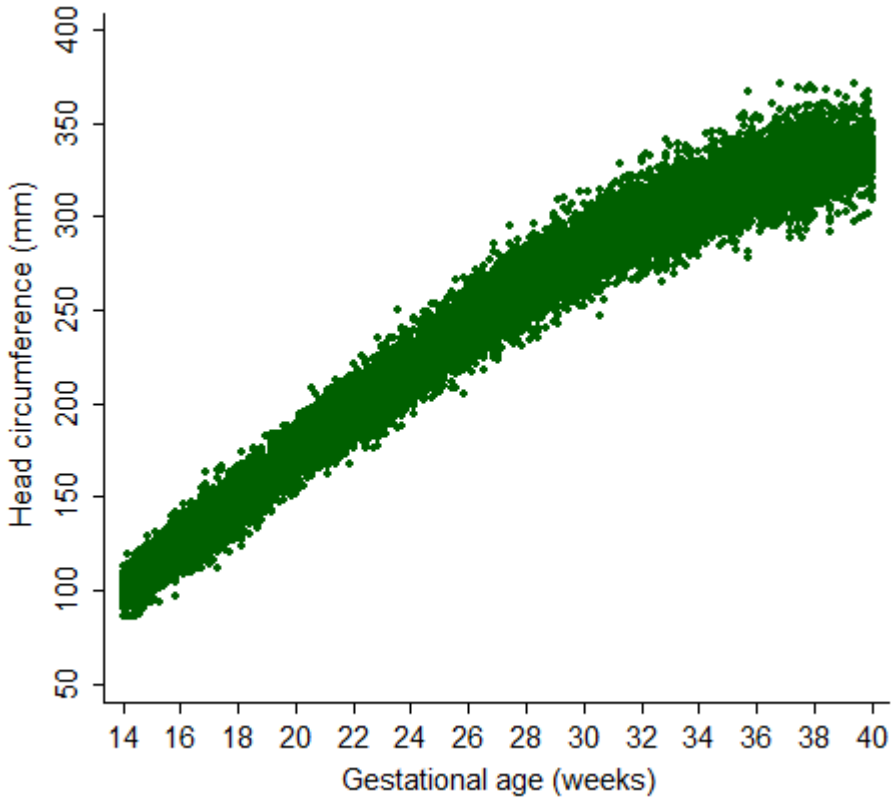


Figure 5.1: Scatter plots of the raw fetal head circumference measurements by gestational age for all of the sites combined.

5.3 Methods

5.3.1 Statistical methodology

To account for repeated measures, multi-level models [199, 200] were applied. To judge the effect of ignoring various design aspects of the repeated FHC data, I used the mean and SD method with FPs based on the assumption that for each GA, the measurement of interest had a normal distribution with a mean and SD that varied smoothly with GA [163], as discussed in Chapter 4. As in Chapter 4, selected models needed to (a) develop smooth centiles that offered a good representation of the raw data, (b) model the data precisely, especially the outer centiles (e.g., the 3^{rd} and 97^{th} centiles) where variability is greatest, (c) produce non-crossing centiles, (d) allow estimates of z-scores and centiles to be calculated, (e) apply continuous age smoothing, not age binning, and (f) offer flexibility to account for both skewness and kurtosis when necessary.

The best fitting powers for the median HC were provided by an FP2 which was incorporated in a multi-level framework to account for repeated measures.

5.3.2 Data analysis

5.3.2.1 Mean and standard deviation method

Three models were created using FPs and different datasets. Models 1 and 2 were used to assess the effect of ignoring the multilevel structure of the data. Both models were fitted to independent datasets of 20,030 HC values. In Model 1, the average of the triplicate measurements HC_1, HC_2, and HC_3 from each visit was used, transforming each triplicate into a single HC value. As using the average of the triplicate measurements could reduce variability, Model 2 instead used a single randomly selected measurement from each triplicate. I randomly selected

6,771 (33.80%) HC_1, 6,566 (32.78%) HC_2, and 6,693 (33.41%) HC_3 measurements from the dataset.

Model 3 was used to assess the effect of transforming the longitudinal data into cross-sectional data. A single measurement was randomly selected from each subject, discarding 80% of the data. Each subject was thus represented by one HC value, which could have been taken at any point during the study. The dataset for Model 3 comprised 4,233 randomly selected single HC measurements, including 1,429 (33.76%) HC_1, 1,408 (33.26%) HC_2, and 1,396 (32.98%) HC_3 measurements.

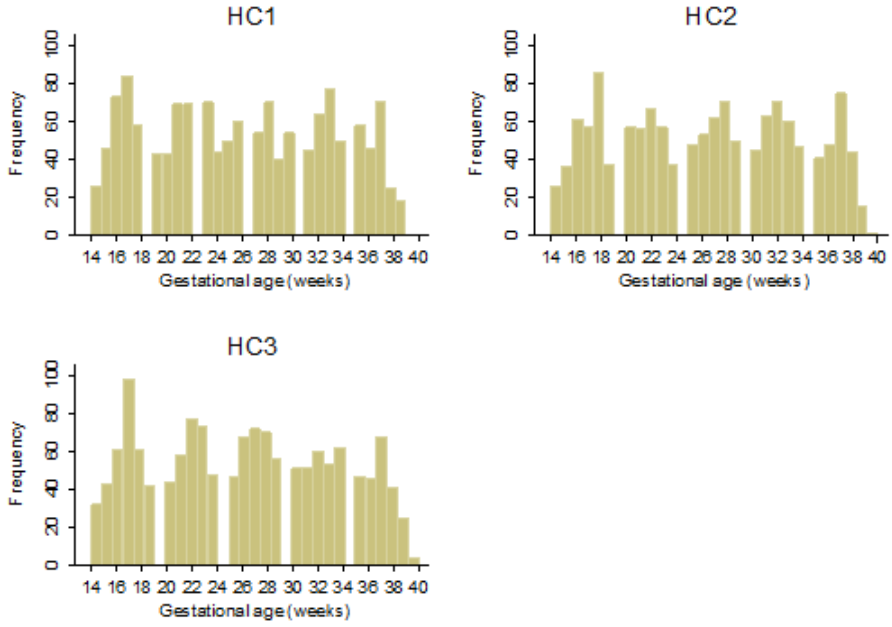


Figure 5.2: Distribution of the randomly selected fetal head circumference measurements by gestational age and order of HC measurement (HC_1, HC_2, and HC_3) in the original triplicate set in the transformed cross-sectional dataset.

5.3.2.2 Multi-level modelling

Multi-level linear models (also commonly referred to as hierarchical or mixed effect models) are increasingly used for longitudinal data [201, 202, 203]. They are regression equations that include both fixed and random components [204]. The fixed components are the same for every subject and the random components differ between subjects according to a normal distribution. These methods are preferred because they allow each subject's growth pattern over time to be characterised and they take into account the correlation structure between measurements from the same individual. The between-subject variability in the specified population can thus be quantified.

The FHC data has three levels that can be expressed by a simple linear regression model. For a given FHC measurement y_i (i = 1, 2, 3) of subject j (j = 1, 2, ..., 4,233 subjects) taken on visit k (k = 1, 2,...,7 visits):

$$y_{ijk} = \beta_0 + \beta_1 X_{ijk} + \varepsilon_{ijk}. \tag{5.1}$$

Equation 5.1 represents the regression of the FHC, y, on the independent variable, GA, in weeks, X. In a typical regression model, the errors ε_{ijk} are assumed to be independent and normally distributed with mean, μ and variance, σ^2. However, the independence assumption does not hold for the HC dataset, as it includes repeated measurements of each fetus. Equation 5.1 assumes that growth across time is the same for all fetuses i.e., that β_0 and β_1 do not vary by fetus. I therefore included individual-specific effects to account for data dependency and characterise the difference in growth between individual fetuses. I used four approaches to factor in these fetus-specific effects and the resulting variation, resulting in four models of increasing complexity.

Two-level random intercept model (Model 4):

In the first multi-level model, Model 4, the first level of the data was collapsed by taking the average of the triplicate FHC measurements HC_1, HC_2, and HC_3 taken at each visit, as in Model 1. The data then comprised two levels, one ultrasound measurement at each visit for each subject during pregnancy (level 1) and all of the measurements from the women at each of the eight sites (level 2). Exploring the influence of each woman on their repeated FHC measurements led to:

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + \upsilon_{0j} + \varepsilon_{ij}, \tag{5.2}$$

where υ_{0j} is the influence of individual i on their repeat measurements. For a given group j, the intercept is $\beta_0 + \upsilon_{0j}$. The model in Equation 5.2 is often partitioned into two components in a multi-level framework. The fixed component (within-subject, level 1) is $\beta_0 + \beta_1 X_{ij}$ and the random component (between-subject, level 2) is $\upsilon_{0j} + \varepsilon_{ij}$. Equation 5.2 indicates that subject i's initial FHC measurement at GA (time) j is influenced by that subject's initial level $\beta_0 + \upsilon_{0j}$ and the population's slope β_1. Using this relation, each individual has their own distinct initial level. The resulting model is commonly referred to as the random intercept model [205]. Model 4 was then created using Equation 5.2 and the same dataset as that used in Model 1, averaging the triplicate FHC measurements taken at each visit (N = 20,030).

Two-level random intercept and slope model (Models 5 and 6):

Equation 5.2 in Model 4 uses the same slope, equal to the population slope β_1, for every fetus. This assumption is too simplistic for our situation, as it is unlikely that every fetus will have the same rate of growth in HC by GA. In Model 5, I relaxed this assumption and assigned each fetus its own initial level (intercept)

and slope that varied with GA:

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + \upsilon_{0j} + \upsilon_{1j} X_{ij} + \varepsilon_{ij}. \tag{5.3}$$

Equations 5.2 and 5.3 model the first level in the same way. Equation 5.3 includes the term υ_{1j}, which represents the slope deviation for each subject i from the average regression slope β_1. As before, ε_{1j} is an independent error term distributed normally with mean 0 and variance σ^2. Model 5 was created using Equation 5.3 and the same dataset as Models 1 and 4, by averaging triplicate FHC measurements taken at each visit (N = 20,030). Model 6 was created using Equation 5.3 and the same dataset as Model 2, by randomly selecting one of HC_1, HC_2, and HC_3 in each triplicate (N = 20,030).

Three-level random intercept and slope model (Model 7):

Models 4–6 considered only two data levels. The full data has three levels: triplicate measurements collected at each visit, HC_1, HC_2, and HC_3 (level 1); repeated ultrasound measurements for each woman across multiple visits during the pregnancy (level 2); and measurements taken from many women (level 3). Considering all three data levels, Equation 5.3 becomes:

$$y_{ijk} = \beta_0 + \beta_1 X_{ijk} + \upsilon_{0jk} + \upsilon_{1jk} X_{ijk} + \varepsilon_{ijk}. \tag{5.4}$$

Model 7 was created using Equation 5.4 and the complete dataset of all triplicate HC measurements (N = 59,973).

5.4 Results

Figure 5.1 shows a scatter plots of the raw FHC data by GA for the data from all eight sites. Table 5.1 shows a descriptive summary of the number of women

and the total number of follow-up visits attended during pregnancy. Most women (93%) attended at least four follow-up visits during pregnancy. Figure 5.2 shows the distribution of HC_1, HC_2, and HC_3 measurements by GA selected for Model 3.

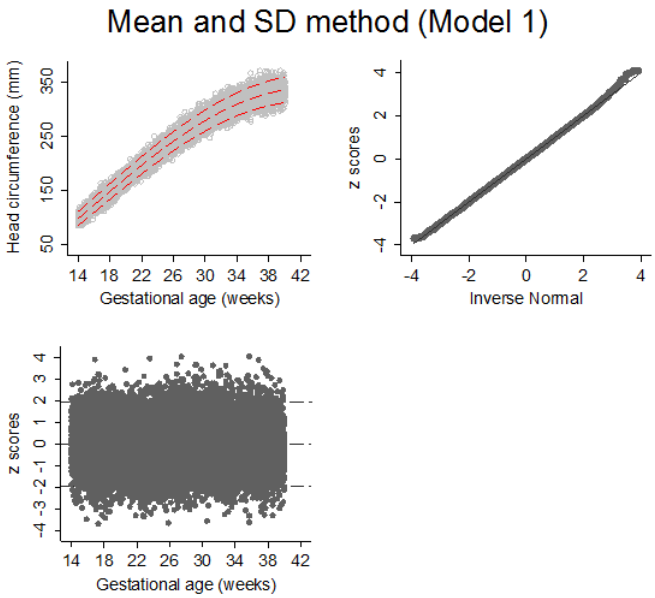
5.4.1 Mean and standard deviation method results

The mean and SD were modelled separately for each of the seven models. In each case the relation between mean FHC and GA was best fit by an FP2. For SD, the relation with GA was linear. Table 5.2 contains the specifications, goodness-of-fit, and comparisons of Models 1–3 fitted to the FHC data using FPs. The fitted centiles based on the three models are shown in Figures 5.3–5.7. The plots show the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles across GA, normal Q-Q plots of the z scores, and z-scores by GA. The models were compared by quantifying maximum absolute differences at the 3^{rd}, 50^{th}, and 97^{th}. Similar results were obtained for all three models, and Models 1 and 2 were indistinguishable (Figure 5.9). The maximum absolute differences between Models 1 and 2 were <1 mm at the 3^{rd}, 50^{th}, and 97^{th} centiles. Model 3 differed by a maximum of 8 mm from Model 1 at the 3^{rd} centile and by 7 mm from Model 2 close to term. These are small differences, as they are <1 cm (Tables 5.2 and 5.9). It is not surprising that there were larger differences between Model 3 and the other models than between Models 1 and 2 as Model 3 used an 80% smaller sample size (4,233 vs. 20,030 observations). All three models had adequate model fit, based on the distribution of the normal Q-Q plots of the z-scores. The z-scores by GA did not show any obvious pattern across GA (Figures 5.3–5.7). The estimated proportions of observations falling below the 3^{rd} centile or above the 97^{th} centile were comparable with the expected 3% (Table 5.2). In terms of precision, the difference between the upper and lower 95% CI divided by two was calculated for the 3^{rd}, 50^{th}, and 97^{th} for Models 1, 2, and 3. Models 1 and 2 were very similar in precision for the respective centiles 5.4 and 5.6. Model 3 had the greatest precision across the three centiles 5.8. As

expected, precision increased with increase in GA 5.10.

				Fractional polynomial powers			Goodness of fit		Maximum absolute difference between models (mm)			Model differences
Model	Model specification	Detail	N	Median	SD	Deviance	Observations $< 3^{rd}$ centile	Observations $> 97^{th}$ centile	3^{rd} centile	50^{th} centile	97^{th} centile	
1	Ignoring the multi-level structure of HC data	Take the mean of the triplicate HC measurements for each visit	20,030	2, 2	1	147173.59	601 (3.00%)	613 (3.06%)				
2	Randomly select one HC measurement from the set of triplicate measurements for each subject at each visit	Randomly select one of the three HC measurements for each visit	20,030	2, 2	1	148012.17	581 (2.90%)	623 (3.11%)	0.74	0.06	0.62	M1 - M2
3	Transform the data from cross-sectional to longitudinal	Randomly select one HC measurement per subject from all of the measurements taken at all of that subject's visit	4,233	2, 3	1	31295.32	135 (3.19%)	131 (3.09%)	7.66 6.91	6.93 6.86	6.20 6.82	M1 - M3 M2 - M3

Table 5.2: Mean and SD method: Model details and results using fractional polynomials



Fitted 3^{rd}, 50^{th}, and 97^{th} smoothed fetal head circumference centile curves (dashed red lines) for fetal head circumference (mm) by ultrasound according to gestational age, showing the actual observations (open grey circles) (top left), quantile-quantile plot (top right), and z-score by gestational age (weeks) (bottom left), of the fractional polynomial model applied to the average of the triplicate fetal head circumference measurements taken at each visit (Model 1).

Figure 5.3: Fractional polynomial model applied to the average of the triplicate fetal head circumference measurements taken at each visit (Model 1).

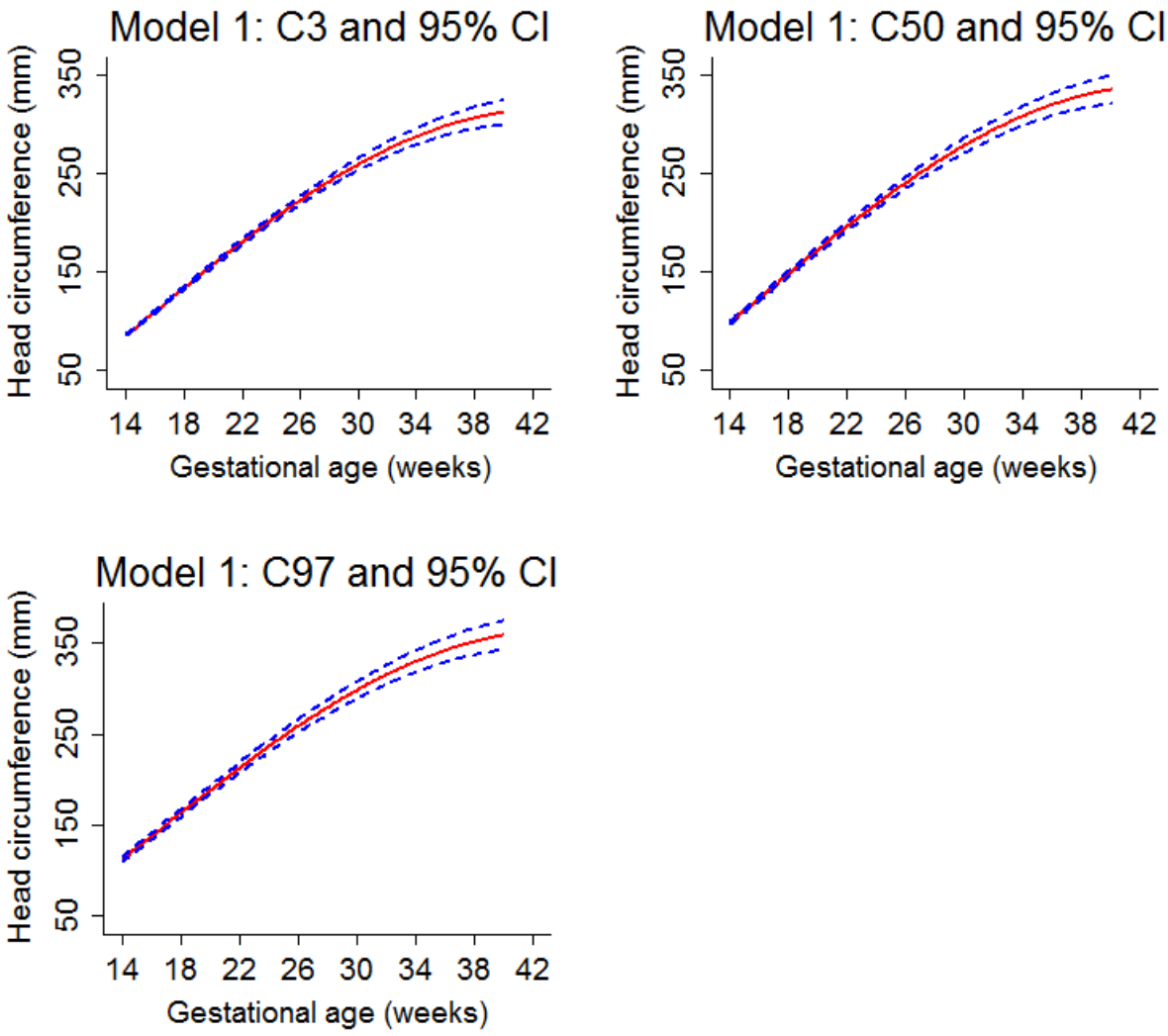


Figure 5.4: Comparisons of the change in SD of the fitted fractional polynomial model applied to the average of the triplicate fetal head circumference measurements taken at each visit (Model 1).

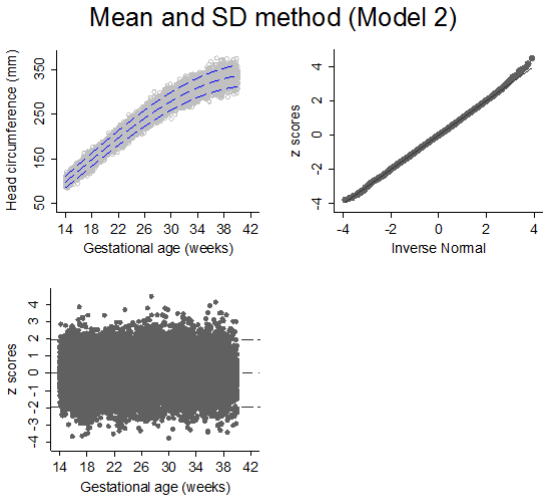


Figure 5.5: Fractional polynomial model applied to a single fetal head circumference measurement selected at random from the set of triplicate fetal head circumference measurements taken at each visit (Model 2).

Fitted 3^{rd}, 50^{th}, and 97^{th} smoothed fetal head circumference centile curves (dashed red lines) for fetal head circumference (mm) by ultrasound according to gestational age showing actual observations (open grey circles) (top left), quantile-quantile plot (top right) and z score by gestational age (weeks) (bottom left)

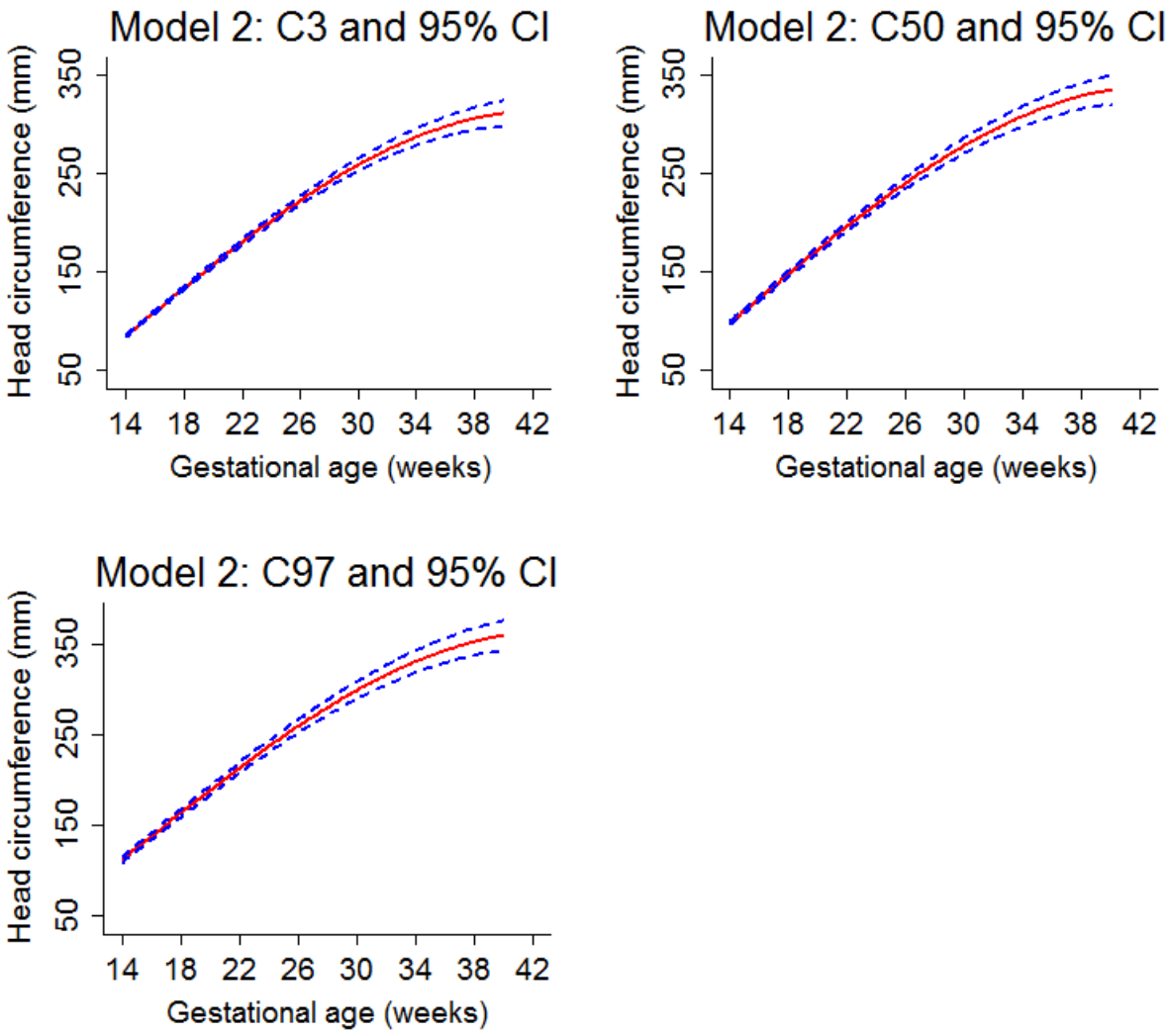


Figure 5.6: Comparisons of the change in SD of the fitted fractional polynomial model applied to a single fetal head circumference measurement selected at random from the set of triplicate fetal head circumference measurements taken at each visit (Model 2).

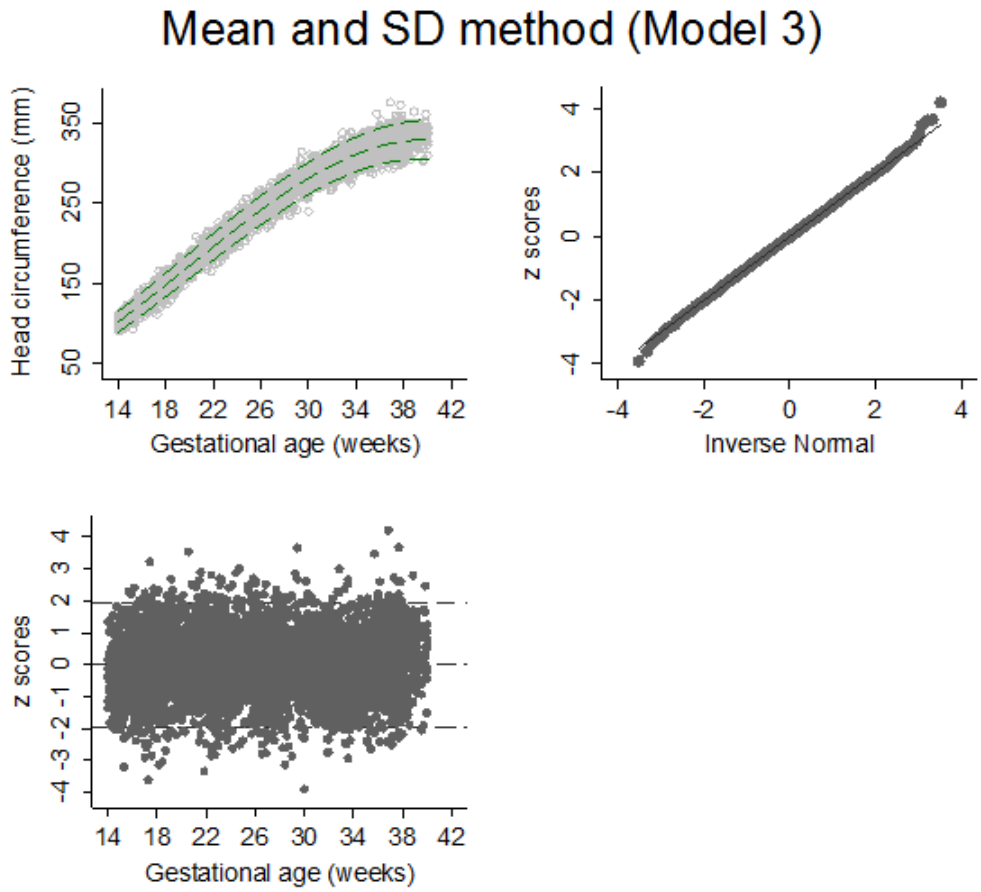


Figure 5.7: Fitted 3^{rd}, 50^{th}, and 97^{th} smoothed fetal head circumference centile curves (dashed red lines) for fetal head circumference (mm) by ultrasound according to gestational age, showing the actual observations (open grey circles) (top left), quantile-quantile plot (top right), and z-score by gestational age (weeks) (bottom left), of the fractional polynomial model applied to cross-sectional data (Model 3).

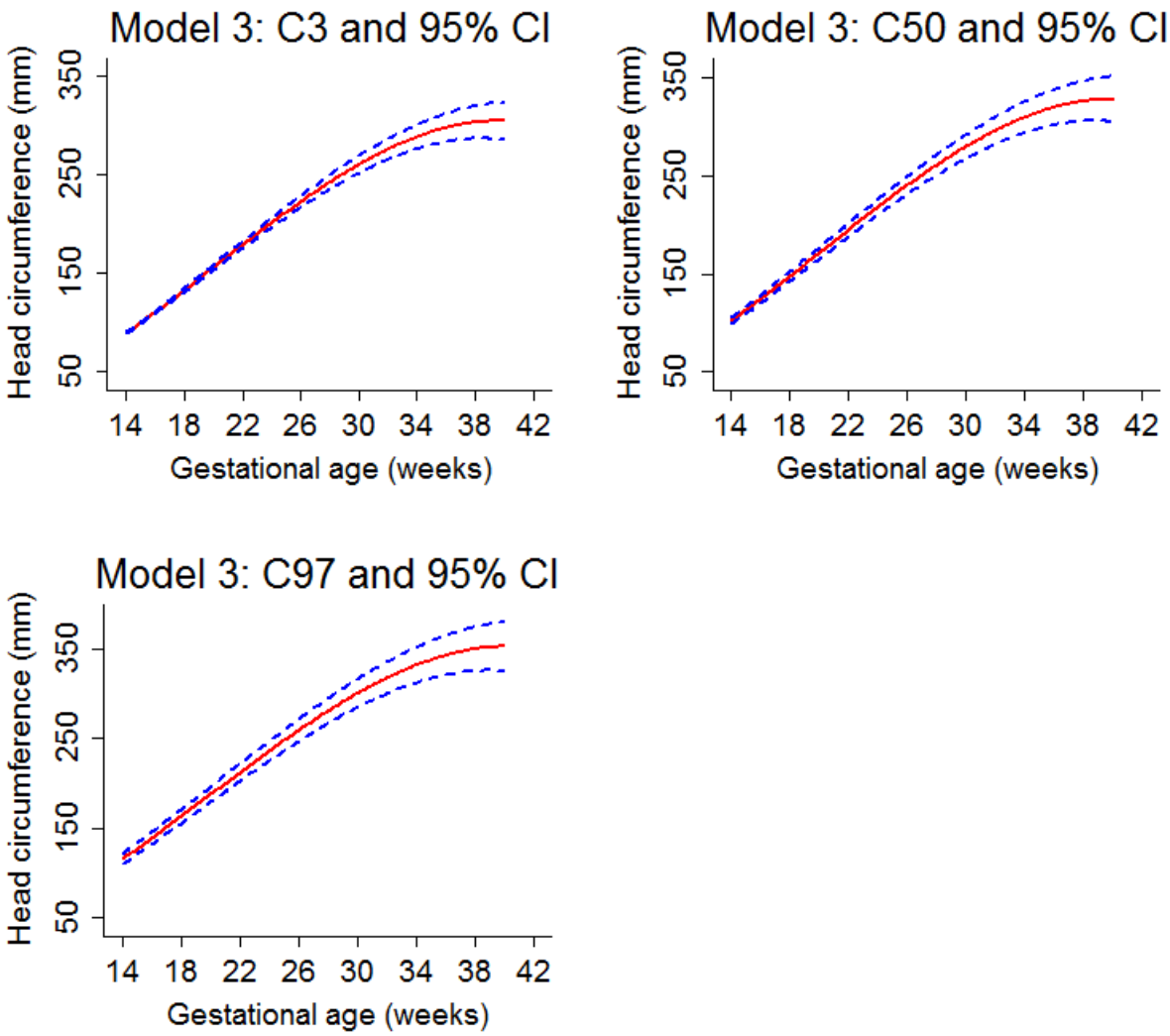


Figure 5.8: Comparisons of the change in SD of the fitted fractional polynomial model applied to cross-sectional data (Model 3).

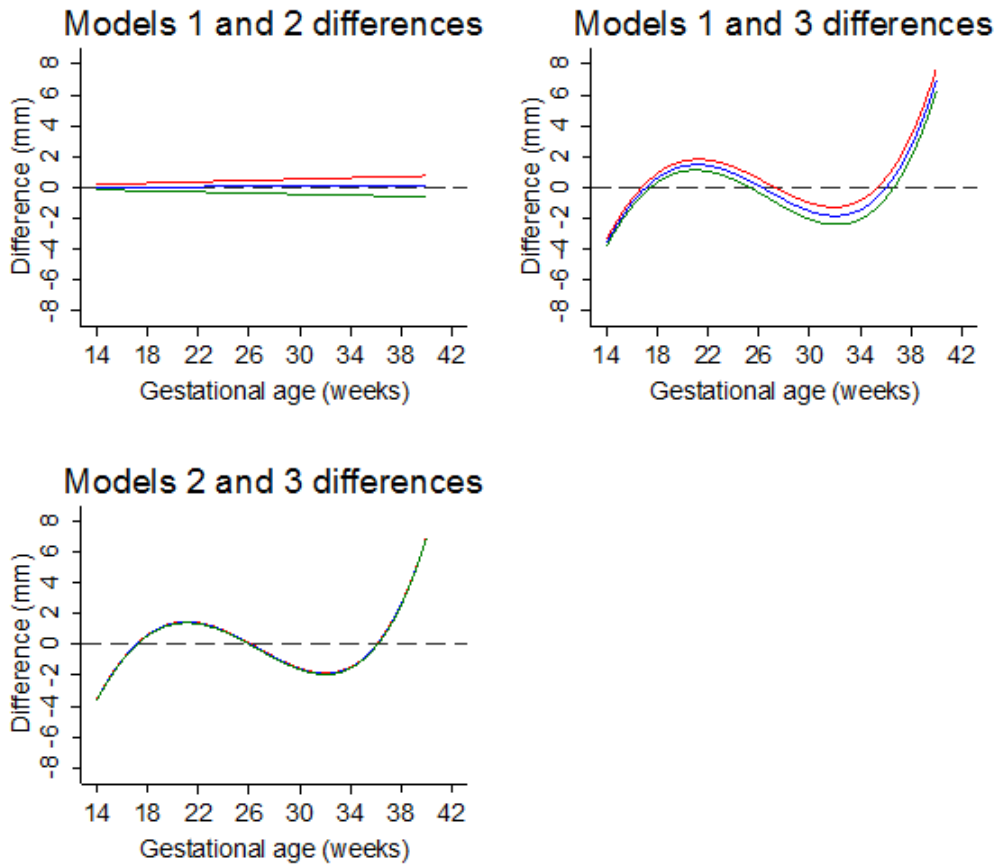


Figure 5.9: Comparison of the maximum absolute differences (mm) between Models 1, 2 and 3, based on the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed fetal head circumference centile curves. Model 1 was based on the average of the triplicate measurements taken at each visit, which ignored the multi-level structure of the data and transformed it into a single fetal head circumference value. Model 2 was based on a single fetal head circumference measurement that was randomly selected from each triplicate, HC_1, HC_2, and HC_3. Model 3 was based on a single randomly selected fetal head circumference measurement taken at any point during the study for each subject, with each subject represented by one fetal head circumference value. The differences between Models 1 and 2 (top left), Models 1 and 3 (top right), and models 2 and 3 (bottom left) are shown.

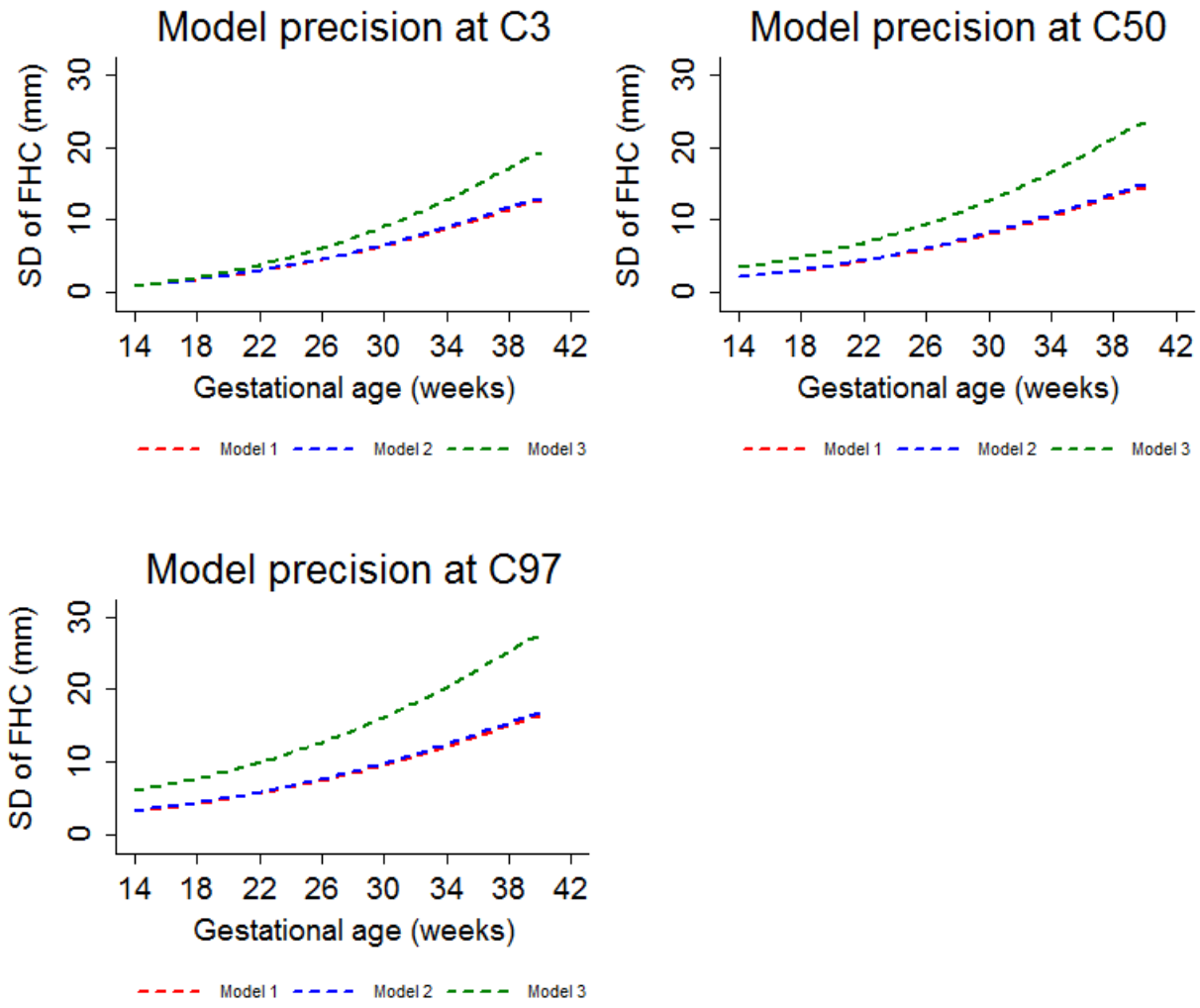


Figure 5.10: Comparison of the change in SD of fetal head circumference (mm) between Models 1, 2 and 3, based on the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed fetal head circumference centile curves. Model 1 was based on the average of the triplicate measurements taken at each visit, which ignored the multi-level structure of the data and transformed it into a single fetal head circumference value. Model 2 was based on a single fetal head circumference measurement that was randomly selected from each triplicate, HC_1, HC_2, and HC_3. Model 3 was based on a single randomly selected fetal head circumference measurement taken at any point during the study for each subject, with each subject represented by one fetal head circumference value.

5.4.2 Multi-level modelling results

Table 5.3 contains the specifications, goodness-of-fit, and comparisons of Models 4–7 fitted to FHC data. As with Models 1–3, I fitted an FP with two powers for the mean and applied it to a multi-level framework to account for repeated measures. The fitted centiles based on the four models are shown in Figures 5.11–5.14. The plots show the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles across GA, normal Q-Q plots of the z-scores, plot of within- and between-subject residuals, and z-scores by GA.

The models were compared by quantifying the maximum absolute differences at the 3^{rd}, 50^{th}, and 97^{th} centiles. The four multi-level models were formulated differently, but had reasonably similar results. The two two-level random intercept and slope models, Models 5 and 6, were very similar, with a maximum absolute difference of $<$0.5 mm at the extreme centiles. In terms of precision, model 5 had better precision compared to Model 6 5.18.

All of the models were compared with Model 1, which ignored the hierarchical structure of the data and was based on taking the mean of the triplicate measurements of HC at each visit. Model 1 differed at most by 2.8 mm and 2.6 mm from the random intercept model (Model 4) and random intercept and slope model (Model 5), respectively, which both used the same dataset as Model 1. Model 1 differed by 2.5 mm from the random intercept and slope model based on randomly selecting one HC measurement from each triplicate (Model 6) and by 2.7 mm from the three-level random intercept and slope model (Model 7).

The random intercept model (Model 4) differed by at most 3.4 mm and 3.8 mm at the 3^{rd} centile from the random intercept and slope two-level models, Model 5 and Model 6, respectively. Models 4–6 all used the same dataset. Model 5 differed by 0.4 mm at the 3^{rd} and 97^{th} centiles from the two-level random intercept and slope model based on randomly selecting one HC measurement from each triplicate

(Model 6) and by 2.2 mm at the 97^{th} centile from the three-level random intercept and slope model (Model 7). Model 4 differed by 4.7 mm at the 3^{rd} centile from the three-level random intercept and slope model (Model 7) and by 2.4 mm at the 97^{th} centiles from the two-level random intercept and slope model based on randomly selecting one HC measurement from each triplicate (Model 6). These results are summarised in Table 5.3. Model comparisons at the 3^{rd}, 50^{th}, and 97^{th} centiles are shown in Figures 5.15–5.17.

The multi-level models though very similar when absolute differences between observed and predicted centiles are compared, they differ in precision with Model 5 having the best precision at the 50^{th} centile 5.18 but not when compared with Model 1.

				Fractional polynomial powers			Goodness of fit		Maximum absolute difference between models (mm)			Model differences
Model	Model specification	Detail	N	Median	SD	Deviance	Observations $< 3^{rd}$ centile	Observations $> 97^{th}$ centile	3^{rd} centile	50^{th} centile	97^{th} centile	
4	Random intercept two-level model	Take the mean of the triplicate HC measurements for each visit	20,030	2, 2	1	136690	638 (3.2%)	647 (3.2%)	2.0	0.3	2.8	M1 - M4
5	Random intercept and slope two-level model	Take the mean of the triplicate HC measurements for each visit	20,030	2, 2	1	132845	866 (4.3%)	923 (4.6%)	2.6 3.4	0.3 0.4	2.5 2.5	M1 - M5 M4 - M5
6	Random intercept and slope two-level model	Randomly select one of the three HC measurements for each visit made at all of that subject's visits	20,030	2, 2	1	135710	926 (4.6%)	1009 (5.0%)	2.4 3.8 0.4 2.7 4.7 1.3	0.2 0.5 0.1 0.8 1.0 0.6	2.5 2.8 0.6 2.6 4.4 2.2	M1 - M6 M4 - M6 M5 - M6 M1 - M7 M4 - M7 M5 - M7
7	Random intercept and slope three-level model	Considering all three data levels	59,973	2, 2	1	346036	2993 (5.0%)	3231 (5.4%)	1.6	0.6	2.4	M6 - M7

Table 5.3: Model details and results of the multi-level modelling

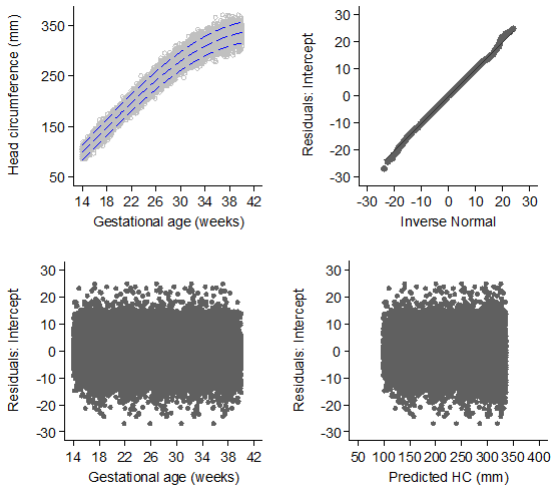


Figure 5.11: Two-level random intercept multi-level model applied to the average of the triplicate fetal head circumference measurements taken at each visit (Model 4).

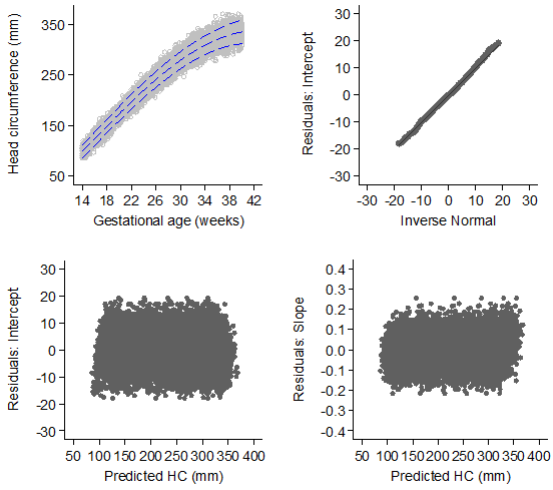


Figure 5.12: Two-level random intercept and slope multi-level model applied to the average of triplicate fetal head circumference measurements taken at each visit (Model 5).

Fitted 3^{rd}, 50^{th}, and 97^{th} smoothed fetal head circumference centile curves (dashed red lines) for fetal head circumference (mm) by ultrasound according to gestational age (weeks) showing the actual observations (open grey circles) (top left), normal plot of intercept residuals against normal scores (top right), plot of intercept residuals against predicted head circumference values (bottom left), and slope residuals against predicted HC values (bottom right).

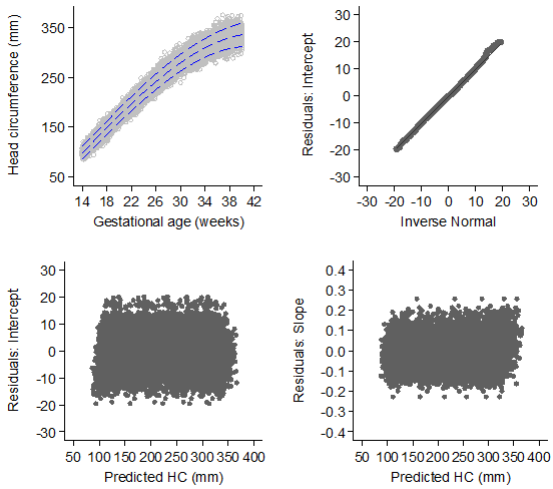


Figure 5.13: Two-level random intercept and slope multi-level model applied to a single fetal head circumference selected at random from the set of triplicate fetal head circumference measurements taken at each visit (Model 6).

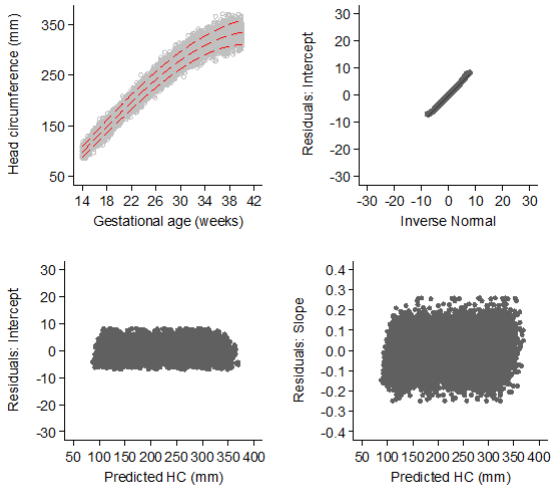


Figure 5.14: Three-level random intercept and slope multi-level model applied to all fetal head circumference triplicate measurements (Model 7).

Fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centile curves (dashed red lines) for fetal head circumference (mm) by ultrasound according to gestational age (weeks) showing the actual observations (open grey circles) (top left), normal plot of intercept residuals against normal scores (top right), plot of intercept residuals against predicted fetal head circumference values (bottom left), and slope residuals against predicted HC values (bottom right).

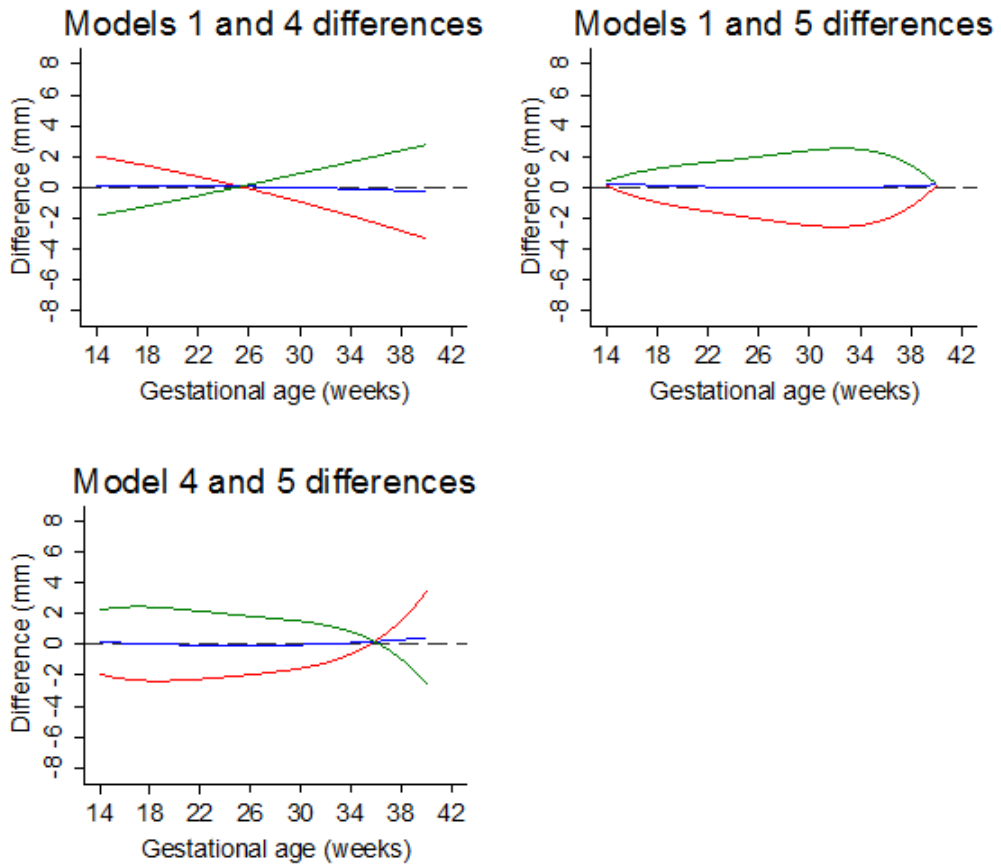


Figure 5.15: Comparison of the maximum absolute differences (mm) between Models 1, 4, and 5, based on the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed fetal head circumference centile curves. Model 1 was based on the average of the triplicate measurements taken at each visit, which ignored the multi-level structure of the data and transformed it into a single fetal head circumference value. Model 4 is a two-level random intercept model based on the average of the triplicate measurements taken at each visit. Model 5 is a two-level random intercepts and slope model based on the average of the triplicate measurements taken at each visit. The differences between Models 1 and 4 (top left), Models 1 and 5 (top right), and Models 4 and 5 (bottom left) are shown.

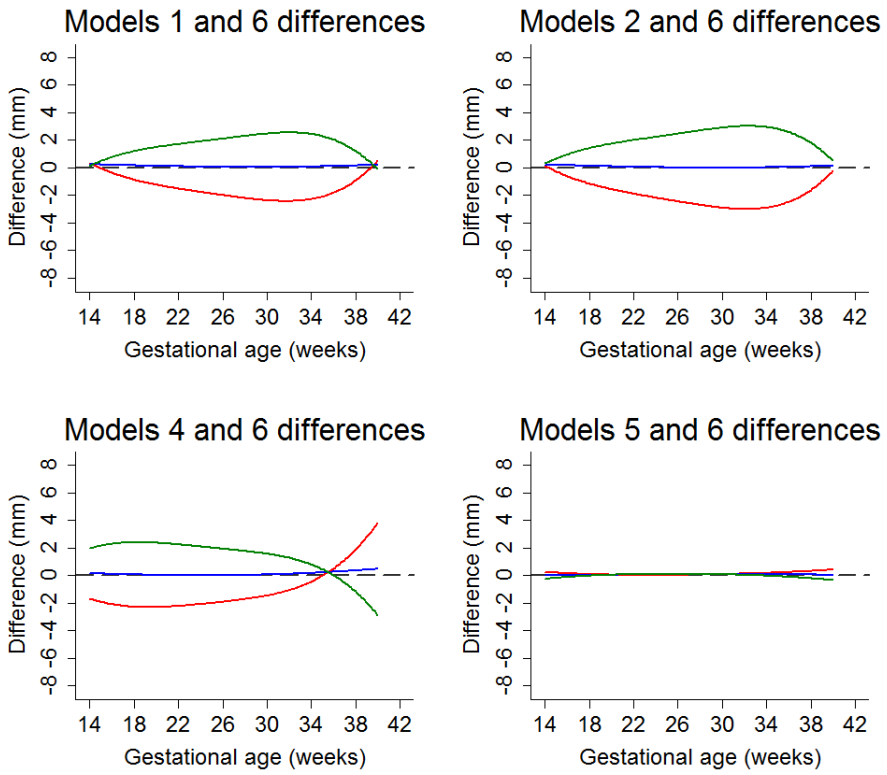


Figure 5.16: Comparison of the maximum absolute differences (mm) between Models 1, 2, 4, 5, and 6, based on the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centile curves. Model 1 was based on the average of the triplicate measurements taken at each visit, which ignored the multi-level structure of the data and transformed it into a single fetal head circumference value. Model 2 was based on a single fetal head circumference measurement randomly selected from each triplicate, HC_1, HC_2, and HC_3. Models 4 and 5 are two-level random intercepts and slope models based on the average of the triplicate measurement randomly selected from each triplicate, HC1, HC2, and HC3. The differences between Models 1 and 6 (top left), Models 2 and 6 (top right), and Models 5 and 6 (bottom left) are showns.

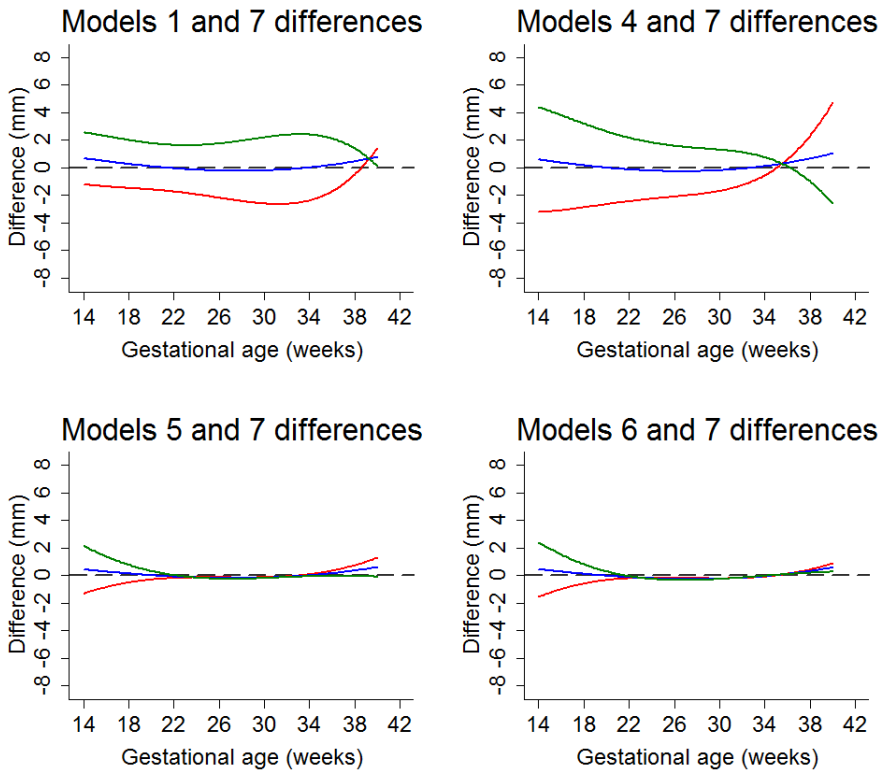


Figure 5.17: Comparison of the maximum absolute differences (mm) between Models 1, 4, 5, 6, and 7, based on the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centile curves. Model 1 was based on the average of the triplicate measurements taken at each visit, which ignored the multi-level structure of the data and transformed it into a single fetal head circumference measurement value. Models 4 and 5 are two-level random intercepts and slope models based on the average of the triplicate measurements taken at each visit. Model 6 is a two-level random intercepts and slope model based on a single fetal head circumference measurement randomly selected from each triplicate, HC_1, HC_2, and HC_3. Model 7 is a three-level random intercepts and slope model based on all fetal head circumference measurements. The differences between Models 1 and 7 (top left), Models 5 and 7 (top right), Models 6 and 7 (bottom left), and Models 3 and 7 (bottom right) are shown.

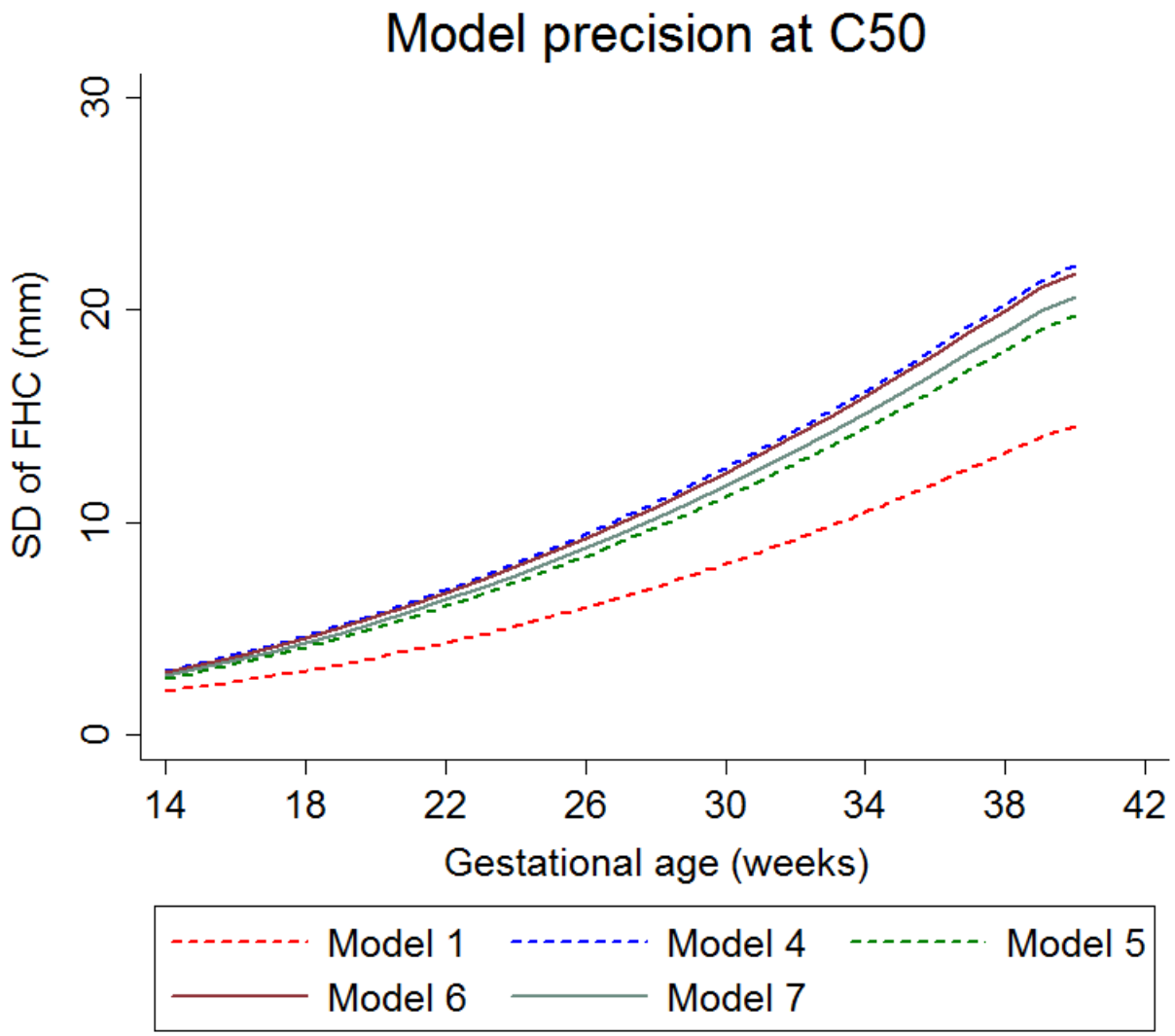


Figure 5.18: Comparison of the change in SD of fetal head circumference (mm) between Models 1, 4, 5, 6, and 7, based on the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centile curves. Model 1 was based on the average of the triplicate measurements taken at each visit, which ignored the multi-level structure of the data and transformed it into a single fetal head circumference measurement value. Models 4 and 5 are two-level random intercepts and slope models based on the average of the triplicate measurements taken at each visit. Model 6 is a two-level random intercepts and slope model based on a single fetal head circumference measurement randomly selected from each triplicate, HC_1, HC_2, and HC_3. Model 7 is a three-level random intercepts and slope model based on all fetal head circumference measurements.

5.5 Discussion

In obstetrics, ultrasound fetal measurements are often assessed at a single time point. The size of the fetus at a given GA is assessed to judge whether it is within the expected size range for that gestation. A well-designed cross-sectional study will suffice to produce reference charts for this purpose. However, sometimes one is interested in monitoring changes in growth (velocity) to track how a fetus is growing or has grown since the last observation. A longitudinal design is then required and more complex statistical models that take into account the correlation structure of repeat observations of an individual must be used. The current absence of good reference charts makes checking growth difficult. It is usually only done by comparing two fetal size assessments with their respective cross-sectional size charts.

The aim of this chapter was to model the data to develop centiles that change smoothly with GA, using the simplest statistical models possible. I aimed to identify modelling approaches that offered a good fit to the raw data and accounted for the increasing variability with GA, which is a phenomenon observed in growth data. I evaluated model goodness of fit both visually and formally using statistical tests [67, 92]. Precision of respective centiles was also evaluated.

Models based on a cross-sectional design are usually based on independent observations where each fetus in the sample contributes only one measurement. Longitudinal data includes multiple measurements per subject. Models based on longitudinal data cannot ignore within-person correlations, or they risk overestimating centile precision, resulting in narrower centiles [206, 207, 208]. However, the degree of bias introduced by not accounting for the correlation structure – by analysing repeated measures data as though they are cross-sectional – has received little attention [192]. This chapter compared different analytical approaches and considerations for working with multi-level data and aimed to demonstrate the loss or gain in

accounting for data dependency for repeated measures data.

I analysed a longitudinal dataset with three levels and modelled the data using multi-level models that accounted for the data hierarchy in different ways. I also ignored the longitudinal structure and analysed the data as purely cross-sectional. I analysed the data by assuming the observations were independent, by transforming the longitudinal dataset into a cross-sectional dataset by selecting one observation per subject, and by taking into consideration the full multi-level structure of the data. In clinical practise, measurements are often taken only once. In contrast, the INTERGROWTH-21st Project took measurements in triplicate at each visit to ensure accurate ultrasound measurements that could be used to create international fetal standards [24]. Models based on the average of the triplicate measurements can be modified to reflect what is usually done in clinical practise by transforming the average of each triplicate to a single value [209].

Royston and Altman demonstrated the utility and flexibility of FPs for modelling growth data that is typically nonlinear [94]. I used a combination of FP and multi-level methods to model these data. Multi-level models account for subject-specific variations in growth by allowing subject-specific random effects [210]. I found little variation in the fitted centiles developed by the methods that accounted for the correlation structure of repeat measurements and those that did not. The model based on the cross-sectional data showed the greatest variability, with a maximum difference at the 3^{rd} and 97^{th} centiles of 8 mm. As this model had an 80% smaller sample size than the rest of the models, this result is perhaps unsurprising and cannot be viewed as a shortcoming of the method. The overall results are not unexpected in the context of The INTERGROWTH-21st study, as the measurements were taken following a standardised protocol, 95% of the subjects were measured four to six times, and the frequency of the measurements was independent of previous measurements. The data were collected from a healthy

cohort that was actively monitored and seen about five times during a pregnancy that resulted in a live singleton birth [24, 11]. The homogeneity of this cohort and the small variation in the number of repeat measurements per woman explain the minimal differences observed when using different modelling approaches.

However, there were significant differences in precision between the models. The model based on cross-sectional data which had the smallest sample size showed to be less precise when compared to the other models. Differences were observed in the various formulations of multi-level models too. It is therefore important during model selection to also consider how precisely a particular centile of interest is estimated based on the model choice.

Wade *et al.* [208] reported similar findings from a dataset that exhibited a strong correlation structure between the CD4 counts of the uninfected children of HIV-1 infected women. They explored the effect of incorporating correlations between measurements into their estimates of age-related centiles. They concluded that there was little effect on model choice, fitted centiles, or precision. In another study, Wade *et al.* [192] re-analysed the previously published fetal abdominal circumference and bi-parietal diameter dataset of Kurmanavicius *et al.* [179, 211]. This was a prospective study of pregnant women who were examined routinely three times during pregnancy and every 2-3 weeks if high risk. The original analysis used only the first of the series of measurements made during pregnancy to create charts. Wade *et al.* re-analysed the data by incorporating all of the measurements and found an 8-fold increase in the abdominal circumference and biparietal diameter measurement sample sizes. The centiles originally reported by Kurmanavicius *et al.* [179, 211] and those based on the re-analysis that incorporated a correlation structure between repeat measurements of the same individual were similar. However, Wade *et al.* still cautioned that incorporating the correlation structure is preferable to transforming the data into cross-sectional data, as this

could severely affect centile precision and accuracy.

Selecting one observation at random from a set of measurements leads to data wastage. Only 20% of the INTERGROWTH data were used in Model 3, which transformed the longitudinal data into cross-sectional data. Kurmanavicius *et al.* used only 25% of their data to develop fetal charts [179, 211]. Wade *et al.* randomly selected one observation per individual to create a dataset of independent measurements of CD4 lymphocyte counts that used only 76% of their CD4 count data [208, 212]. Discarding data cannot be justified, considering the effort and resources required to obtain it. There are many effective methods to deal with such data complexities.

During study planning, the overall goal of the study must be determined and an appropriate study design must be chosen to answer the specific questions or hypotheses to avoid either wasted effort or later data wastage. For example, if a study aims to develop references or standards, a well-designed cross-sectional study with data collected specifically for this purpose is sufficient. Model choice is dependent on the study aim and the question one is trying to answer rather than the richness of the data.

5.6 Conclusion

In this chapter, I have demonstrated the application of different statistical methods for using repeated measures data on the FHC dataset collected during the INTERGROWTH-21st FGLS. These methods, are not restricted to fetal data and can be applied to other repeated measures data. The methodology and statistical considerations discussed here have also been applied to other commonly measured fetal dimensions, such as biparietal diameter, occipito-frontal diameter, abdominal circumference, and femur length.

The use of all available data is encouraged as discarding data cannot be justified and

effective methods are available to deal with most data complexities. It is important to determine the main aim of a study from the outset and choose an appropriate study design. These considerations and methodology were key to the development of the recently published international standards for fetal growth [24].

6

Estimation of gestational age in early pregnancy from crown-rump length when gestational age range is truncated

6.1 Background

Fetal ultrasound scanning is considered to be essential part of routine antenatal care. First-trimester scans are recommended for confirming viability, determining the number of fetuses, and accurately estimating GA [35, 36]. A reliable estimate of GA is needed as it underpins clinical care and allows the expected date of delivery to be estimated. GA in pregnancy can be measured using (a) a reliable first day of the LMP alone, (b) CRL measured using an early (9^{+0} to 13^{+6} weeks) ultrasound alone, (c) LMP and ultrasound combined, (d) the height of the uterine fundus, or (e) methods for estimating the timing of ovulation based on changes in basal body temperature and alteration in sex hormone profiles. The implications of these

methods on research findings have recently been discussed [213, 28]. Ultrasound can accurately determine the day of conception to within 5 days either way in 95% of cases and may be on average 2–3 days more accurate than LMP in predicting the date of a spontaneous delivery [35, 78, 214, 215, 39, 216].

Routine ultrasound measurements of fetal biometry are now common in many settings and form the basis for GA estimation in clinical practice to date pregnancies, either alone or in combination with LMP. It is not possible to measure the time of conception precisely in spontaneously conceived pregnancies. LMP is widely used as a proxy indicator for pregnancy dating and is based on the assumption that pregnancy has a constant duration from the first day of the LMP, with ovulation on the 14^{th} day [34]. This method of dating pregnancies, has been shown to be unreliable, even for women with a certain menstrual history [32, 33]. The LMP approach is often criticised for its susceptibility to recall bias. It is dependent on a woman's ability to recall her dates. Accurate recall rates vary from as high as 79% [217] to as low as 32% [218]. Caution is recommended when using LMP alone for dating because up to 50% of women are uncertain of their dates, have an irregular cycle, have recently stopped taking the oral contraceptive pill, are lactating, or did not have a normal LMP [31].

Assessment of GA based on ultrasound biometry was first introduced in 1969 by Campbell [219], who used biparietal diameter. Ultrasound biometry has become the preferred method for dating pregnancy. In 1973, Robinson and colleagues developed pregnancy dating charts using CRL [220]. CRL is now widely accepted as the method of choice for GA of a fetus. Ultrasound measured GA assumes that fetal growth in early pregnancy is sufficiently uniform within a population that the length of a fetus can be used as a proxy for GA [221]. The extent to which this assumption is justified depends on: (a) the GA at which the biometry is measured, (b) the choice of fetal biometry (e.g., CRL, height of the uterine fundus), and (c)

a variety of technical factors related to image acquisition, sonographer skill, and image quality [27]. It is important to acknowledge that there may be considerable variation in the methods used to obtain apparently similar findings.

Robinson's formula has been validated by subsequent studies in a range of settings [27] using high-resolution ultrasound technology, including data acquired transvaginally [222]. On the basis of such findings, ultrasound is now widely accepted as a more accurate predictor of GA than LMP and is recommended as the standard for dating pregnancies throughout the developed world [48, 223, 222].

In the UK, the National Institute for Health and Care Excellence Guideline for Routine Antenatal Care (2008) and the International Society of Ultrasound in Obstetrics and Gynaecology recommend that all pregnant women be offered an early ultrasound examination to date pregnancies [35, 31, 224]. This should ideally be performed by measuring CRL between 10 and 13^{+6} weeks, which can reduce the need for induction of labour after 41 weeks of gestation. Although there is always a margin of error in ultrasound-based estimates [225], this error is relatively small compared with LMP-based estimates [224, 226]. The clinical consequences of this error are unclear and it is generally accepted that any error associated with ultrasound is less than that with LMP or other dating methods.

Since the validation of Robinson's formula, many pregnancy dating charts using CRL have been developed and are in use. However, they have been developed using different populations, resulting in discrepancies when they are compared or applied to a specific population. This has led to the need for an international reference dating equation and chart [28, 213, 27, 227, 47]. The INTERGROWTH-21^{st} Project aimed to generate fetal growth charts and a new dating chart. GA was based on first day of the LMP and corroborated with CRL using a known dating equation [43]. Only women between 9^{+0}–13^{+6} weeks gestation whose estimates from the two methods agreed within 7 days were recruited into the FGLS component

of the INTERGROWTH-21st Project.

6.2 Problem statement

CRL is modelled as a function of GA when developing charts of fetal size. For dating charts, the variables are interchanged so that GA is modelled as a function of CRL. McLennan and Schluter illustrated [46] the difference between the two concepts using scatter plots derived from the same population. They reported CRL (the independent variable) against GA to represent the regression analysis fitting model for deriving the equation of GA estimation. They then reported GA (the independent variable) against CRL for a size chart [78]. Sahota *et al.* [47] elegantly demonstrated how the assessment of size and maturity should not be considered to be interchangeable, as simply flipping a regression can lead to an over- or underestimation of GA, especially at the extremes of the CRL range.

Regression modelling to derive an equation for GA as a function of CRL is problematic if the available data are constrained by a restricted GA range [92]. One aim of the longitudinal study of the INTERGROWTH-21st Project was to develop a new GA estimation equation based on the CRL from women recruited between 9^{+0} and 13^{+6} weeks. The CRL data was truncated at both ends (i.e. at 9^{+0} and 13^{+6} weeks) which was a challenge in terms of modelling given the need for obtaining estimates in the truncated regions. Ignoring the GA truncation would lead to heavily biased estimates of GA. This restriction was part of the design of the INTERGROWTH-21st study, as CRL measurements are less reliable outside this GA range [35, 31, 40, 41, 42]. Restriction commonly exists, as fetal curling prevents accurate measurement of CRL after 13^{+6} weeks.

Other examples in the literature of truncated measurements of interest originate from military conscription registers. Conscription was introduced on a large scale throughout Europe at the beginning of the 19th century. It was common for most

armies to only admit conscripts whose height exceeded a certain threshold. As the height of those not meeting the height threshold was rarely recorded, the shape of the left tail of the height distribution was unknown. The problem is illustrated in Figure 6.1, where the height distribution of a typical cohort of conscripts is plotted. All observations below the truncation point, 157 cm in the plot, are not available [228]. The truncation of the distribution has important implications for estimating our parameter of interest (the mean) as the left tail of the distribution may contain considerable probability mass.

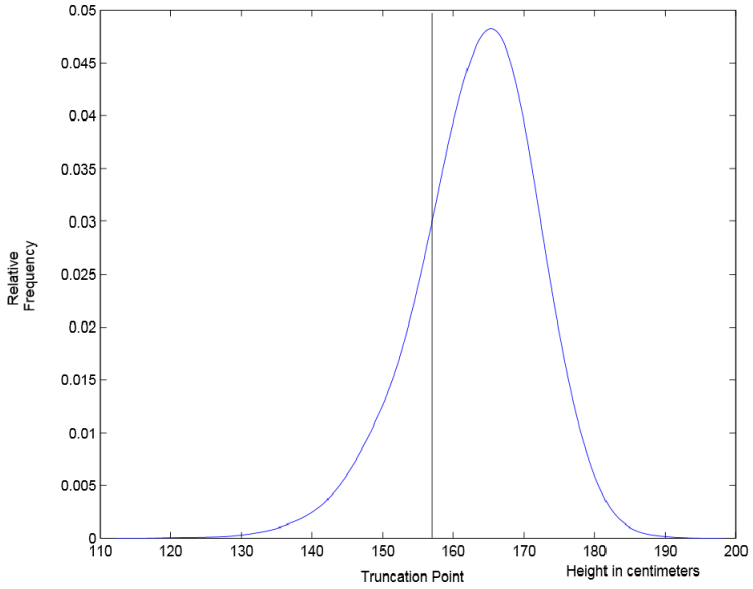


Figure 6.1: Height distribution and truncation

The aim of this chapter is to explore strategies to overcome GA truncation when developing equations and charts for dating pregnancies from CRL measurements. I use the CRL data from the INTERGROWTH-21st Project to investigate this. I explore three statistical approaches to overcome the truncation of GA. To evaluate these strategies, I generate a dataset with no GA truncation that is similar to the INTERGROWTH-21st CRL dataset. I use the simulated data to explore the performance of different methods of analysis by imposing truncation at 9 and 14 weeks of gestation. The three methods are first tested in a simulation-based

study using a previously published dating equation from Verburg *et al.* [36]: the performance of each test model is compared with the performance of the model that generated the data. The best approach is then tested on a sample of the INTERGROWTH-21st data to estimate GA from CRL.

6.3 Methodology

Several reliable statistical methods already exist for developing age-related reference centiles [67, 92, 163]. These can be applied in a straightforward way for developing equations for fetal size as a function of GA. However, dating requires estimating GA as a function of fetal size, measured with fetal CRL. The INTERGROWTH-21st Project is the largest prospective study to collect good quality data on CRL in geographically diverse populations to date and with a high level of quality control measures in place. I used the INTERGROWTH-21st data to develop centiles for the distribution of GA for CRL values between 15 and 100 mm, which was the range of measurements for the CRL data. The statistical challenge was this: How should the data be modelled when the outcome variable (GA) was truncated at both ends (i.e. at 9^{+0} and 13^{+6} weeks)?.

I explored three statistical approaches to overcome the GA truncation. I generated a dataset with no GA truncation that was similar to the INTERGROWTH-21st Project CRL data and used it to evaluate the performance of the three methods after imposing truncation at 9^{+0} and 13^{+6} weeks of gestation.

6.4 Statistical methods

Data were explored visually with scatter plots of CRL by GA and vice versa. The relationship between GA and CRL is nonlinear, although the distribution of CRL is conditionally normal at any given GA. In contrast, GA has a positively skewed distribution for a given CRL [78]. I applied FP models by fitting separate models

to the mean and SD of GA to account for the increase in variance with greater CRL and gestation. I used FP models that have been shown to fit fetal data very well, are very flexible, and extrapolate well [67, 163]. Equations of the mean and SD allow any desired centiles and z-scores to be calculated:

$$p^{th}\ centile = Median\ CRL + (K \times SD), \tag{6.1}$$

where K is the normal equivalent deviate (z-score) corresponding to a particular centile (e.g., K = 1.88 for the 97^{th} centile and -1.88 for the 3^{rd} centile) and SD is the predicted estimate from the regression analysis.

Fitted curves for the $3^{rd}, 50^{th}$, and 97^{th} centiles from each model were assessed visually and by comparing deviances to check for a good fit. The choice of centiles presented was based on what is commonly reported in the literature and used in clinical practice as standard centiles. The INTERGROWTH-21^{st} Project aimed to complement the WHO-MGRS, which produced reference standards for children aged 0–5 years and presented the 3^{rd} and 97^{th} centiles [43]. Goodness of fit was assessed with a scatter plot of the distribution of residuals in z-scores by CRL and by counting the number of observations below the 3^{rd} and above the 97^{th} centiles.

The three approaches explored to deal with GA truncation at 9 and 14 weeks were: (a) simulation, restriction and extrapolation; (b) simulation; and (c) inversion of the model for predicting CRL from GA. Extrapolation was applied to obtain reliable estimates between 9 and 14 weeks in the presence of truncation before 9 weeks and after 14 weeks. The resultant equation cannot be used for dating before 9 weeks or after 14 weeks as this is not recommended in clinical practice. The reliability of FP models for extrapolation has been discussed previously by Royston and Altman who showed that these models extrapolate well for fetal measurements [163]. The three methods were first tested in a simulation-based study using a previously published dating equation by Verburg *et al.* [36]. I evaluated how well each of the three

approaches performed compared with the model that generated the data.

The Verburg *et al.* equation was selected from the many dating equations in use as it is one of the four preferred dating equations according to a recent systematic review of the methodology used for creating dating charts [27]. It is also recommended by the International Society of Ultrasound in Obstetrics and Gynaecology. The great strength of performing a simulation study based on a known dating equation is that the performance of the proposed methods can be evaluated in a situation where the 'truth' is known (i.e., the equations from which the simulated data were obtained). After evaluating the three approaches using simulated data based on the Verburg *et al.* equations, the best approach was applied to the INTERGROWTH-21st Project data to estimate GA from CRL.

Data were simulated from the Verburg *et al.* dating equation [36]:

$$Mean\ of\ log\ GA = 1.4653 + 0.001737 \times CRL + 0.2313 \times log\ CRL, \tag{6.2}$$

$$SD\ of\ log\ GA = 0.04590 \tag{6.3}$$

Here and throughout, all logarithms are natural logarithms.

Equations 6.2 and 6.3 assume that log GA has a normal distribution for any value of CRL. From these equations, I simulated 100 observations of GA for each CRL value from 5 mm to 110 mm (the range of CRL measurements in the INTERGROWTH-21st Project) in 1 mm increments, resulting in 10,600 observations. A sample size of 100 was chosen as it represented the average number of CRL observations in each complete week in the INTERGROWTH-21st data and was large enough to remove sampling variation effects. The GA was between 5 and 17 weeks which is the GA range of the original Verburg data from which the equations were obtained. GA was log-transformed in all of the analyses to stabilise variance [36, 47, 67, 93].

6.5 Validation of the simulated data

I modelled the simulated data using FP regression of log-transformed GA on CRL and compared the FP terms and predicted median GA from the obtained equation to the original dating equation reported by Verburg *et al.*. The equations obtained from simulated data were remarkably similar to the original Verburg *et al.* equations:

$$Mean\ of\ log\ GA = 1.4612 + 0.001693 \times CRL + 0.2332 \times log\ CRL, \quad (6.4)$$

$$SD\ of\ log\ GA = 0.0458114 - 0.00000198 \times CRL, \quad (6.5)$$

Both equations for the median were FP models of degree 2 with powers 0 and 1 (i.e. terms in CRL and log CRL). The equation for SD was an FP model of degree 1, power 1 (linear), whereas the SD obtained by Verburg *et al.* was a constant. The predicted GA from the two equations agreed within 0.08 days (Figure 6.2, Table 6.1).

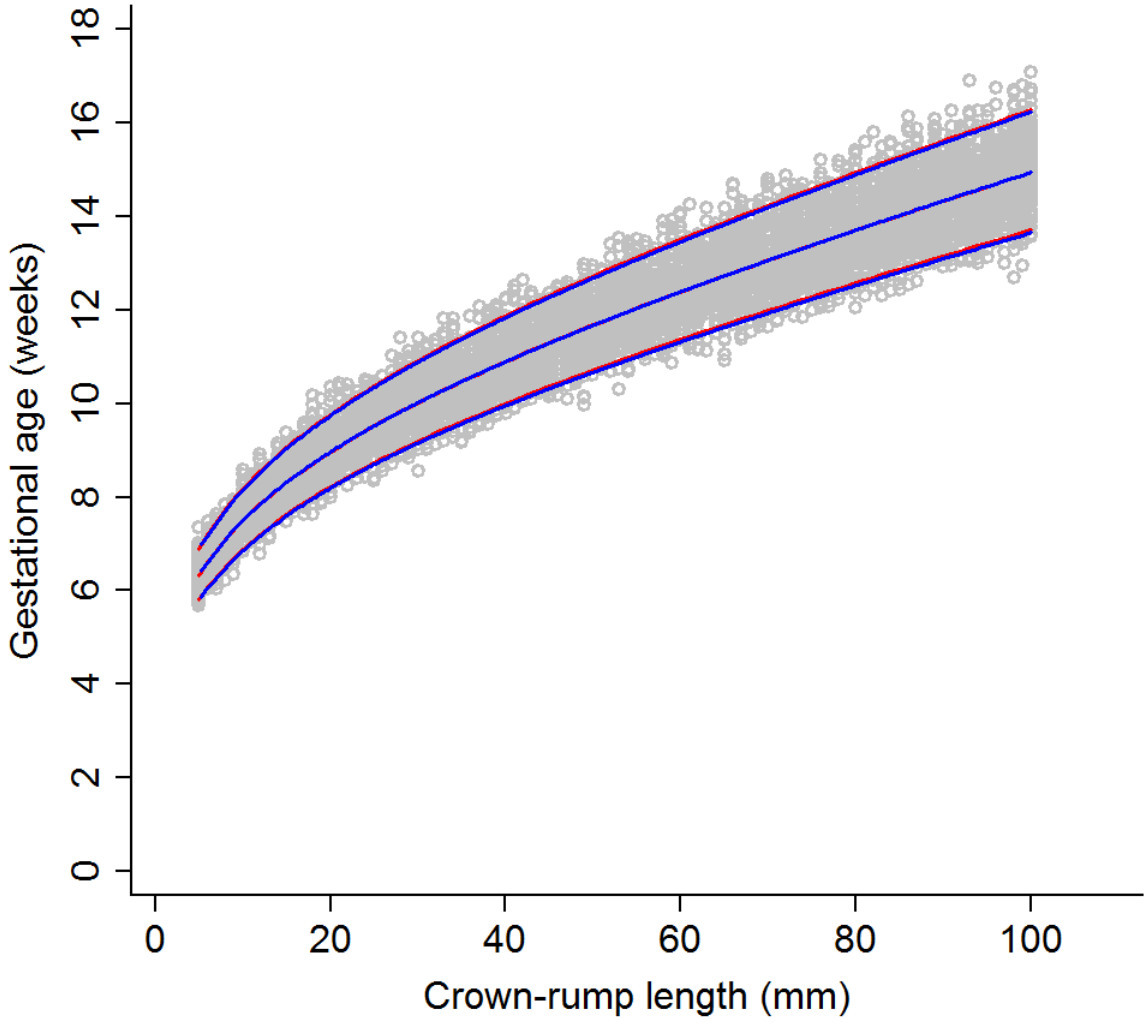


Figure 6.2: Simulated data for crown-rump length measurements in relation to gestational age (grey circles) with 3^{rd} and 97^{th} fitted centiles. Blue continuous lines represent the original equation fit reported by Verburg *et al.* [36] and from which the simulated data were derived, whereas the red continuous lines represent model fit of the simulated data.

CRL (mm)	Original Verburg *et al.* equation Median GA (weeks) predicted from CRL	Equation from the simulated data Median GA (weeks) predicted from CRL	Difference GA (days)
5	6.336	6.324	0.082
10	7.503	7.497	0.041
15	8.312	8.310	0.015
20	8.962	8.962	-0.003
25	9.519	9.521	-0.017
30	10.015	10.019	-0.026
35	10.469	10.474	-0.032
40	10.892	10.897	-0.035
45	11.290	11.296	-0.036
50	11.670	11.675	-0.036
55	12.034	12.039	-0.033
60	12.386	12.390	-0.029
65	12.727	12.731	-0.023
70	13.060	13.063	-0.016
75	13.386	13.387	-0.008
80	13.706	13.706	0.001
85	14.021	14.019	0.012
90	14.331	14.328	0.023
95	14.638	14.633	0.036
100	14.942	14.935	0.050

Table 6.1: Crown-rump length (CRL) measurements in relation to gestational age (GA) for the original equation fit reported by Verburg *et al.* [36] compared with the model fit of the simulated data.

After successful validation of the simulated data, I truncated GA at 9 and 14 weeks to match the INTERGROWTH 21^{st} data set. As discussed previously, truncation is only a problem when modelling GA as a function of CRL and not when modelling CRL as a function of GA (size chart) (Figure 6.3). All three suggested approaches make use of this fact, but in different ways. I applied the three proposed approaches to the truncated simulated data shown in Figure 6.3. Figure 6.4 shows a flow diagram summarising the three methods.

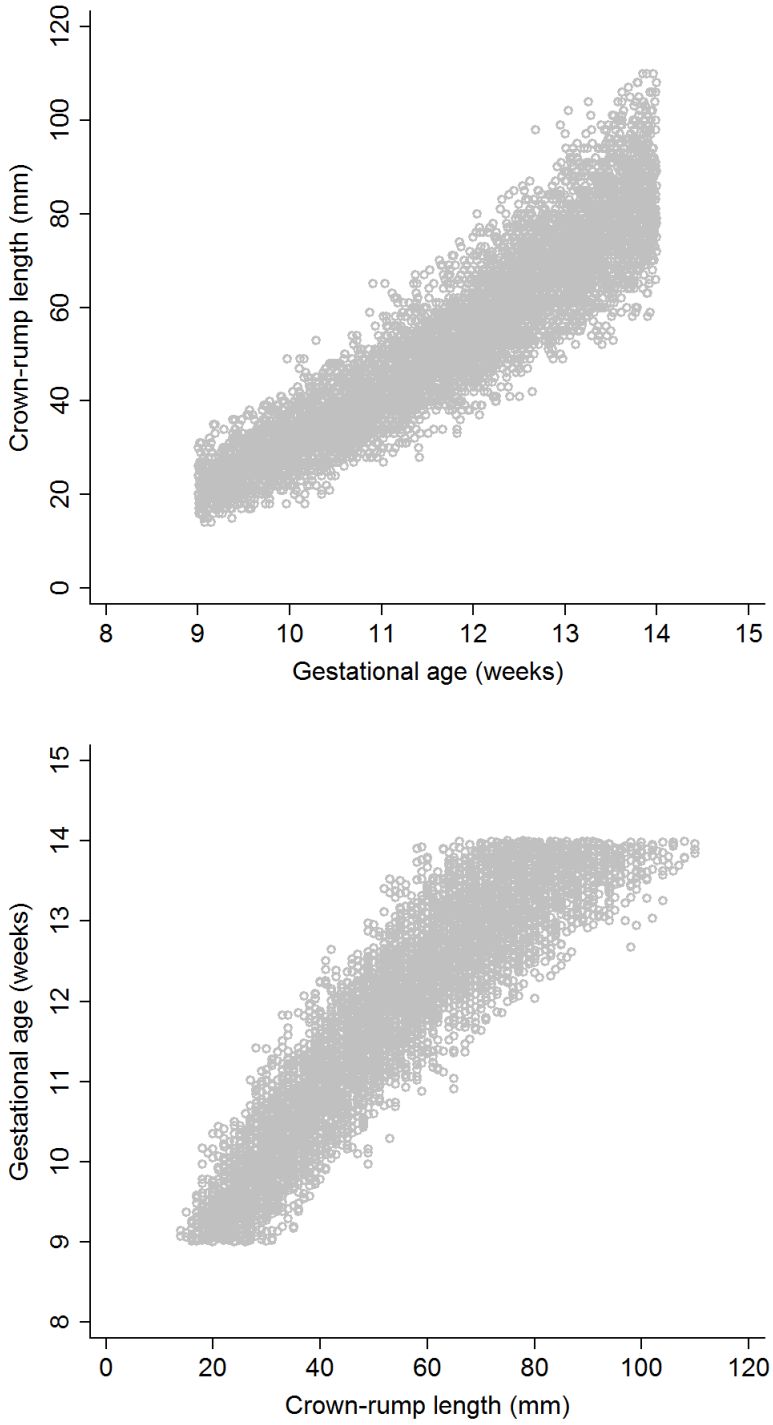


Figure 6.3: Simulated data generated from the dating equation by Verburg *et al.* and truncated at 9 and 14 weeks. Top figure shows crown-rump length versus gestational age for creating a size chart and bottom figure shows gestational age versus crown-rump length for creating a dating chart.

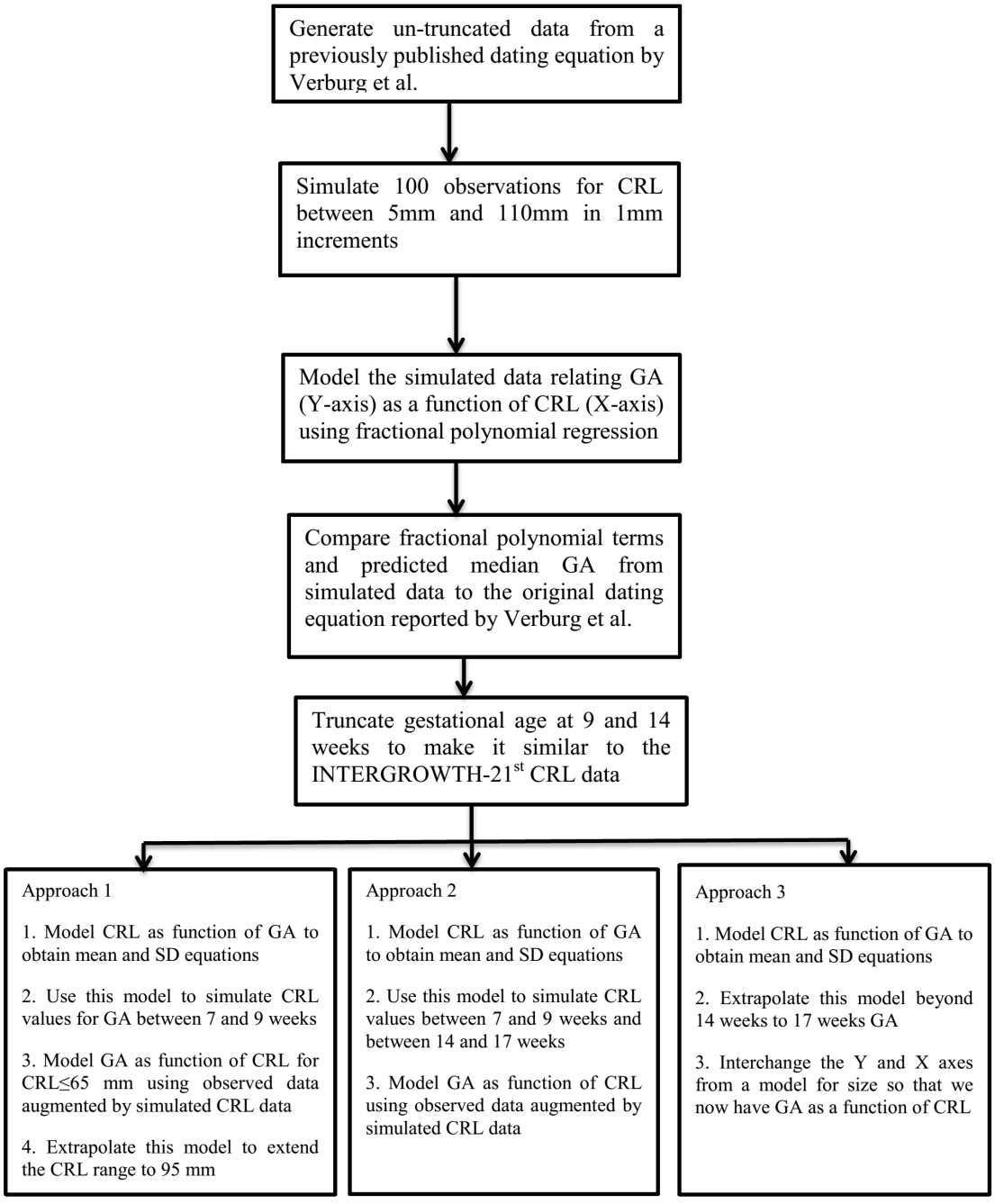


Figure 6.4: A flow diagram summarising the process and methodology of the simulation study to evaluate three methods to overcome the truncation problem inherent in the data set.

6.6 Approach 1: Simulation for small crown-rump length, restriction and extrapolation

The simulation, restriction, and extrapolation approach was based on modelling CRL as a function of GA (Figure 6.5, top figure). From the obtained equation of the median GA, I simulated 100 CRL observations for each day of gestation between 7 and 9 weeks, to overcome the truncation at the bottom end of the distribution of CRL measurements. There were also 100 observations for each day of GA in the untruncated dataset. The choice of 7 weeks as a lower limit for extrapolation was based on the desire to obtain a good fit to the data at 9 weeks, where the actual data was truncated. It was also the lowest limit at which the fitted equations and range of GA remained plausible when extrapolated. Using the augmented dataset, I modelled GA as a function of CRL, with CRL restricted to lowest CRL measurement reported at 14 weeks in the INTERGROWTH-21st data set, 65 mm, as there remained a truncation problem at the upper end of the CRL distribution (Figure 6.5, middle figure). I then extrapolated the obtained mean and SD equations to the rest of the data (Figure 6.5, bottom figure).

The predicted GA from this approach was compared with the GA reported by Verburg *et al.* (Table 6.2). A sensitivity analysis was performed to establish whether truncating CRL at a lower cut-off of 10 mm, 15 mm or 20 mm would give the best prediction. It was performed by comparing the predicted GA obtained using the derived equation to that reported by Verburg *et al.*. The choice of a cut-off affects the fit for large CRLs and so has clinical implications, because it is desirable to be able to predict GA from CRL between 15 mm and 95 mm (Table 6.2).

Summary of Approach 1

1. Model CRL as a function of GA to obtain the mean and SD equations
2. Use this model to simulate CRL values for GA between 7 and 9 weeks
3. Model GA as function of CRL for CRL $\leq$ 65 mm using observed data augmented by simulated CRL data
4. Extrapolate this model to extend the CRL range to 95 mm

Fitting the Approach 1 model to the simulated data gave the following equations:

$$Mean\ log\ GA = 0.81350 + 3.62375 \times CRL^{-1} + 0.40202 \times log\ CRL, \quad (6.6)$$

$$SD\ of\ log\ GA = 0.04926 - 0.0000946 \times CRL, \quad (6.7)$$

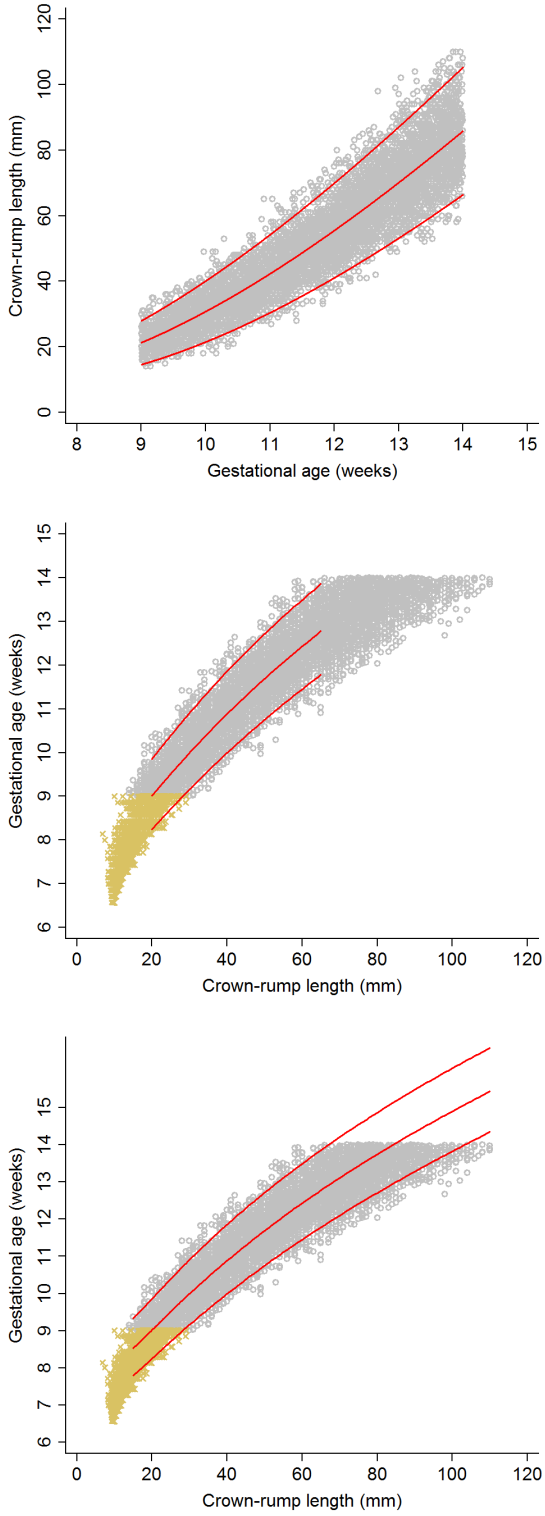


Figure 6.5: Crown-rump length measurements in relation to gestational age (grey circles) with 3^{rd}, 50^{th}, and 97^{th} fitted centiles (top figure). Yellow small crosses in the middle and bottom figures represent data simulated from the fitted equation of the mean and standard deviation from the top figure. The middle figure shows the model fit relating gestational age and crown-rump length with crown-rump length restricted to ≤ 65 mm and the bottom figure shows the model fit from the middle figure extrapolated to the full range of crown-rump length (Approach 1).

	Original Verburg *et al.* equation Estimated GA (weeks)			Approach 1 Estimated GA (weeks)			Difference (days)		
CRL (mm)	3^{rd} centile	Median	97^{th} centile	3^{rd} centile	Median	97^{th} centile	3^{rd} centile	Median	97^{th} centile
10	6.88	7.50	8.18	6.85	8.18	8.22	0.21	-4.76	-0.28
15	7.63	8.31	9.06	7.60	8.53	9.09	0.21	-1.54	-0.21
20	8.22	8.96	9.77	8.20	9.02	9.80	0.14	-0.42	-0.21
25	8.73	9.52	10.38	8.72	9.51	10.40	0.07	0.07	-0.14
30	9.19	10.02	10.92	9.18	9.99	10.93	0.07	0.21	- 0.07
35	9.60	10.47	11.41	9.60	10.45	11.41	0.00	0.14	0.00
40	9.99	10.89	11.87	10.00	10.88	11.86	-0.07	0.07	0.07
45	10.36	11.29	12.31	10.37	11.30	12.29	-0.07	-0.07	0.14
50	10.70	11.67	12.72	10.73	11.69	12.69	-0.21	-0.14	0.21
55	11.04	12.03	13.12	11.08	12.07	13.07	-0.28	-0.28	0.35
60	11.36	12.39	13.50	11.41	12.43	13.44	-0.35	-0.28	0.42
65	11.67	12.73	13.87	11.74	12.77	13.80	-0.49	-0.28	0.49
70	11.98	13.06	14.24	12.05	13.11	14.15	-0.49	-0.35	0.63
75	12.28	13.39	14.59	12.37	13.43	14.49	-0.63	-0.28	0.70
80	12.57	13.71	14.94	12.67	13.74	14.82	-0.70	-0.21	0.84
85	12.86	14.02	15.28	12.98	14.04	15.15	-0.84	-0.14	0.91
90	13.15	14.33	15.62	13.27	14.34	15.47	-0.84	-0.07	1.05
95	13.43	14.64	15.96	13.57	14.62	15.79	-0.98	0.14	1.19
100	13.71	14.94	16.29	13.86	14.90	16.10	-1.05	0.28	1.33

Table 6.2: Crown-rump length (CRL) measurements in relation to gestational age (GA) for the original equation fit reported by Verburg *et al.* compared to the model fit of the simulated data (Approach 1).

6.7 Approach 2: Simulation for small and large crown-rump length

Approach 2 was very similar to Approach 1, as data were again simulated from fitting a size equation and using the mean and SD equations of CRL by log GA (Figure 6.6, top figure). I used the model for CRL to simulate 100 observations of CRL for each day of gestation at both ends of the distribution, which was about the same number of observations for each day of GA in the untruncated dataset. Data were simulated for below 9 weeks (between 7 and 9 weeks) and above 14 weeks (between 14 and 17 weeks) of gestation (Figure 6.6, middle figure). I chose a lower limit of 7 weeks and an upper limit of 17 weeks so that a good fit to the data between 9 and 14 weeks, where the actual data were truncated, could be obtained. The 7 and 17 weeks cut-offs were also the limits at which the fitted equations and GA remained plausible when extrapolated.

The simulated CRL measurements below 9 weeks and above 14 weeks overcame the truncation problem presented by the data, allowing GA to be modelled as a function of CRL more efficiently and the respective median and SD equations to be obtained (Figure 6.6, bottom figure). The predicted GA from this approach was compared with the Verburg *et al.* reported GA (Table 6.3). A sensitivity analysis assessment was performed on the value of the lower CRL cut-off.

Summary of Approach 2

1. Model CRL as a function of GA to obtain the mean and SD equations
2. Use this model to simulate CRL values below 9 weeks (GA between 7 and 9 weeks) and above 14 weeks (GA between 14 and 17 weeks)
3. Model GA as function of CRL using observed data augmented by simulated CRL data

Fitting the Approach 2 model to the simulated data gave the following equations:

$$Mean\ log\ GA = 0.595705 + 1.507363 \times CRL^{-0.5} + 0.421304 \times log\ CRL \quad (6.8)$$

$$SD\ of\ log\ GA = 0.044991 \quad (6.9)$$

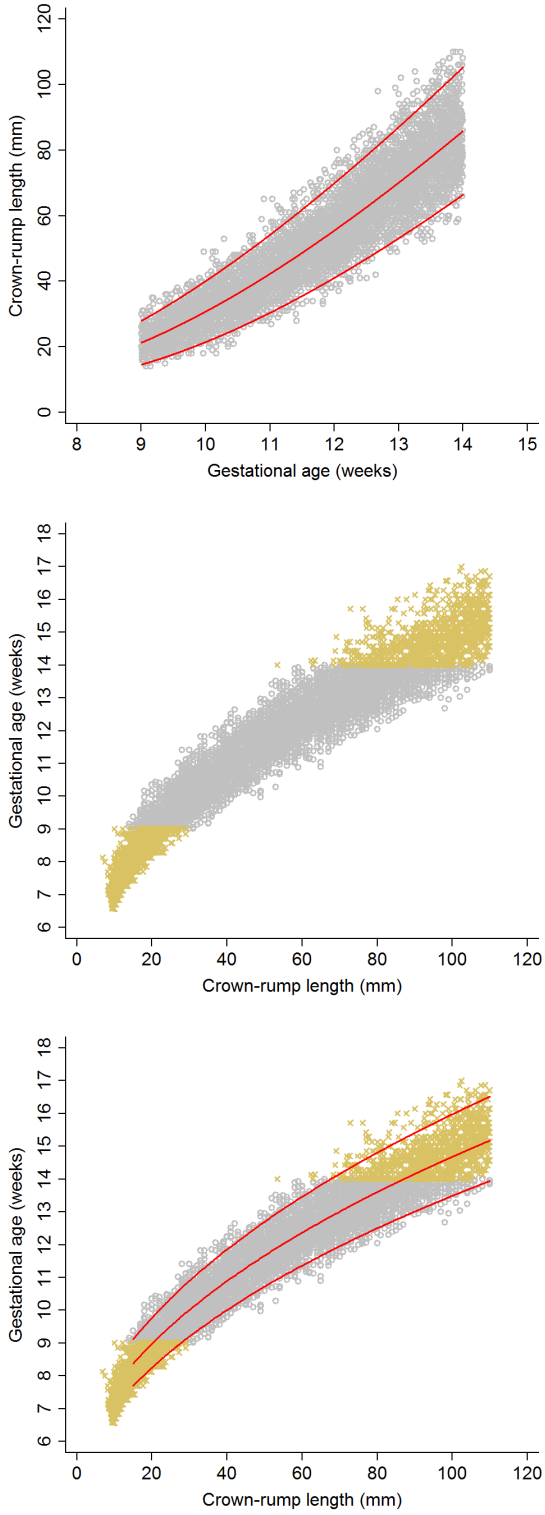


Figure 6.6: Crown-rump length measurements in relation to gestational age (grey circles) with 3^{rd}, 50^{th}, and 97^{th} fitted centiles (top figure). Yellow small crosses in the middle and bottom figures represent data simulated from the fitted equations of the mean and standard deviation from the top figure. The bottom figure shows the model fit relating gestational age and crown-rump length (Approach 2).

	Original Verburg *et al.* equation Estimated GA (weeks)			Approach 2 Estimated GA (weeks)			Difference (days)		
CRL (mm)	3^{rd} centile	Median	97^{th} centile	3^{rd} centile	Median	97^{th} centile	3^{rd} centile	Median	97^{th} centile
10	6.88	7.50	8.18	7.08	7.71	8.39	-1.41	-1.45	-1.48
15	7.63	8.31	9.06	7.70	8.38	9.12	-0.52	-0.47	-0.41
20	8.22	8.96	9.77	8.25	8.98	9.77	-0.21	-0.12	-0.02
25	8.73	9.52	10.38	8.75	9.52	10.36	-0.11	0.00	0.12
30	9.19	10.02	10.92	9.20	10.01	10.90	-0.09	0.02	0.15
35	9.60	10.47	11.41	9.62	10.47	11.39	-0.11	0.01	0.14
40	9.99	10.89	11.87	10.01	10.89	11.86	-0.13	-0.01	0.13
45	10.36	11.29	12.31	10.38	11.29	12.29	-0.14	-0.02	0.12
50	10.70	11.67	12.72	10.72	11.67	12.70	-0.13	0.00	0.15
55	11.04	12.03	13.12	11.05	12.03	13.09	-0.09	0.04	0.20
60	11.36	12.39	13.50	11.37	12.37	13.46	-0.03	0.11	0.28
65	11.67	12.73	13.87	11.67	12.70	13.82	0.06	0.21	0.39
70	11.98	13.06	14.24	11.96	13.01	14.16	0.18	0.35	0.54
75	12.28	13.39	14.59	12.23	13.31	14.49	0.33	0.51	0.73
80	12.57	13.71	14.94	12.50	13.60	14.81	0.50	0.71	0.95
85	12.86	14.02	15.28	12.76	13.89	15.11	0.71	0.94	1.20
90	13.15	14.33	15.62	13.01	14.16	15.41	0.95	1.20	1.49
95	13.43	14.64	15.96	13.25	14.42	15.70	1.22	1.50	1.81
100	13.71	14.94	16.29	13.49	14.68	15.98	1.51	1.82	2.17

Table 6.3: Crown-rump length (CRL) measurements in relation to gestational age (GA) for the original equation fit reported by Verburg *et al.* compared to the model fit of the simulated data (Approach 2).

6.8 Approach 3: Interchanging the X- and Y-axes from a model for size

The third approach did not require data simulation. As before, I modelled CRL (the Y-axis) as a function of GA (the X-axis) using all of the available data. I extrapolated the obtained equations to larger GAs to cover the desired CRL range (Figure 6.7, top figure) and interchanged the X- and Y-axes to give GA (the Y-axis) as a function of CRL (the X-axis) (Figure 6.7, middle figure). There were no equations for the median and SD describing the relationship between GA to CRL. Instead, three sets of X, Y coordinates of GA gave the predicted 3^{rd}, 50^{th}, and 97^{th} centiles for CRL. A new equation for the median was obtained by regressing GA on the predicted median CRL. Similarly, equations for the 3^{rd} and 97^{th} centiles were obtained (Figure 6.7, bottom figure). The predicted GA from this approach was compared with that originally reported by Verburg *et al.* (Table 6.4). As there was no equation for the SD, the full model cannot be written down simply. I will describe how to obtain an equation for the SD as a function of CRL that also allows any desired centiles to be predicted.

6.8.1 Computing an equation for the standard deviation

The equations for the 3^{rd}, 50^{th} and 97^{th} centiles that relate log GA and CRL can be used to estimate the SD. Two estimates of the SD can be obtained at a given CRL from the difference between 97^{th} and 50^{th} centiles and the difference between the 50^{th} and 3^{rd} centiles. The two centiles will not be exactly the same but will be very similar because GA was modelled on the log scale. It is thus reasonable to estimate the SD for each value of CRL by simply taking the average of the two SDs. An equation for SD relating GA to CRL was obtained by regressing the estimated SD

of GA on CRL. Estimates of any desired centiles can then be obtained using:

$$p^{th}\ centile = Median\ CRL + (K \times SD), \tag{6.10}$$

where K is the normal equivalent deviate (z-score) corresponding to a particular centile (e.g., K = 1.88 for the 97^{th} centile and -1.88 for the 3^{rd} centile) and SD is the predicted estimate from the regression analysis just described.

Summary of Approach 3

1. Model CRL as a function of GA to obtain the mean and SD equations
2. Extrapolate this model beyond 14 weeks to 17 weeks GA
3. Interchange the Y- and X-axes from a model for size so that the plot becomes GA (Y-axis) as a function of CRL (X-axis)

Fitting the Approach 3 model to the simulated data gave the following equations:

$$Mean\ log\ GA = 2.12294 - 1.11877 \times CRL^{-0.5} + 0.068754 \times CRL^{0.5}, \tag{6.11}$$

$$SD\ of\ log\ GA = 0.0000063 - 0.00000011 \times CRL, \tag{6.12}$$

I have shown that these rather *ad hoc* approaches corresponded very closely to the real-world data reported by Verburg *et al.* (Figure 6.8). The Verburg *et al.* dataset has similarities to the INTERGROWTH 21^{st} Project CRL dataset (Figure 6.9). Figure 6.9 shows data from 1600 fetuses ($\backsim$35% of the overall target sample) included in INTERGROWTH 21^{st} study, in the same format as Figure 6.3. The close similarity between the two datasets is apparent.

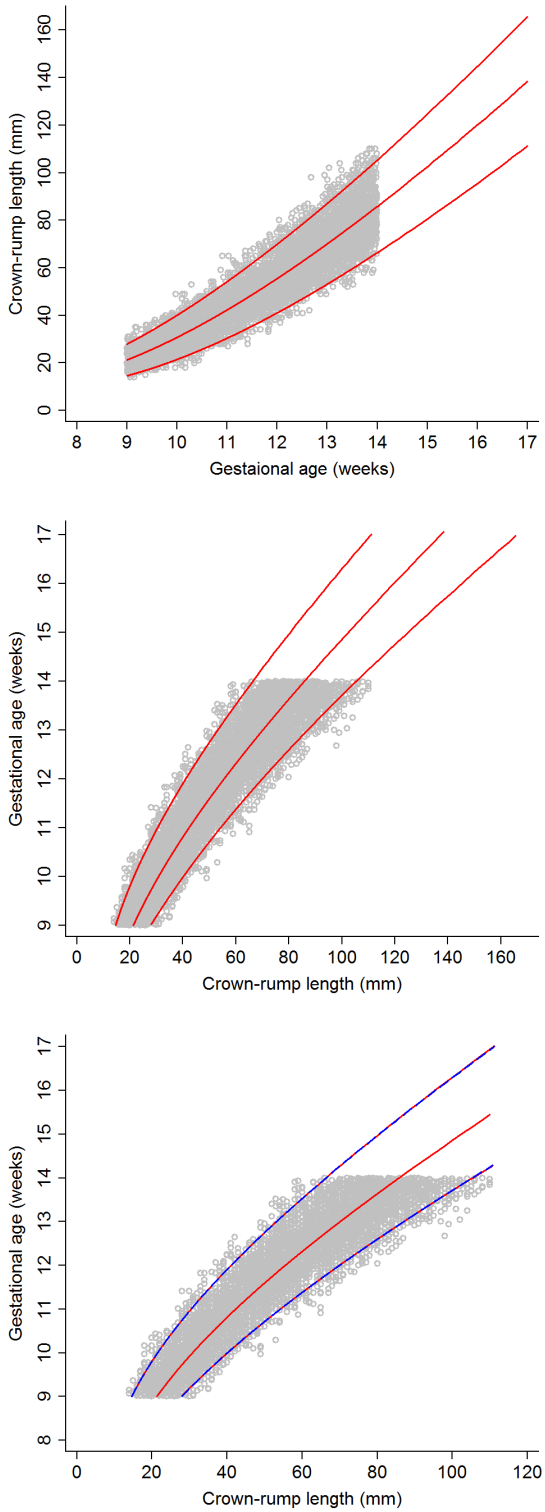


Figure 6.7: Crown-rump length (CRL) measurements in relation to gestational age (GA) (grey circles) with 3^{rd}, 50^{th}, and 97^{th} fitted centiles (top figure). The middle figure shows the relationship between GA and CRL after interchanging the axes and fitting new models to the three sets of coordinates. The bottom figure shows the model obtained by simply taking the average of two standard deviations (SDs). An equation for the SD relating GA to CRL can then be obtained by regressing this SD of GA on CRL and estimating outer centiles by combining the model for SD with that for the median (Approach 3).

	Original Verburg *et al.* equation Estimated GA (weeks)			Approach 3 Estimated GA (weeks)			Difference (days)		
CRL (mm)	3^{rd} centile	Median	97^{th} centile	3^{rd} centile	Median	97^{th} centile	3^{rd} centile	Median	97^{th} centile
10	6.88	7.50	8.18	6.97	7.29	8.15	-0.60	1.49	0.23
15	7.63	8.31	9.06	7.66	8.17	9.08	-0.26	1.00	-0.14
20	8.22	8.96	9.77	8.23	8.85	9.81	-0.10	0.79	-0.25
25	8.73	9.52	10.38	8.73	9.42	10.42	-0.02	0.68	-0.29
30	9.19	10.02	10.92	9.19	9.93	10.96	0.00	0.62	-0.29
35	9.60	10.47	11.41	9.60	10.39	11.45	0.00	0.57	-0.28
40	9.99	10.89	11.87	9.99	10.81	11.91	-0.02	0.54	-0.26
45	10.36	11.29	12.31	10.36	11.22	12.34	-0.05	0.52	-0.24
50	10.70	11.67	12.72	10.72	11.60	12.75	-0.08	0.50	-0.22
55	11.04	12.03	13.12	11.05	11.96	13.15	-0.11	0.48	-0.20
60	11.36	12.39	13.50	11.38	12.32	13.53	-0.14	0.47	-0.18
65	11.67	12.73	13.87	11.70	12.66	13.90	-0.16	0.47	-0.15
70	11.98	13.06	14.24	12.01	12.99	14.25	-0.17	0.47	-0.12
75	12.28	13.39	14.59	12.30	13.32	14.61	-0.18	0.47	-0.09
80	12.57	13.71	14.94	12.60	13.64	14.95	-0.18	0.48	-0.05
85	12.86	14.02	15.28	12.88	13.95	15.29	-0.16	0.49	-0.01
90	13.15	14.33	15.62	13.17	14.26	15.62	-0.14	0.52	0.04
95	13.43	14.64	15.96	13.44	14.56	15.94	-0.11	0.55	0.10
100	13.71	14.94	16.29	13.72	14.86	16.27	-0.06	0.58	0.16

Table 6.4: Crown-rump length (CRL) measurements in relation to gestational age (GA) for the original equation fit reported by Verburg *et al.* compared to the model fit of the simulated data (Approach 3).

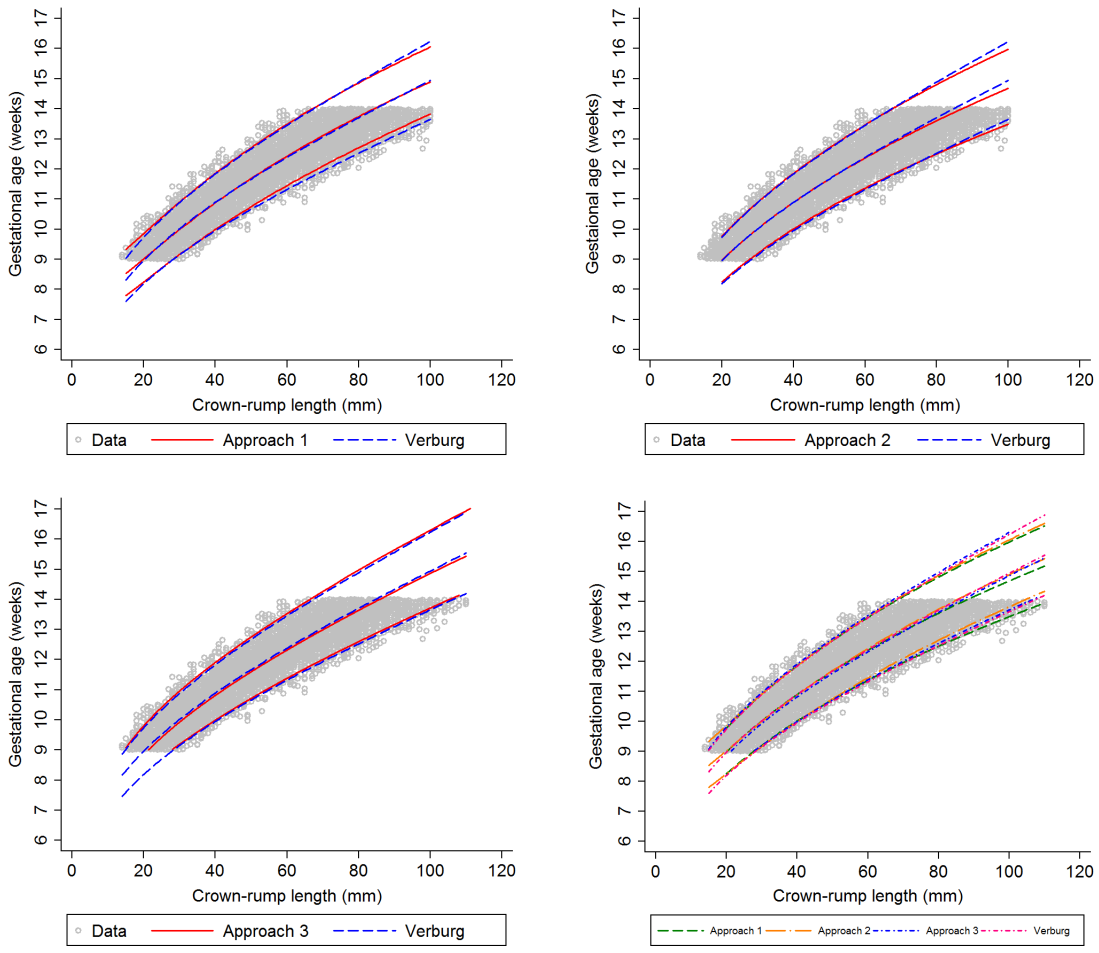


Figure 6.8: Crown-rump length measurements in relation to gestational age for the simulated data for crown-rump length from 9^{+0} to 13^{+6} weeks gestational age, comparing each of the three approaches with Verburg *et al.* (top left, top right, and bottom left figures) and all three approaches with Verburg *et al.* (bottom right figure).

Results from the simulation study using truncated data have demonstrated that these approaches can deal with the problem of truncation at 9 and 14 weeks. In order to obtain reliable estimates based on the truncated INTERGROWTH-21st CRL data, the three approaches were applied as demonstrated in the next section (Figures 6.10–Figure 6.13). For demonstration purposes, I have used 35% of the overall target sample in the FGLS data.

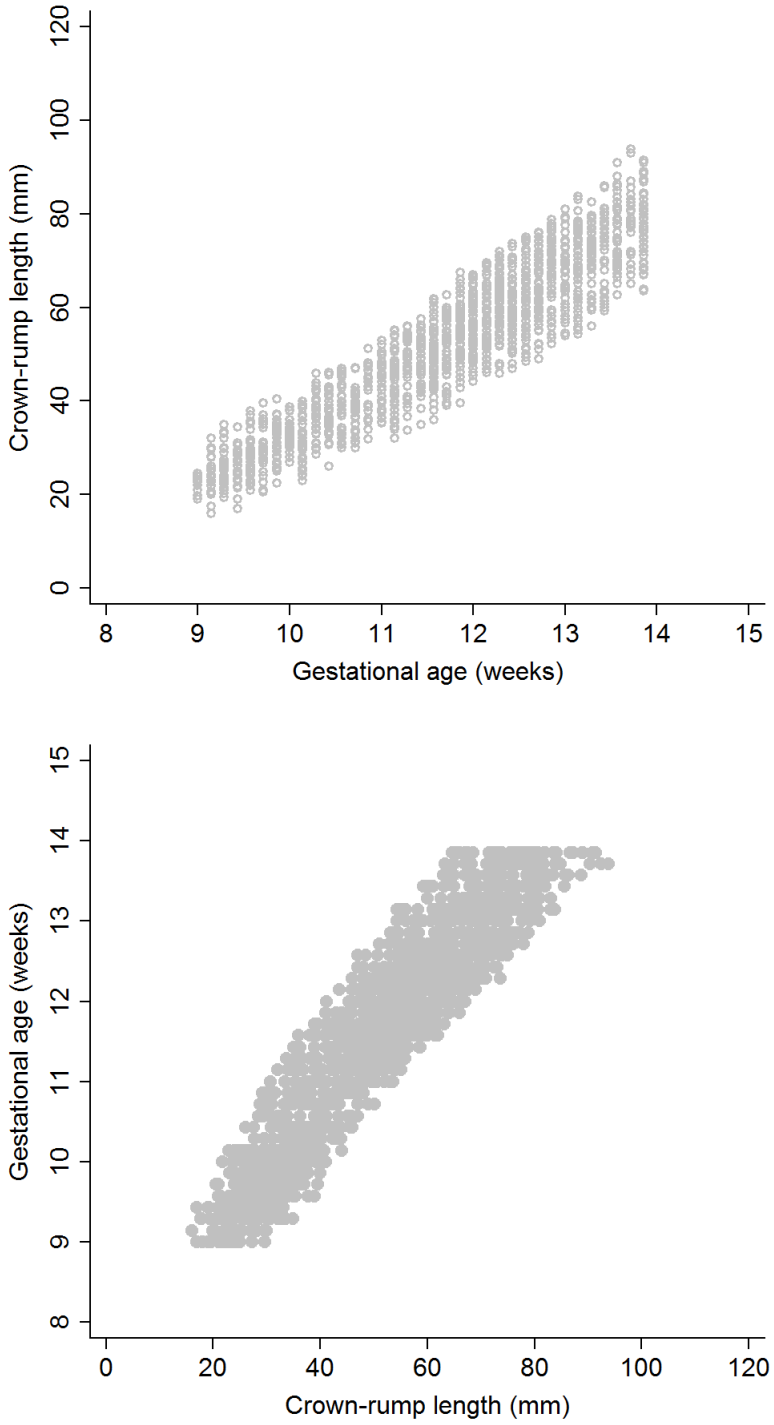


Figure 6.9: Crown-rump length (CRL) versus gestational age (GA) for creating a size chart (top figure) and GA versus CRL data for creating a dating chart (bottom figure) using a sample of the INTERGROWTH-21st project data ($\backsim$35% of the overall target sample) for CRL from 9^{+0} to 13^{+6} weeks GA.

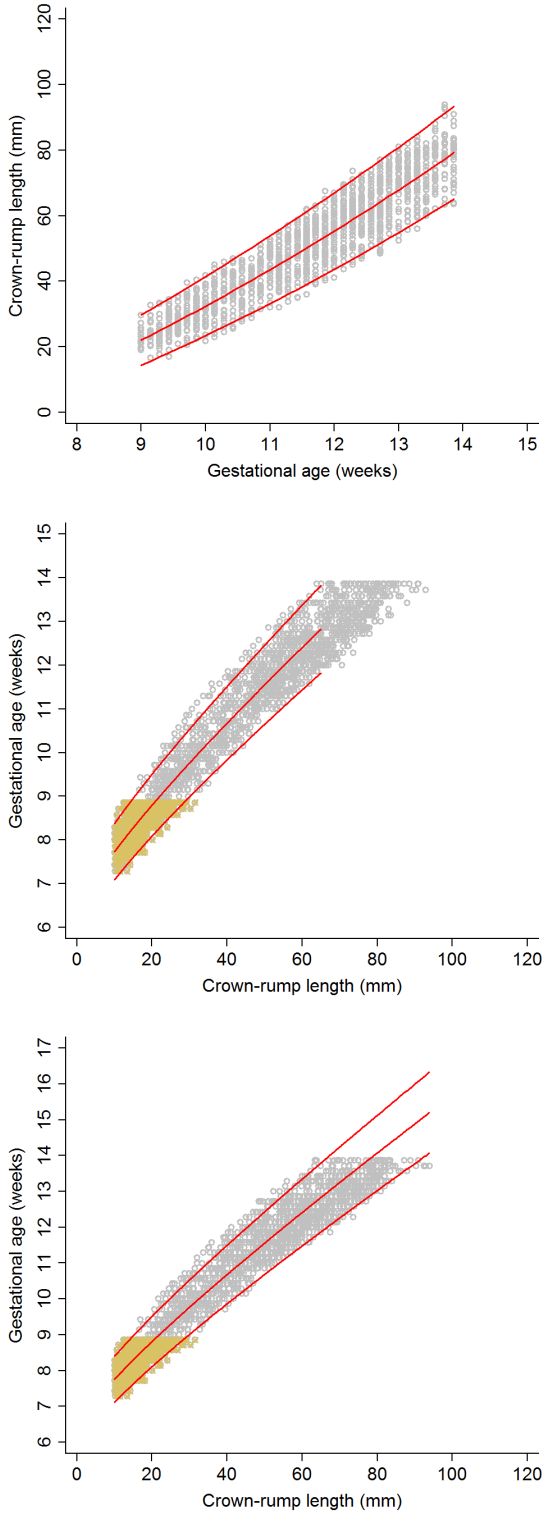


Figure 6.10: Crown-rump length measurements in relation to gestational age (grey small hollow circles) with 3^{rd}, 50^{th}, and 97^{th} fitted centiles (top figure). Yellow small crosses in the middle and bottom figures represent the INTERGROWTH-21^{st} project data for CRL from 9^{+0} to 13^{+6} weeks GA of the fitted equation of the mean and SD from the top figure. The middle figure shows the model fit relating GA and CRL with CRL restricted to $\leq$ 65 mm and the bottom figure shows the extrapolated model fit from the middle figure to the rest of the data (Approach 1).

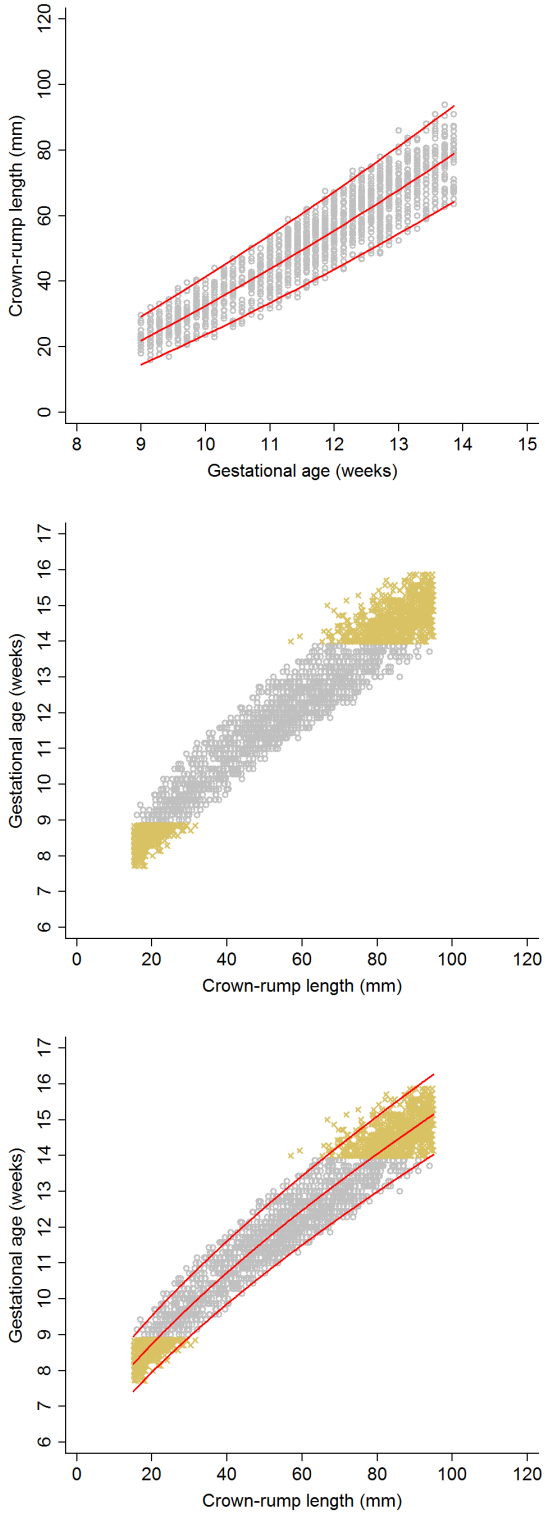


Figure 6.11: Crown-rump length measurements in relation to gestational age (grey small hollow circles) with 3^{rd}, 50^{th}, and 97^{th} fitted centiles (top figure). Yellow small crosses in the middle and bottom figures represent the INTERGROWTH-21^{st} project data for CRL from 9^{+0} to 13^{+6} weeks GA of the fitted equation of the mean and SD from the top figure. The bottom figure shows the model fit relating GA and CRL (Approach 2).

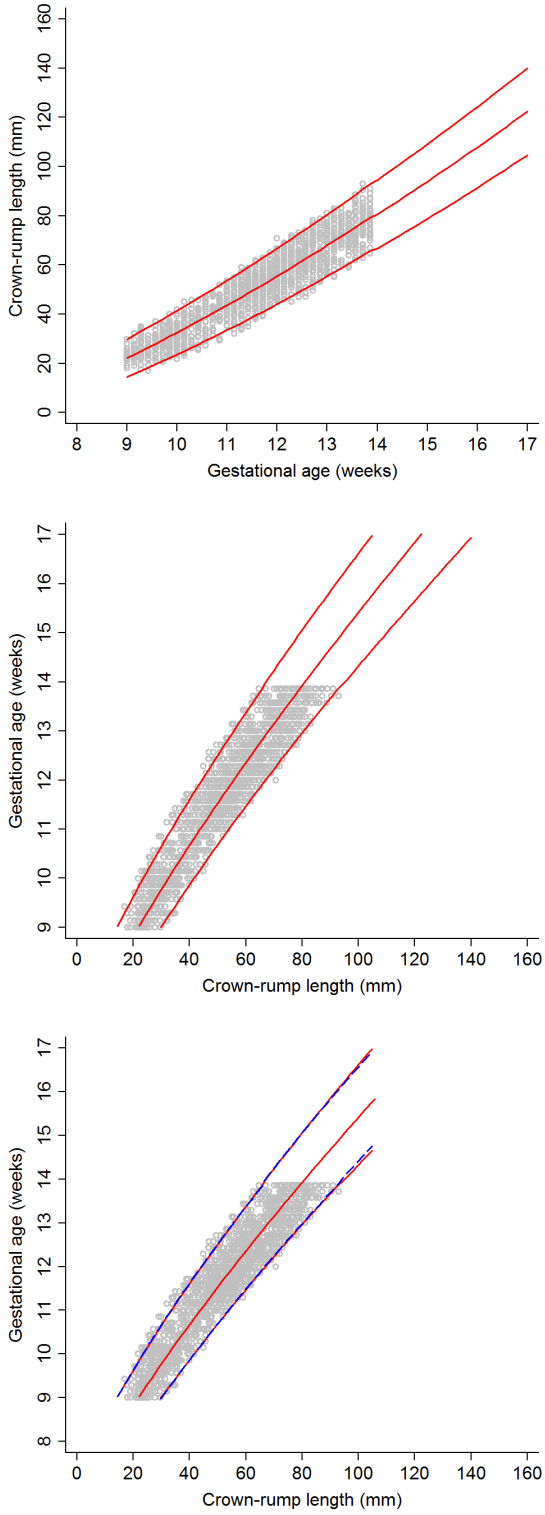


Figure 6.12: Crown-rump length measurements in relation to gestational age (grey small hollow circles) with 3^{rd}, 50^{th} and 97^{th} fitted centiles (top figure). The middle and bottom figures shows the relation between gestational age and crown-rump length of the INTERGROWTH-21^{st} project data after interchanging the axes and refitting the model (Approach 3).

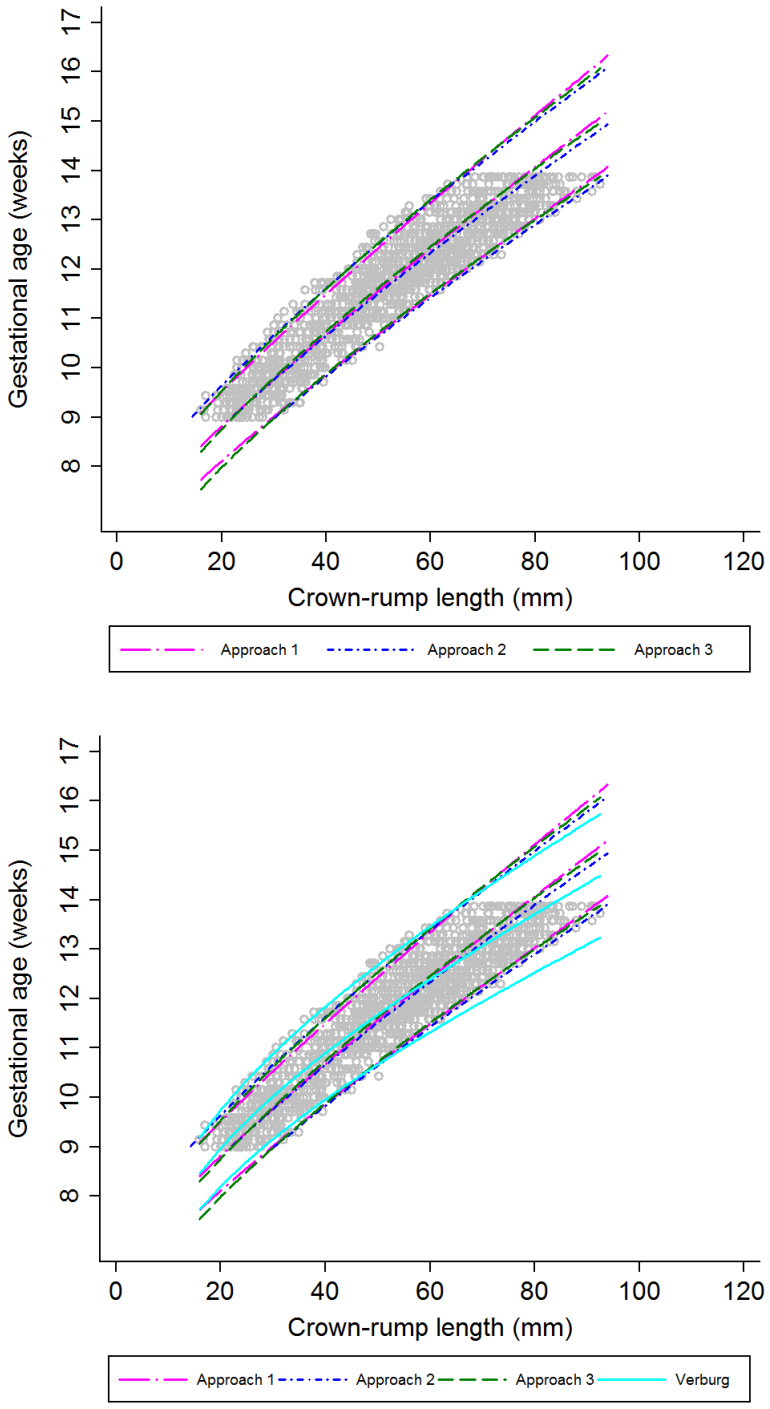


Figure 6.13: Crown-rump length measurements in relation to gestational age for the INTERGROWTH-21st project data for crown-rump length from 9^{+0} to 13^{+6} weeks gestational age comparing all the three approaches with one another (top figure) and with Verburg *et al.* (bottom figure).

6.9 Summary of results

The agreement between the median GA estimated by Approach 1 and by the original fit reported by Verburg *et al.* was within 0.4 days when CRL was between 20 mm and 100 mm. The largest difference between the estimates was at the lower range of CRL, with 4.8 days for CRL = 10 mm and 1.5 days for CRL = 15 mm (Figure 6.5, Table 6.2, Figure 6.8). This pattern arose because the model was first fit for CRL between 20 mm and 65 mm, then extrapolated to the rest of the data. Model fits beginning with lower CRL values like 10 mm and 15 mm did not perform as well when extended to the rest of the data. There were 135/4,600 (2.9%) observations below the 3^{rd} centile and 120/4,600 (2.6%) above the 97^{th} centile for CRL between 20 mm and 100 mm (Figure 6.5).

The predicted values of median GA from Approach 2 and from Verburg *et al.* were within 1 day of each other for CRL between 15 mm and 85 mm. The largest differences were seen at the two extremes of CRL, with 1.5 days for CRL = 10 mm and 1.8 days for CRL = 100 mm (Figure 6.6, Table 6.3, Figure 6.8). There were 207/7,640 (2.7%) observations below the 3^{rd} centile and 232/7,640 (3.0%) above the 97^{th} centile for CRL between 20 mm and 100 mm (Figure 6.6).

The predicted values from Approach 3 and from Verburg *et al.* were within 1 day of each other for CRL between 15 mm and 100 mm. The largest difference was of 1.5 days, which was observed at CRL = 10 mm. Approach 3 underestimated the predicted median GA across the whole range by 0.6 days (Figure 6.7, Table 6.4, Figure 6.8). There were 128/6,448 (2.0%) observations below the 3^{rd} centile and 221/6,448 (3.4%) above the 97^{th} centile for CRL between 20 mm and 100 mm (Figure 6.7). The estimates obtained from calculating SD in Approach 3 were remarkably similar to those obtained from the three sets of X, Y coordinates of GA and the predicted 3^{rd}, 50^{th}, and 97^{th} centiles for CRL (Figure 6.7, middle and bottom figures).

6.10 Discussion

The main aim of this chapter was to explore the best methodology for modelling data when the outcome variable (GA) was truncated at both ends, i.e. at 9 and 14 weeks. The statistical methodology work carried out in this chapter formed the background for the INTERGROWTH-21st clinical paper detailing the construction of an international equation for pregnancy dating. I have not discussed the results of the INTERGROWTH-21st CRL data here as the final results of the full sample and the new international dating equation have already been published [229].

To overcome the truncation problem, I evaluated three approaches by generating data from an existing equation Verburg *et al.* [36]. Each approach provided a good fit to the data (Figure 6.7) when compared with the original equation reported by Verburg *et al.*. It is difficult to justify selecting one of these approaches over the others through formal statistical testing alone. Approaches 1 and 2 both gave excellent results. However, Approach 2 gave better results than Approach 1 at the extreme ends of the CRL distribution: the Approach 2 estimates agreed within 1 day for CRL between 15 mm and 85 mm with the largest difference of 1.8 days at the very extreme ends of the distribution, whereas Approach 1 had a largest difference of 4.7 days at the lower end of the CRL distribution. Approach 3 consistently underestimated GA by about half a day over the entire range of CRL. Approach 2 can therefore be considered the best approach.

The INTERGROWTH-21st systematic review of CRL dating equations and charts showed large variations in reporting quality between studies. Very few of the studies reported complete information on inclusion/exclusion criteria, maternal demographics, ultrasound quality control, last menstruation reliability, or sample selection [27]. The resulting potential for bias, methodological heterogeneity, and limitations can affect clinical decision-making depending on the equation used;

hence the need for an international dating equation and chart.

The INTERGROWTH-21st population was carefully selected, actively followed up during pregnancy, and had a known outcome at birth. It is therefore an ideal population for developing an international standard equation and chart. The INTERGROWTH-21st project is the biggest study so far to prospectively collect data on CRL across eight geographically diverse sites. The data are of very high quality, with ultrasound measurements made by highly trained sonographers following a standardised protocol and using the latest ultrasonography equipment.

The lack of an international standard for relating CRL to GA was the motivation for producing the first international standards for early fetal size and ultrasound dating of pregnancy based on CRL measurement. The methodological work to address the truncation problem presented here was background work for a clinical paper targeting obstetricians. The methodologies discussed were later applied to the full INTERGROWTH-21st CRL dataset to develop international standards for early fetal size and pregnancy dating [24]. I produced an international prescriptive standard for early fetal linear size and ultrasound dating of pregnancy in the first trimester that can be used throughout the world. The relationship between CRL and GA was given by the following two equations (in which GA is in days and CRL in mm):

$$Mean\,CRL = -50.6562 + (0.815118 \times GA) + (0.00535302 \times GA^2), \quad (6.13)$$

$$SD\ of\ CRL = -2.21626 + (0.0984894 \times GA). \quad (6.14)$$

GA estimation is carried out according to two equations:

$$GA = 40.9041 + (3.21585 \times CRL^0.5) + (0.348956 \times CRL), \quad (6.15)$$

$$SD\ of\ GA = 2.39102 + (0.0193474 \times CRL). \quad (6.16)$$

GA estimation is an important component of clinical care and epidemiological studies. I believe that, as in other fields of medicine, all available information should be used for assessment: both LMP and ultrasound should be taken into account and agreement between the two required to be certain of their validity. Rather than automatically assuming that a discrepancy between LMP and ultrasound means incorrect dates and simply re-dating, clinicians should be aware that discrepancies can also indicate disturbances in early fetal growth.

There is wide agreement that CRL is the best measure for assessing GA, at least up to 14 weeks GA, as LMP is affected by random error, a systematic tendency to overstate the duration of gestation, biological variability, and method errors, including recall bias, digit preference, and additional bleeding after conception [32, 214, 215, 230, 231, 232, 233]. Ultrasound-based methods measure fetal size and use reliable LMP-based formulas (of which many are in use) to estimate GA. However, these formulas assume no biological variability as all fetuses of a given size are estimated to have the same GA. Biological variability exists and is compounded by variability due to measurement error from both equipment and sonographer. Thus, accurate measurements of CRL require rigorous standardisation before initiation of any study and continuous quality control measures should be implemented similar to those routinely used in laboratory practices.

The unusual problem of truncation encountered in the INTERGROWTH-21st CRL data has been present in other studies, but has never been adequately addressed. This feature of the data had the potential to introduce considerable bias, mostly at the extremes of CRL, unless analysed carefully. Altman *et al.* [92] addressed a similar problem in the estimation of GA using FHC by restricting the range of measurements included in the regression analyses. Their FHC data spanned 12–42 weeks GA. In contrast, the INTERGROWTH-21st CRL data spanned only 5 weeks. I therefore could not simply use only the CRL data unaffected by truncation, as

that would have led to a large loss of data and limited clinical usefulness.

6.11 Conclusion

Although the three approaches discussed here do not follow standard statistical analysis paradigms for modelling, I have shown empirically that the results of these rather *ad hoc* statistical methods correspond very closely to real-world data from Verburg *et al.* [36] who used a dataset similar to the INTERGROWTH-21^{st} CRL dataset. These approaches are more suitable for reducing the effect of sampling variation and ensuring reasonable extrapolation in large datasets. I am thus confident that these approaches can be used to get reliable estimates based on the INTERGROWTH-21^{st} CRL data. Although the approaches were only examined in the CRL context, they may offer a solution to other truncation problems involving similar data. Their applicability to other settings would need to be evaluated. Choosing the best approach is hard to justify through formal statistical testing and is likely to depend on the data being analysed.

6.12 Future work

There is an existing body of literature on the truncated height distributions that are common in the military following the purposeful recruitment of adults of a certain height threshold. Some researchers have attempted to model these distributions and some of the methods applied have followed similar reasoning to the approaches presented here. For example, Jacobs *et al.* [234] analysed historical height data for the Dutch province of Drenthe during 1826–1860. Although the height data was collected for conscription purposes, height measurements of those not recruited to the military were also recorded. This enabled the authors to calculate the true sample mean in the untruncated distribution of height data (the true sample mean, μ). They then discarded all observations below the truncation point and

re-analysed the data. They used various estimation procedures to estimate the mean on the truncated sample and compared the results compared to the true sample. They explored six methods based on overviews by Komlos [235] and A'Hearn [236]: the quantile bend estimator, truncated least squares, the Komlos and Kim estimator, truncated maximum likelihood, restricted truncated maximum likelihood, and converted truncated least squares.

As described in Chapter 4, Rigby and Stasinopoulos [237] developed the GAMLSS framework for fitting regression type models in which the distribution of the response variable does not have to belong to the exponential family and includes highly skewed and kurtotic continuous and discrete distribution. The GAMLSS framework is very flexible and offers several ways to extend the GAMLSS family of 45 distributions. Among the capabilities are (a) the creation of a new GAMLSS family distribution, (b) truncating new distributions, (c) using a censored version of an existing distribution, and (d) mixing different GAMLSS family distributions to create a new finite mixture distribution. To create a new distribution, the probability density function of the distribution needs to be known and easy to evaluate. The add-on package *gamlss.tr* also allows existing distributions to be truncated to the left, right, or in both tails of the range of the response y variable.

As part of future work, I would like to explore these methodologies and their application to these data. The performance of these methods will be compared with the three approaches described in this chapter.

7

Fetal growth velocity standards

7.1 Background

Chapters 4 and 5 discussed in detail the design, methodological considerations, and statistical methodology for cross-sectional and longitudinal studies for the construction of human growth charts. Charts constructed from cross-sectional data do not provide guidance for a longitudinal context. Clinicians sometimes need to know, given a fetus's current FHC, abdominal circumference or femur length, how well it has grown since the last head circumference or femur length measurement. Answering this question requires a velocity increment or velocity gain reference constructed from longitudinal data collected prospectively [91, 238].

The INTERGROWTH-21st Project collected longitudinal data to construct such references [11]. Its longitudinal component measured serial fetal growth scans every 5 ± 1 weeks from recruitment at 9^{+0}–13^{+6} weeks of gestation until, but not beyond, 42^{+0} weeks of gestation [24]. The project comprised eight geographically diverse health institutions [12]. For the data collected from the difference centres

to be of the highest quality and comparable within and between the study sites, the sonographers and anthropometrists were given standardised training and all ultrasound measurements were performed in a standardised manner following strict protocols. In Chapter 2, I discussed the procedures and methods that were employed by the INTERGROWTH-21st Project to ensure the data collected was of good quality [23]. For clarity, the table below summarises the terminology used in this chapter with a brief description of what is implied.

Terminology	Definitions
Velocity	Gain in attained fetal size adjusted for the time span between any two measurements.
Reference	Anthropometry of a given population at a particular time and place, such as a hospital, region, or country. It includes an unselected group of women with minimal exclusion criteria regarding risk factors for optimal health.
Standard	Anthropometry of a population considered to be of optimal health, with good education, socioeconomic status, adequate nutritional status, and at low risk of abnormal growth. It shows how humans should grow independent of time and place.
Velocity increment	Changes in anthropometry obtained by calculating the difference between consecutive measurements (or z-scores) in fetal biometry (e.g., FHC) divided by the time elapsed. The velocity increment is usually plotted at the mean gestational age of the two measurements.
Velocity gain	Changes in the z–scores of fetal biometry (e.g., FHC) between any two time-points. The change in z-score is assessed to evaluate the fetus's centile or z-score position, given its known previous centile or z-score position. Calculating a velocity gain depends on (a) the fetal starting point at time 1, (b) the time interval between time 1 and time 2, and (c) the correlation of the fetal measurements between the two time-points. This approach takes into account the effect of gestational age and regression to the mean.

Various approaches have been described to evaluate the rate of change (change in size) in fetal, neonatal or child growth. Milani *et al.* [239] published a review of parametric models describing the velocity curves of an individual. Gasser *et al.* [240] developed nonparametric velocity and acceleration curves for height and weight focussing on methods that assessed changes in fetal size measurements using centiles or z-scores. Argyle *et al.* [241] reviewed methodology for developing growth velocity in infancy and childhood using changes in centiles or z-scores.

Previous work on velocity has focused on postnatal centile-crossing and tracking the growth that occurs after birth [242, 243]. Studies have particularly focused on the first year of life, when infants, released from the effects of the prenatal environment, shift up (catch-up) and down (catch-down) across growth reference centiles as they fine-tune their growth rates in the postnatal environment and become channelled into their genetic potential for growth [244, 245, 246]. The prenatal triggers for these patterns have received less attention. Some studies have suggested that the relationship between upwards centile-crossing and small birth size [247], and the association between preterm labour and small birth size [248] show that catch-up growth in early infancy follows fetal growth faltering. The prenatal triggers of postnatal growth are thus a logical future area of investigation.

7.2 Introduction

Using ultrasound to monitor fetal size – measured by femur length, FHC, and abdominal circumference – is an accepted part of routine obstetric care. This practice is based on the premise that some growth disorders, such as fetal growth restriction, are manifestations of malnutrition, metabolic disorders, or infection [19] and can be treated. Fetal biometry measurements are monitored to assess whether a fetus's attained size or growth is 'normal' when compared with a defined reference population of the same age [249, 250]. These reference charts describe the average

pattern of growth and the distribution of the measurement of interest (e.g., FHC), and also show trends with GA in a prespecified population. A fetal chart usually depicts a series of smooth fitted curves of selected centiles of a distribution, usually the normal distribution. Fetal measurements falling in the extreme ends of the distribution (for example, below the 3^{rd}, 5^{th}, or 10^{th} centiles or above the 90^{th}, 95^{th}, or 97^{th} centiles) are deemed to indicate fetuses at increased risk of a growth disorder or with existing pathological conditions, such as intra-uterine growth restriction. Villar *et al.* [3] classified intrauterine growth restriction into three types depending on the timing of the fetal insult: (a) in the first trimester of pregnancy, due to a reduction in sustenance, (b) around the 30th week of pregnancy when negative factors can develop, and (c) in the last month of pregnancy, due to a reduction in food supplies and a depletion of the fetal stored fat. Longitudinal data are recommended for making such evaluations [59, 60, 91]. The INTERGROWTH-21^{st} Project recently published charts of fetal size standards based on repeated measurements [24].

The distinction between fetal size and growth is frequently ignored or misunderstood [91], which has resulted in the inappropriate use of fetal size charts for monitoring growth over time and the inappropriate use of cross-sectional data for classifying growth retardation [73, 251, 252]. Some studies have reported that velocity charts are more sensitive at detecting small-for-GA or large-for-GA fetuses than attained size charts [80].

When considering true growth, the question of interest is: 'Given a measurement, of some biometric variable, X, at a specified visit, what is the expected value of X after a certain time?' Answering this question requires a series of fetal measurements taken on multiple occasions to assess changes in size based on longitudinal data [79, 80, 253].

When the INTERGROWTH-21^{st} Project was planned, fetal velocity standards did not yet exist. These standards could be used to make judgements about fetal progress

based on previous ultrasound or anthropometric measurements [80, 254, 255]. They may not have been developed yet because it is not easy to illustrate or plot on a chart, there may have been uncertainty about how standards would be used clinically, there are inherent methodological complexities that must be considered when constructing such standards, and there was lack of high-quality longitudinal data at the time.

A key difference between velocity and distance standards is that velocity standards trace an individual's growth pattern between any two time points. Those most at risk and requiring medical intervention can be identified using the changes in growth observed between two points. Serial measurements of fetal biometry are in principle superior to single measurements, especially for classifying fetal growth restrictions [73].

Recent studies on the assessment of fetal growth by serial ultrasound examinations performed every 4 weeks from mid-gestation to birth have identified potential consequences for postnatal health. Body composition and hormonal status were associated with a decline in fetal growth velocity, irrespective of the final birthweight [256, 257].

Cameron *et al.* [258] investigated how much recovery can be expected after stunting. They advocated for catch-up growth to be estimated using z-scores rather than actual infant measurements and suggested that catch-up was only possible when the change in z-score exceeded that predicted by regression to the mean. A WHO expert committee recommended the use of z-scores was in a technical report on the use and interpretation of anthropometry [77] that has been widely adopted. Z-scores are thus recommended for assessing growth velocity as they can be used to compare across GAs and different biometric measures [259, 260].

7.2.1 Overall aim

This chapter presents work aiming to develop fetal growth velocity standards from ultrasound measurements of FHC, abdominal circumference, and femur length

based on data from the FGLS component of the INTERGROWTH-21st Project [24, 11]. These standards were developed using a different methodology from, but should complement, the INTERGROWTH-21st fetal size standards that were the primary output for the project. I used a two-stage approach. I first carried out a systematic review of the literature on velocity charts and methodology applied to fetal and child growth. I then used the identified methodology to develop fetal velocity standards using the INTERGROWTH-21st dataset.

7.2.2 Justification

Velocity standards will form an important complement to the published INTERGROWTH-21st International fetal growth standards (size charts) [24]. I used the same dataset that was used to construct the international fetal standards to develop the new velocity standards. Altman and Hytten [91] highlighted the clear and urgent need for true measures of fetal growth from which deviations indicating genuine growth retardation can be derived. Sovio *et al.* [261] sought to determine the diagnostic effectiveness of universal ultrasonic fetal biometry in the third trimester as a screening test for small-for-GA infants. They found that serial assessment of fetal biometry in all pregnancies improved the detection of small-for-GA neonates and identifies a subset at risk for morbidity and mortality. The detection rate of small-for-GA neonates was tripled by screening women in the third trimester. The authors recommended a combined approach using both fetal biometry and fetal growth velocity (measured with abdominal circumference) to identify a subset of small-for-GA fetuses that were at increased risk of neonatal morbidity, mortality, or adverse outcomes.

Rates of growth differ across time and between individuals. A better characterisation of insults during pregnancy requires identifying and characterising periods of rapid growth [262, 3]. Although human growth is particularly rapid during fetal life,

standard obstetric care does not include monitoring fetal growth in women with low-risk pregnancies. A meta-analysis of seven trials that evaluated the effects of routine late pregnancy ultrasound, defined as occuring at more than 24 weeks gestation, on obstetric practice and pregnancy outcome in women with either unselected or low-risk pregnancies concluded that there was no evidence of a beneficial effect [57]. The results of this systematic review led to the recommendation that late pregnancy ultrasound should not be offered routinely in the third trimester [263, 264]. Sovio *et al.* [261] conducted a prospective cohort study of unselected nulliparous women with a singleton viable gestation and reliable dating in early pregnancy. They confirmed the already-overwhelming evidence that fetal growth disorders are risk factors for adverse neonatal outcomes and can predispose infants to adult chronic diseases [265, 266, 267, 268]. The Royal College of Obstetricians and Gynaecologists evidence-based recommendations for managing suspected fetal growth restriction includes fetal monitoring, timing the induction of labour, and how to undertake delivery. In light of these findings, a programme of screening that includes universal ultrasonography and intervention has the potential to reduce the number of adverse perinatal outcomes caused by fetal growth restriction [269]. In planning public health activities such as nutrition interventions for developing countries, the type of intrauterine growth retardation present in the target population should be considered to determine which type of intervention would be most appropriate and establish the correct timing.

To address some of the issues raised regarding fetal monitoring and the benefits that can be deduced from prospective ultrasound measurements of fetuses in improving detection of fetal growth restriction, a longitudinal study of a large number of women that takes scheduled repeat measurements of fetal dimensions such as FHC, abdominal circumference, and femur length using ultrasound has been recommended [91].

7.2.3 Velocity or increment reference values

Velocity reference charts (also known as increment reference charts) are the simplest form of describing growth velocity. Velocity charts represent the distribution of velocity as a function of GA. An alternative to the chart-based approach is the use of incremental tables [270, 271, 271, 272, 273]. These tables present age range, for example 0–1 months, versus the mean and SD of increments, for example birthweight (g or kg/day) or length (cm/day), along with selected centiles (e.g. the 5^{th}, 10^{th}, 50^{th}, 90^{th}, and 95^{th} centiles) for that age range.

According to Tanner [30], velocity represents what is happening now, whereas attained size represents a summary of all that has happened in the past. The main restriction in using velocity increment charts and tables is that fixed measurement intervals are used. Healy [274, 106, 100] highlighted other shortcomings of this approach in three papers. He noted that differences between any two measurements may be unduly large because either the final or initial measurement is far from its expected value. The usual practice is to take the difference between two values, express it as a velocity, and compare it with a reference or standard chart. Growth velocity is calculated as the difference between two measurements, Y_1 and Y_2, taken at times t_1, and t_2 [274, 275, 276]. Velocity references are influenced by the time interval between measurements, as the rate of growth is not constant over time. For example, fetal growth is rapid in early pregnancy and slows down towards term [277]. Relative measurement error is greater over shorter periods and during saltatory growth [278, 277, 279]. Velocity charts are thus usually based on fixed time periods. Velocity charts of height and weight during child growth have been recommended to be based on minimum intervals of 1 year to avoid seasonal effects. In contrast, intervals of between 1 week and 3 months have been recommended for velocity charts for fetal and infant growth [80, 274, 280, 77, 281, 282]. The choice of an appropriate interval is a trade-off between the noise (measurement

error) and signal (ability to detect true growth velocity) [283]. A limitation of these charts is that they can only be used for measurements taken very near the specified intervals, e.g. 1-year intervals for the Tanner charts and weekly for the Owen and Campbell fetal biometry charts.

The WHO, as part of the MGRS, published growth velocity increment charts for birthweight, BL, and BHC for monitoring infants in the first 2 years of life [273]. The increment references were derived from a longitudinal study of a cohort of children who were examined in 21 visits from birth to 2 years of age [159]. The increments on which the velocity standards were based were calculated using a longitudinal sample of 882 children (428 boys and 454 girls) whose mothers complied fully with the MGRS infant-feeding and no-smoking criteria and completed 24 months of follow-up. The children were measured at birth; at weeks 1, 2, 4, and 6; monthly at 2-12 months; and bi-monthly in the second year. Weight increments were calculated and presented for 1-month periods from birth to 12 months, and for 2- to 6-month periods from birth to 24 months. Weight increments were also presented from birth to 60 days in 1-week and 2-week intervals. Velocity references for length were presented for 2- to 6-month periods from birth to 24 months. References for BHC were presented for 2- and 3-month increments from birth to 12 months, and 4- to 6-month increments from birth to 24 months [273].

7.2.4 Velocity gain z-scores

Instead of incremental velocities, Healy suggested a reference approach that results in centiles at time t_2 given measurements at time t_1. He argued that such centiles are more valuable as they potentially have greater sensitivity for detecting fetal growth restriction than increment velocities obtained by merely taking the difference between any pair of measurements without accounting for the size of the previous measurements. The approach also avoids ascribing abnormally low velocities to

children who happen to be above their expected measurement at the start of the interval or vice versa [106].

Another important aspect to consider when constructing velocity is the well-known phenomenon of regression to the mean, first described by Galton in 1886. He related the heights of children to the average height of their parents, which he called the mid-parent height [284]. He noted, for example, that there was a tendency for tall fathers to have shorter sons. This is a statistical phenomenon, whereby if individuals or groups of individuals are measured on two occasions, the second measurement is likely, on average, to be nearer the mean than the first measurement [242, 243, 285].

Regression to the mean occurs because there is never perfect correlation mainly because of measurement error [285]. If two measures are weakly correlated, then regression to the mean has a greater effect [286]. Suppose the z-score, Z_1, of a fetus is 2 at time t_1, the fetus is 2 SD above the mean ($\sim 97^{th}$ centile) and if the correlation between Z_1 and the z-score Z_2 at time t_2 is 0.75. At time t_2, I would expect the fetus to be $2 \times 0.75 = 1.5$ SD above the mean, on average ($\sim 93^{rd}$ centile). This reduction by 4 centiles is commonly referred to as 'catch-down' growth. The reverse, 'catch-up growth', also occurs [287]. On average, most, but not all, light fetuses will 'catch-up' and most, but not all, heavy fetuses will 'catch-down' [243]. Centile crossing, both upwards and downwards, has been used to describe growth rate changes with respect to reference growth charts after growth retardation following illness and acceleration of growth during recovery [288]. Centile-crossing has been used more often in studies that assess postnatal growth patterns [244, 245, 247] and less frequently in studies of prenatal growth [257, 289]. The term 'tracking' is commonly used to refer to the maintenance of rank order within a group of peers over time [248]. The absence of centile crossing is called perfect tracking [290] and can sometimes be used to refer to future predictions. A correlation coefficient between two measurements can give an indication of centile tracking [290].

Calculating a velocity gain z-score requires (a) a set of reference fetal size standards to convert respective fetal measurements to z-scores and (b) the correlation between any pair of z-scores for each fetal measurement [277]. The value of the correlation will depend on the time interval between the pair of z-scores and the respective GAs, which indicates whether the measurements were taken in early, mid, or late pregnancy. The INTERGROWTH-21st dataset provides all of the required information, except an estimate of the correlation. An estimate of correlation is easy to calculate for situations in which data are collected at fixed time intervals. However, this is seldom the case in normal practice as fetuses are seen and measured at arbitrary time points over time. For pragmatic reasons it is impossible to see and measure everyone on the same day i.e., at fixed time intervals.

Fetal size measurements tend to have a close to normal distribution at any GA [92]. Data that are normally distributed can be summarised using the mean and SD, from which any desired centiles or z-scores can be calculated. By definition, z-scores are normally distributed, with mean = 0 and SD = 1. If the distribution of reference values follows a normal distribution, the centiles and z-scores are related through a mathematical transformation. The commonly used z-scores of -3, -2, and -1 correspond to the 0.13th, 2.28th, and 15.8th centiles, respectively. Similarly, the 1st, 3rd, and 10th centiles correspond to the -2.33, -1.88, and -1.38 z-scores, respectively.

Once again, the question is: 'Given a fetus's previous FHC, abdominal circumference, and femur length measurements, how likely are the current measurements?'

7.2.5 Presentation of size versus growth velocity standards

It is common practice to present fetal size standards in relation to GA as centile curves, tabulated centiles, or both. Centile curves provide an easy way to track individual growth patterns, but do not offer the same representation for growth velocity. Growth velocity is highly variable between fetuses. It is not unusual for

a fetus whose increment at one time point is on the 5^{th} centile to be at the 75^{th} centile at the next measurement [273]. Interpreting fetal velocity is different from interpreting the more commonly used fetal size charts. Fetuses do not necessarily track on a fixed velocity curve, except perhaps in the median range [274, 281]. Tanner [291] recommended that velocity increments be used in conjunction with size charts. An example of such an effort is the 3-in-1 weight monitoring chart proposed by Cole [292], amongst others [242, 283].

7.3 Data

The FGLS component of the INTERGROWTH-21^{st} Project generated a unique dataset as it was the largest prospective study to date to collect data on fetal ultrasound measurements among optimally healthy pregnant women in geographically diverse populations, with a high level of quality control measures in place. To ensure that they collected accurate, reproducible ultrasound measurements, the study centres used uniform methods, identical ultrasound equipment, and standardised methodology to take fetal measurements, and they employed locally accredited ultrasonographers who underwent standardisation training and monitoring [293]. Of the 13,108 pregnant women who were screened, 4,607 met the clinical eligibility criteria and were enrolled in the study. All of the women were closely followed up throughout pregnancy by the study team until delivery and discharge from hospital. The target sample for the proposed analyses was formed by the enrolled 4,321 women had live singleton births in the absence of severe maternal conditions or fetal congenital abnormalities detected by ultrasound or at birth. Individual ultrasound measurements of FHC, biparietal diameter, occipitofrontal diameter, abdominal circumference and femur length were taken at regular intervals (every 4 to 6 weeks) between the 14^{th} and 40^{th} week of gestation with between four and seven measures per fetus. These data are suitable for evaluating ultrasound estimated

fetal growth rates for the prediction of adverse perinatal outcomes and for studying the rate of fetal growth between two defined time points (velocity).

The median number of ultrasound scans (excluding the dating scan) was 5.0 (mean = 4.9, SD = 0.8, range from 4 to 7) and 97% of the women had ≥ 4 scans (mean = 5.0, SD = 0.6, range from 4 to 7), indicating that the participants adhered well to the protocol. Eighty-five percent of the 20,313 ultrasound scans were performed within the expected GA window of the protocol (range from 76% in India to 93% in Oman) [24]. Although measurements were taken across GA from 14 to 40 weeks, the high protocol adherence rates in the FGLS meant that the intervals between the adjacent measurements were mostly 4- (n = 3,836), 5- (n = 8,871), or 6- (n = 2,411) week intervals. As most of the women were seen in 4- , 5- , and 6- week intervals, shorter or longer follow-up intervals were unusual and were therefore not considered in the analyses. The two diameters of the head, the biparietal and occipito-frontal diameters, were not considered relevant for the velocity analyses as they were already represented by the FHC. The 4,321 fetuses contributed 20,030 fetal measurements. These data were used to construct fetal growth velocity and to assess changes in fetal size standards [24]. As the main focus of this thesis is methodology, I only present analyses and results based on the two approaches using FHC data as an example. The same methodological framework and analyses were applied to abdominal circumference and femur length data without any problems (results are not shown).

Table 7.1 summarises the number of women and total number of follow-up visits of FHC from 14 to 40 weeks of GA. Scatter plots of increments in raw FHC, abdominal circumference, and femur length data (mm/week) according to GA (weeks) for all of the sites combined are shown in Figure 7.1.

Number of visits	Number of women	Percentage	Number of observations	Percentage
1	39	0.92	4,233	21.13
2	55	1.30	4,194	20.94
3	203	4.80	4,139	20.66
4	810	19.14	3,936	19.65
5	2,724	64.35	3,126	15.61
6	402	9.50	402	2.01
Total	4,233	100.00	20,030	100.00

Table 7.1: Summary of the number of women and total number of follow-up visits measuring fetal head circumference.

Figure 7.1: Increments in fetal head circumference, abdominal circumference and femur length data (mm/week) according to gestational age (weeks) for all of the sites combined.

7.4 Data analyses

The INTERGROWTH-21st fetal growth data were used to model velocity increment centiles on a continuous scale. Centile curves were generated both graphically and in tabular format for 4-, 5-, and 6-weeks intervals starting at 14 weeks of gestation. Diagnostic plots of the model fit are shown, empirical values and fitted centiles are compared across GA, and the differences between the empirical and fitted centiles across GA are quantified for selected centiles. Quantile-quantile plots of the fitted model and scatter plots of resultant z-scores from the fitted model by GA are also presented.

7.4.1 Velocity or increment reference values

Pairs of consecutive measurements (Y_1 and Y_2) were considered for each fetus and the time elapsed (time interval) between each pair of measurements calculated. The average daily growth rate for each individual was obtained from the difference in the measurement between the two time points ($Y_2 - Y_1$) and was plotted at the mid-time-point of the two measurements, i.e., at time $(t_2 - t_1)/2$. The average daily rate of growth was thus given by:

$$\frac{Y_2 - Y_1}{(t_2 - t_1)/2} \tag{7.1}$$

Growth was modelled as a function of gestation age by expressing FHC velocity at the mid-time-point between any pair of observations using FPs [163]. The centiles and z-scores for increment velocities were computed in the same way as in attained size charts [24]. The mean and SD were fitted using FPs that were based on the normality assumption. Each desired centile or z-score of increment x was calculated using:

$$\text{z-score of } x = \frac{x - \mu(x)}{\sigma(x)}, \tag{7.2}$$

where x refers to an increment in the size of a fetus at a specified time point, $\mu(x)$ is the FP fitted mean velocity of all fetuses observed at that time point, and $\sigma(x)$ is the FP fitted average SD of all fetuses observed at that time point. For example, to calculate a z-score for an 18-week-old fetus with an increment in FHC of 42 mm between 14 and 18 weeks: increment $= 42$ mm, $\mu(\text{x}) = 50.71$ mm, and $\sigma(\text{x}) = 6.07$ mm,

$$\text{z-score of increment} = \frac{(42 - 50.71)}{6.07} = -1.43. \tag{7.3}$$

Any desired centile estimate can be calculated using the relation:

$$p^{\text{th}} \text{ centile} = \text{increment} \pm \text{K}\times\text{SD(x)}, \tag{7.4}$$

where K is the normal equivalent deviate (z-score) corresponding to a particular centile, e.g. $K = 1.88$ for the 97^{th} centile and -1.88 for the 3^{rd} centile, and the mean and SD are the predicted estimates from the FP regression analysis. For example, to convert the z-score calculated above to a corresponding velocity increment:

$$\mu(\text{x}) = 50.71 \text{ mm}, \ \sigma(\text{x}) = 6.07 \text{ mm}, \text{ and } \text{z} - \text{score} = -1.43,$$

$$\text{Increment} = 50.71 + (-1.43 \times 6.07) = 42.03 \text{ mm}. \tag{7.5}$$

7.4.2 Velocity gain z-scores

The velocity increment calculated above does not take into account an individual's size at time t_1 and the correlation between the distribution of z-scores at the two time points t_1 and t_2. In contrast, velocity gain does take these considerations into account. All of the ultrasound data from the FGLS were converted to z-scores using the international fetal growth standards [24]. The z-score values were normally

distributed as the fetal standards were constructed using the same data. Using FHC as an example, velocity can be assessed by considering the difference in z-scores between any two time points, Z_2 - Z_1. Change is assessed to predict the fetus's centile or z-score position, given its known previous centile or z-score position. The aim is to predict Z_2 from Z_1 and compare it with the observed Z_2. A negative value indicates a fall in the fetus's position relative to the population. However, the value of the difference should be interpreted considering the value of Z_1 used in its calculation. Fetuses with a small initial velocity, such as $Z_1 <$ -1 z-score, will on average show an increased velocity when their FHC measurement is taken later, and vice versa. Z_2 is expected to become smaller (nearer to zero, the median) with time. The correlation is a direct measure of regression to the mean [242, 243, 285].

Consider the distribution of the FHC z-scores Z_1 and Z_2 at time points t_1 and t_2, and the correlation r between them. If Z_2 is regressed on Z_1, then:

$$\mathrm{Z_2 = a + b \times Z_1 + \varepsilon}, \tag{7.6}$$

where a, b, and ε are the intercept, gradient, and residual error. As Z_1 and Z_2 are z-scores derived from fetal size standards constructed using the same data, Z_1 and Z_2 both have normal distributions with mean = 0 and SD = 1. The regression coefficient, b, is equal to the correlation r between Z_1 and Z_2. The expected value of Z_2 given Z_1 can thus be simplified to:

$$\mathrm{Z_2 = r.Z_1 + \varepsilon}. \tag{7.7}$$

The expression $\mathrm{Z_2 - r.Z_1}$ is a measure of the difference between Z_2 and its expected value. The SD of $\mathrm{Z_2 - r.Z_1}$ is given by $\sqrt{(1-r^2)}$ [287]. This leads to the

velocity gain z-score:

$$\text{Velocity gain z-score} = \frac{Z_2 - r.Z_1}{\sqrt{(1-r^2)}} \tag{7.8}$$

where $t_1 < t_2$, Z_1 is the initial FHC z-score at t_1, Z_2 is the later FHC z-score at t_2, and r is the correlation between Z_1 and Z_2 [243].

7.4.2.1 Implementation

Empirical correlations were calculated for all possible z-score pairs. Estimates of correlation are unstable for small sample sizes, so the greater the number of observations, the better the sample correlation approximates the true correlation. Correlations were therefore only calculated for samples of at least 10 measurement pairs. Fisher demonstrated that samples from a bivariate normal distribution with sample sizes of 10 or more had z-scores with an approximately normal distribution and a remarkably robust transformation [294]. Fisher's transformation is a variance-stabilising transformation:

$$\text{Fisher's } z\text{-statistic} = 0.5 \times \log_e \left\{ \frac{1+r}{1-r} \right\} \tag{7.9}$$

The empirical correlations were transformed with Fisher's transformation and modelled using FP regression [163]. The transformed correlations were then modelled as a function of the two GAs, the mean GA of each pair of observations, $(t_1 + t_2)/2$, and their difference, $(t_2 - t_1)$ [243]. This resulted in smoothed estimates of the correlation as a function of GA. The predicted values from the FP regression model were then back-transformed to the original scale for correlation using:

$$\text{Correlation coefficient, } r = \frac{\exp(2z) - 1}{\exp\ (2z) + 1} \tag{7.10}$$

where z is Fisher's transformed correlation.

7.4.3 Comparing the velocity increment approach with the velocity gain z-score approach

The velocity increment and velocity gain approaches were compared by (a) plotting empirical densities of the distribution of the z-scores from each method and (b) comparing the distributions of all pairwise differences between the z-scores from the two methods. The fetal data were also used to check for evidence of regression to the mean. The fetuses were categorised as either below the lower cut-off (fetal z-scores between -2.5 and -1.5) or above the upper cut-off (fetal z-scores between +1.5 and +2.5) of the attained fetal sizes at t_1 of 4-, 5-, and 6-weekly intervals. The change in z-scores, $Z_2 - Z_1$, of the 4-, 5-, and 6-week intervals were examined to assess the extent of regression to the mean in the data.

7.5 Results

7.5.1 Velocity or increments reference values approach

FHC velocity increments were modelled for 4-, 5-, and 6-week intervals, accounting for GA. Tables generated from the 4-week increment curves contain estimated centiles for GAs 14–18, 15–19, . . . , 36–40 weeks. Tables generated from the 5-week increment curves contain estimated centiles for 14-19, 15-20, . . . , 35-40 weeks. Tables generated from the 6-week increment curves contain estimated centiles for 14–20, 15–21, . . . , 34–40 weeks. Figure 7.1 shows the FHC velocity increments in mm/week.

7.5.1.1 Four-week increments

There were 3,836 4-week increments. The best FP model was an FP2 with powers for the mean = 1, 1 and SD = 1. Repeat powers are allowed in FPs. Each time a power repeats, one of the two terms is multiplied by $\log_e(x)$ [163]. A summary table with descriptive statistics for 4-week intervals showing their sample sizes is

presented in Table 7.2. Figure 7.2 shows the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles in mm/week according to GA. The associated diagnostic results are shown in Figures 7.3 and 7.4. Figure 7.5 compares the fitted smoothed centiles with their corresponding empirical centiles for selected centiles.

There was no evidence of bias when the fitted and empirical centiles were compared or when the residuals for the selected centiles were assessed, including the extreme centiles (e.g., the 3^{rd} and 97^{th} centiles) where the differences were expected to be greatest (Figures 7.5 and 7.6). The differences between the fitted and empirical centiles were within 3 mm from 14 to 39 weeks and were in the range 4 mm to 8 mm at 40 weeks. The greater range at 40 weeks was largely attributable to the small sample size (n = 9). Table 7.3 presents the predicted 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles for 4-week increment velocities between 14 and 40 weeks.

Gestational age (completed weeks)	Minimum (mm)	Maximum (mm)	Mean (mm)	SD (mm)	Total (n)
14–18	39.09	70.08	53.08	6.14	58
15–19	38.46	67.98	52.85	6.00	126
16–20	35.11	67.93	50.40	5.63	218
17–21	32.58	70.22	49.03	6.57	217
18–22	36.13	66.59	49.90	6.01	139
19–23	34.99	66.88	49.62	5.75	99
20–24	32.86	61.69	48.08	6.00	147
21–25	29.38	63.63	48.02	6.62	151
22–26	31.81	69.52	48.26	6.40	144
23–27	34.60	66.96	48.03	6.57	124
24–28	28.23	58.73	43.31	6.87	105
25–29	25.92	59.96	43.54	7.29	144
26–30	20.01	56.51	40.46	6.68	195
27–31	20.17	58.55	37.76	7.45	160
28–32	13.69	54.67	34.45	7.71	180
29–33	13.58	59.58	31.99	7.00	195
30–34	11.48	50.04	28.43	7.43	193
31–35	3.38	54.92	25.69	8.04	230
32–36	1.64	41.76	23.84	7.23	208
33–37	0.77	50.18	20.31	6.98	462
34–38	1.77	40.96	18.94	7.95	191
35–39	1.04	35.58	15.99	7.23	141
36–40	5.15	25.48	15.87	6.67	9
Total	0.77	70.22	36.33	14.10	3836

Table 7.2: Descriptive statistics for the 4-week increments in head circumference (mm).

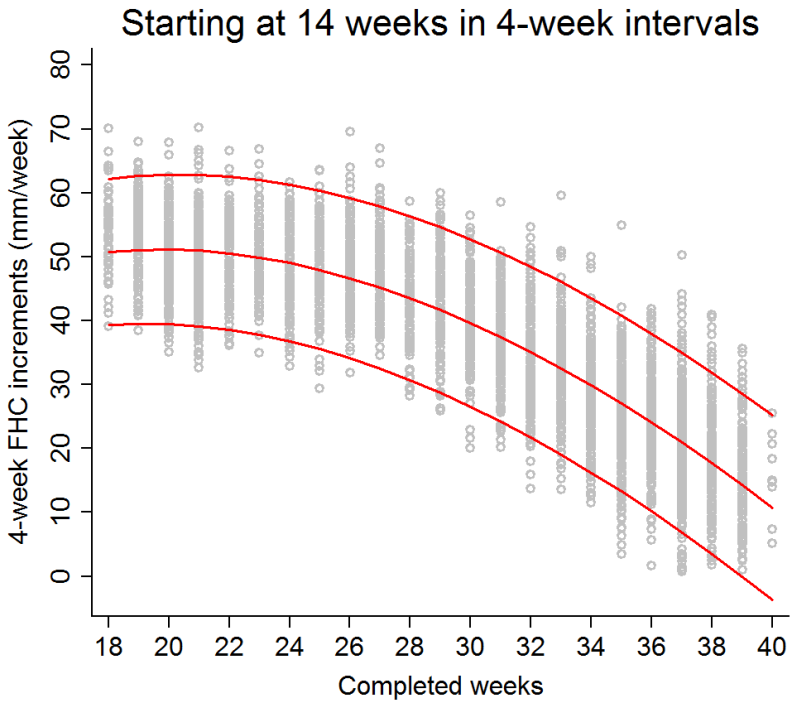


Figure 7.2: Individual 4-week increments in fetal head circumference (mm) (FHC), with the 3^{rd}, 50^{th}, and 97^{th} centiles superimposed according to gestational age (weeks).

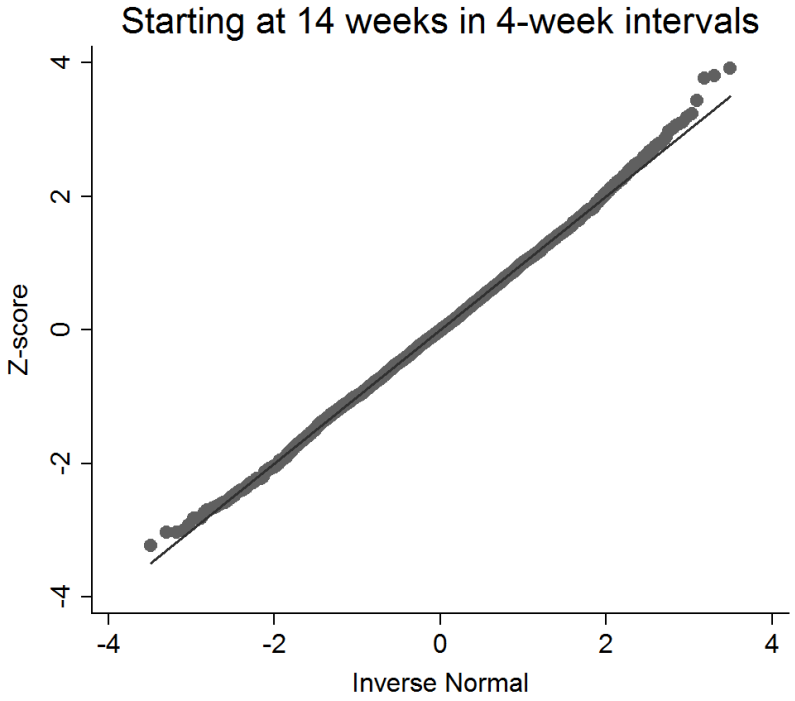


Figure 7.3: Quantile-quantile plot of the fitted model of 4-week increments in fetal head circumference (mm).

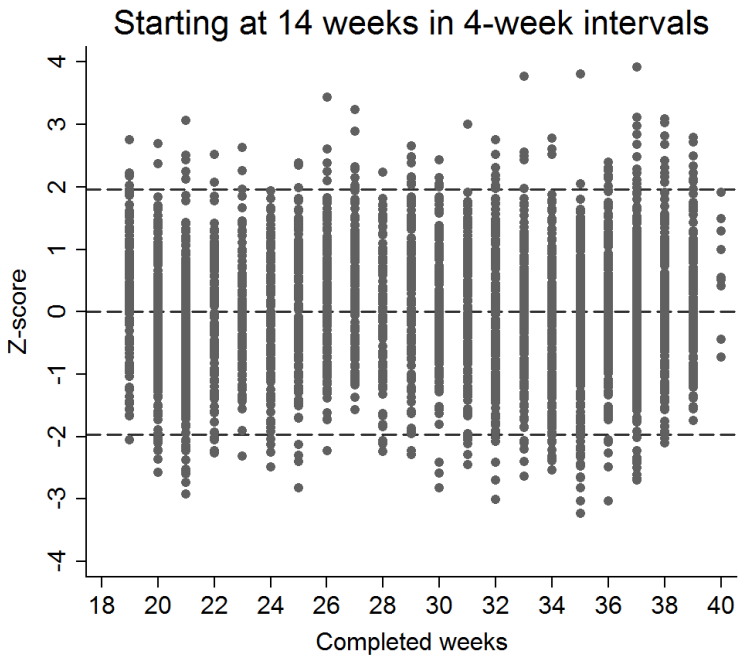


Figure 7.4: Scatter plot of individual 4-week increments in fetal head circumference (mm) from the fitted model according to gestational age in completed weeks.

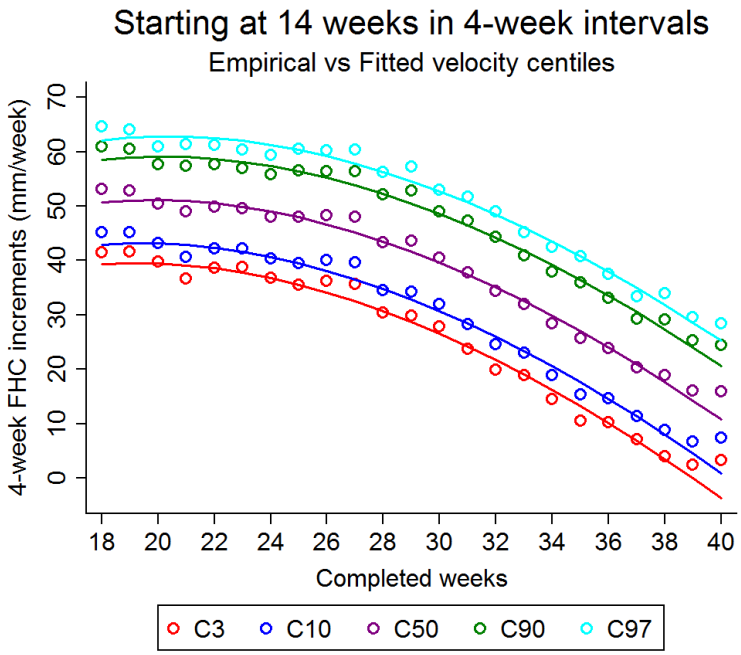


Figure 7.5: Comparison of fitted (solid lines) and empirical (open circles) 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles for the 4-week increments in fetal head circumference (mm) according to gestational age (weeks).

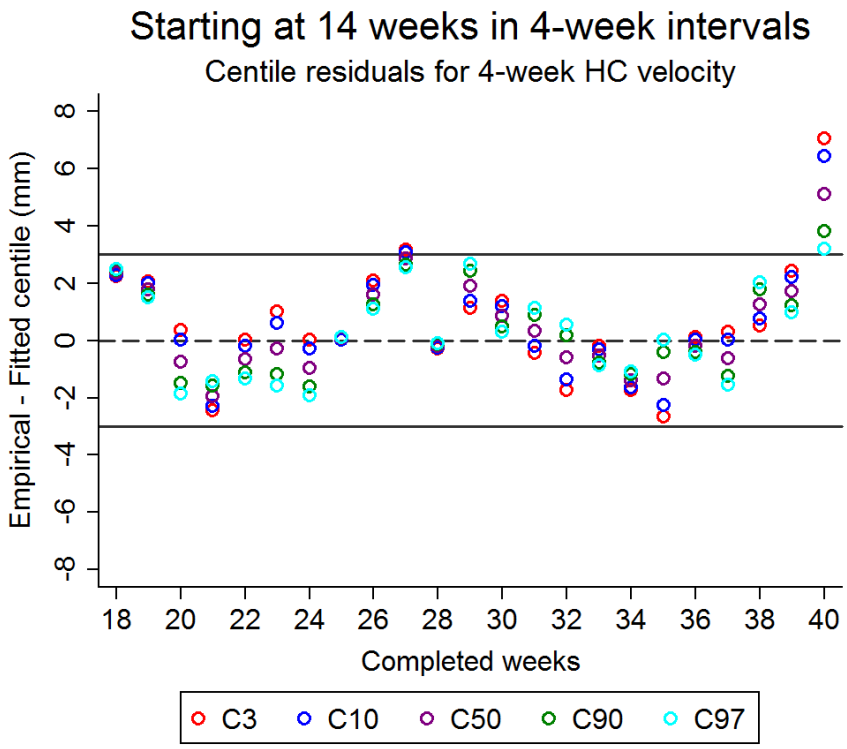


Figure 7.6: Differences in fitted and empirical centiles (residuals) for 4-week increments in fetal head circumference (mm) at 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles according to gestational age (weeks).

Gestational age (completed weeks)	Total (n)	3^{rd} centile	10^{th} centile	50^{th} centile	90^{th} centile	97^{th} centile
14–18	58	39.29	42.93	50.71	58.50	62.14
15–19	126	39.50	43.18	51.06	58.94	62.63
16–20	218	39.44	43.17	51.14	59.12	62.85
17–21	217	39.13	42.90	50.97	59.04	62.81
18–22	139	38.57	42.39	50.55	58.71	62.53
19–23	99	37.78	41.65	49.90	58.16	62.02
20–24	147	36.78	40.68	49.03	57.38	61.29
21–25	151	35.55	39.50	47.95	56.39	60.34
22–26	144	34.13	38.12	46.66	55.20	59.19
23–27	124	32.51	36.54	45.18	53.81	57.85
24–28	105	30.69	34.78	43.50	52.23	56.31
25–29	144	28.70	32.83	41.65	50.47	54.60
26–30	195	26.53	30.70	39.62	48.54	52.71
27–31	160	24.19	28.40	37.42	46.43	50.64
28–32	180	21.69	25.94	35.05	44.16	48.42
29–33	195	19.02	23.33	32.53	41.73	46.03
30–34	193	16.20	20.55	29.85	39.14	43.49
31–35	230	13.24	17.63	27.02	36.41	40.80
32–36	208	10.12	14.56	24.04	33.52	37.96
33–37	462	6.87	11.34	20.92	30.50	34.98
34–38	191	3.47	8.00	17.67	27.34	31.86
35–39	141	-0.06	4.51	14.28	24.04	28.61
36–40	9	-3.71	0.90	10.76	20.62	25.23
Total	3,836					

Table 7.3: Fitted 4-week increments in fetal head circumference (mm) at the 3^{rd}, 10^{th}, 50^{th}, and 97^{th} centiles according to gestational age (weeks).

7.5.1.2 Five-week increments

There were 8,871 5-week increments. The best FP model was an FP2, with powers for the mean = 2, 2 and SD = 1. A summary table with descriptive statistics for the intervals and their sample sizes is presented in Table 7.4. Figure 7.7 shows the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles in mm/week according to GA. The associated diagnostic results are shown in Figures 7.8 and 7.9. Figure 7.10 compares the fitted smoothed centiles with their corresponding empirical centiles for selected centiles. The difference were within 4 mm from 14 to 40 weeks (Figures 7.10 and 7.11). Table 7.5 presents the predicted 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles.

Gestational age (completed weeks)	Minimum (mm)	Maximum (mm)	Mean (mm)	SD (mm)	Total (n)
14–19	44.08	79.91	62.73	5.70	190
15–20	43.53	80.63	61.46	5.84	367
16–21	40.21	78.44	59.01	6.29	465
17–22	38.39	77.30	58.74	6.15	558
18–23	42.40	74.28	58.39	5.59	532
19–24	35.59	80.53	56.76	6.37	300
20–25	30.04	75.39	56.59	6.69	425
21–26	35.86	80.44	56.44	6.64	509
22–27	34.42	82.81	55.89	6.51	592
23–28	29.86	76.21	54.65	6.81	593
24–29	28.08	72.98	51.70	7.45	360
25–30	28.10	77.34	48.52	7.65	386
26–31	24.12	75.81	46.40	7.62	444
27–32	21.70	67.72	42.94	7.32	552
28–33	11.73	63.91	39.48	7.53	576
29–34	11.32	60.55	35.59	8.34	369
30–35	2.55	57.64	32.77	8.66	385
31–36	8.10	51.33	30.75	7.88	352
32–37	7.24	56.71	26.63	7.67	486
33–38	0.84	51.82	22.99	8.50	282
34–39	0.17	51.06	20.45	9.12	133
35–40	2.33	31.50	15.61	8.67	15
Total	0.17	82.81	47.84	13.91	8871

Table 7.4: Descriptive statistics for the 5-week increments in fetal head circumference (mm).

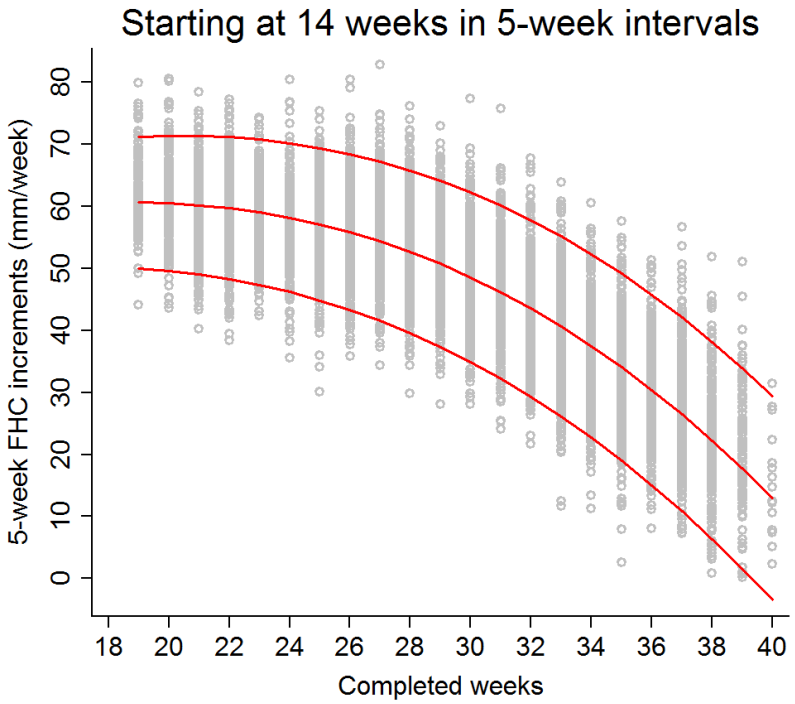


Figure 7.7: Individual 5-week increments in fetal head circumference (mm) (FHC), with the 3^{rd}, 50^{th}, and 97^{th} centiles superimposed according to gestational age (weeks).

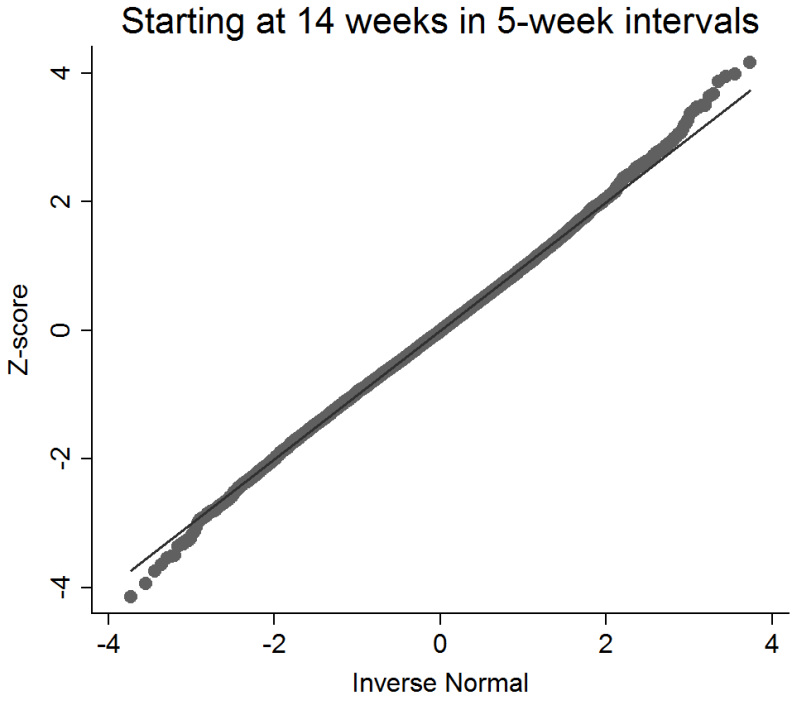


Figure 7.8: Quantile-quantile plot of the fitted model of 5-week increments in fetal head circumference (mm).

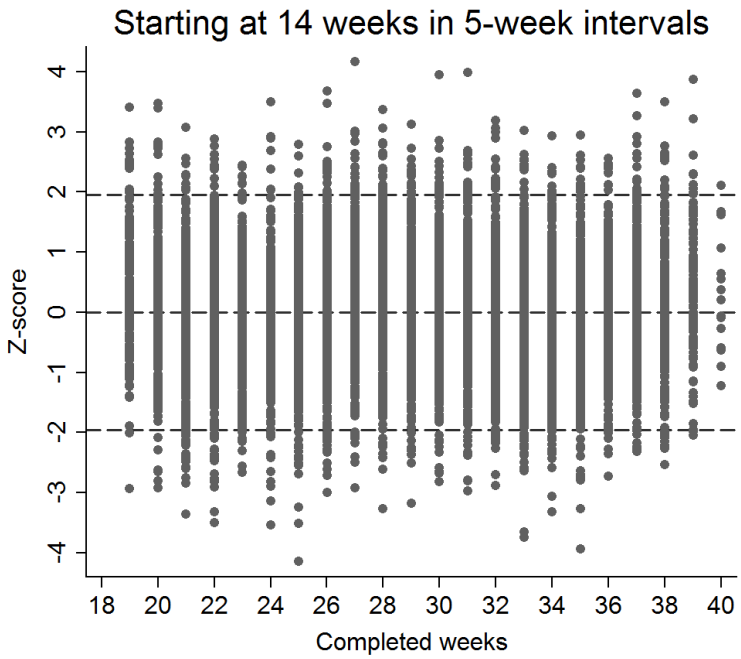


Figure 7.9: Scatter plot of individual 5-week increments in fetal head circumference (mm) from the fitted model according to gestational age in completed weeks.

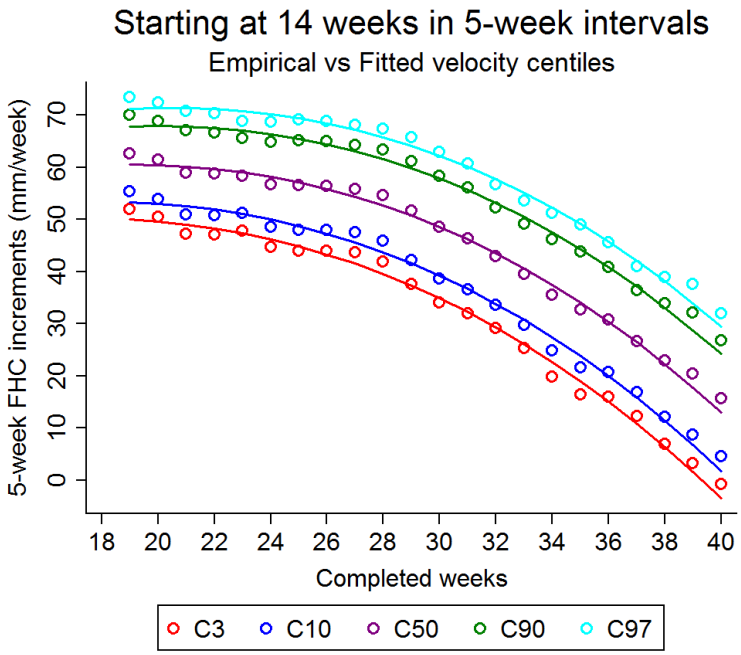


Figure 7.10: Comparison of fitted (solid lines) and empirical (open circles) 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles for the 5-week increments in fetal head circumference (mm) according to gestational age (weeks).

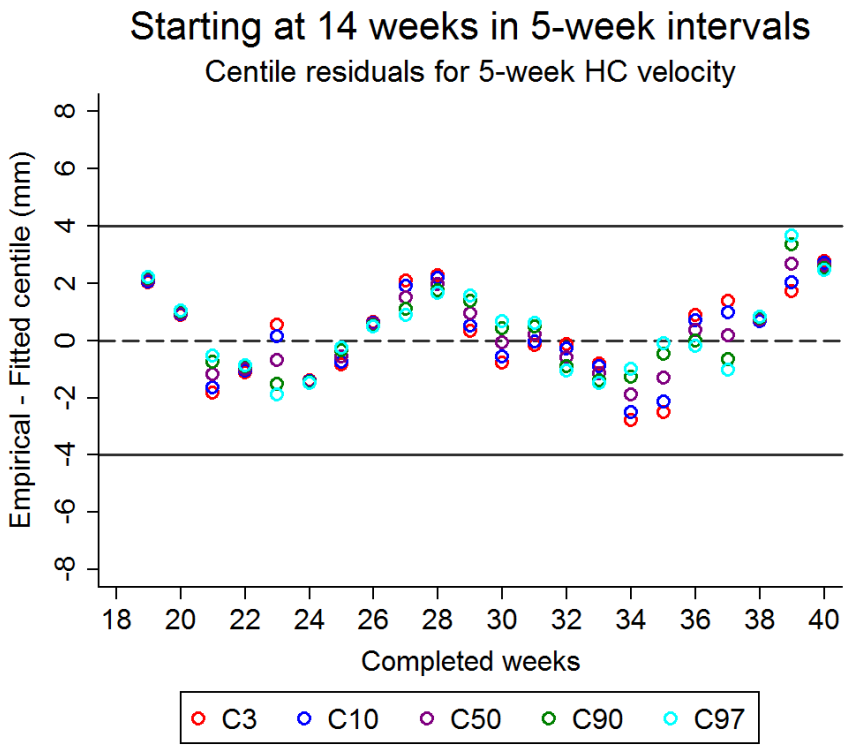


Figure 7.11: Differences in fitted and empirical centiles (residuals) for 5-week increments in fetal head circumference (mm) at the 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles according to gestational age (weeks).

Gestational age (completed weeks)	Total (n)	3^{rd} centile	10^{th} centile	50^{th} centile	90^{th} centile	97^{th} centile
14–19	190	50.01	53.39	60.63	67.86	71.25
15–20	367	49.61	53.08	60.50	67.93	71.40
16–21	465	49.03	52.59	60.21	67.82	71.38
17–22	558	48.28	51.93	59.73	67.53	71.18
18–23	532	47.33	51.07	59.07	67.06	70.80
19–24	300	46.20	50.02	58.21	66.39	70.22
20–25	425	44.86	48.77	57.14	65.52	69.43
21–26	509	43.31	47.31	55.87	64.44	68.44
22–27	592	41.54	45.64	54.39	63.14	67.23
23–28	593	39.56	43.74	52.68	61.62	65.80
24–29	360	37.35	41.62	50.75	59.88	64.15
25–30	386	34.90	39.26	48.58	57.90	62.26
26–31	444	32.22	36.67	46.18	55.69	60.13
27–32	552	29.30	33.83	43.53	53.23	57.76
28–33	576	26.12	30.75	40.63	50.52	55.15
29–34	369	22.69	27.41	37.48	47.56	52.27
30–35	385	19.01	23.81	34.08	44.34	49.14
31–36	352	15.06	19.95	30.40	40.86	45.75
32–37	486	10.84	15.82	26.46	37.11	42.09
33–38	282	6.35	11.42	22.25	33.09	38.16
34–39	133	1.58	6.74	17.77	28.79	33.95
35–40	15	-3.46	1.78	13.00	24.21	29.46
Total	8,871					

Table 7.5: Fitted 5-week increments in fetal head circumference (mm) at the 3^{rd}, 10^{th}, 50^{th}, and 97^{th} centiles according to gestational age (weeks).

7.5.1.3 Six-week increments

There were 2,411 6-week increments. The best FP model was an FP2, with powers for the mean = 2, 3 and SD = 1. A summary table with descriptive statistics for these intervals and their sample sizes is presented in Table 7.6. Figure 7.12 shows the fitted 3^{rd}, 50^{th}, and 97^{th} smoothed centiles in mm/week according to GA. The associated diagnostic results are shown in Figures 7.13 and 7.14. Figure 7.15 compares the fitted smoothed centiles with their corresponding empirical centiles for selected centiles. Differences between empirical and fitted centiles are shown in Figure 7.16. The differences were within 6 mm from 14 to 39 weeks

and were in the range of 2 mm to 8 mm at 40 weeks. Again, the greater range at 40 weeks was largely attributable to the small sample size (n = 7). Table 7.7 presents the predicted 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles for 6-week increment velocities between 14 and 40 weeks.

Gestational age (completed weeks)	Minimum (mm)	Maximum (mm)	Mean (mm)	SD (mm)	Total (n)
14–20	55.12	85.96	71.47	5.99	90
15–21	55.88	89.14	71.08	6.11	114
16–22	54.85	94.58	69.70	6.04	209
17–23	50.84	88.92	67.83	6.49	252
18–24	54.17	81.64	66.75	6.49	119
19–25	46.87	79.97	63.54	7.60	65
20–26	51.13	87.76	66.14	7.06	123
21–27	46.51	78.24	65.47	6.40	147
22–28	44.09	79.35	63.31	6.63	183
23–29	43.14	86.58	63.48	7.16	166
24–30	37.77	75.20	58.14	7.36	102
25–31	35.80	74.97	55.87	7.69	112
26–32	35.89	81.23	52.62	8.23	122
27–33	25.44	68.64	46.74	7.95	132
28–34	15.13	72.88	45.90	9.49	113
29–35	16.78	58.91	39.41	9.38	94
30–36	14.99	59.75	37.49	7.99	86
31–37	8.53	60.18	33.72	8.88	87
32–38	14.09	45.62	30.46	7.34	72
33–39	11.25	46.63	23.59	8.68	16
34–40	5.43	35.12	21.97	10.60	7
Total	5.43	94.58	58.36	14.31	2411

Table 7.6: Descriptive staistics for the 6-week increments in fetal head circumference (mm).

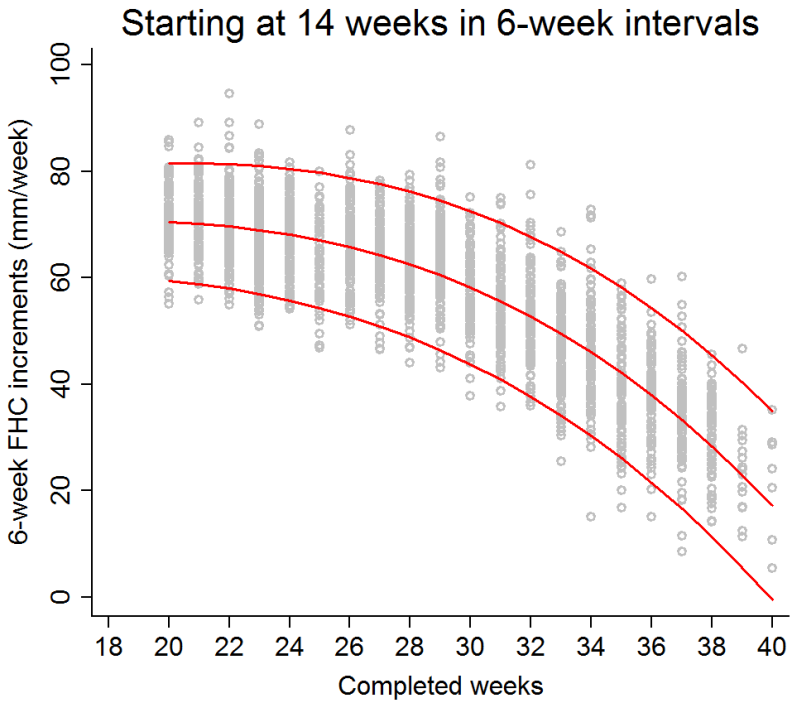


Figure 7.12: Individual 6-week increments in fetal head circumference (mm) (FHC), with the 3^{rd}, 50^{th}, and 97^{th} centiles superimposed according to gestational age (weeks).

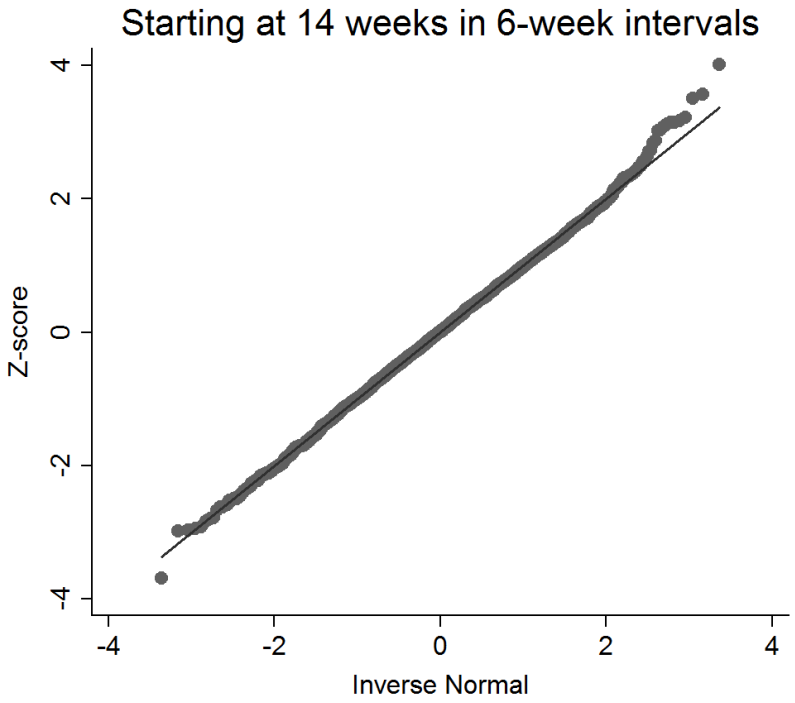


Figure 7.13: Quantile-quantile plot of the fitted model of 6-week increments in head circumference (mm).

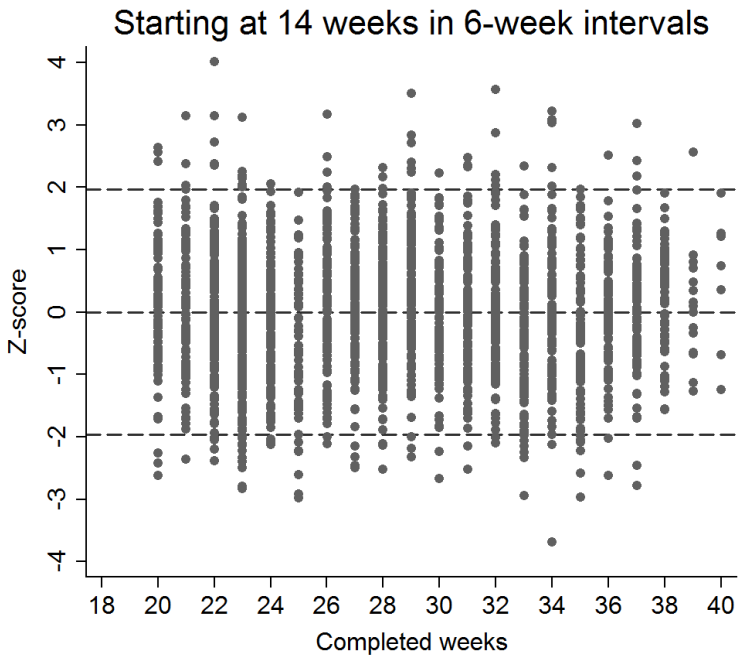


Figure 7.14: Scatter plot of individual 6-week increments in fetal head circumference (mm) from the fitted model according to gestational age in completed weeks.

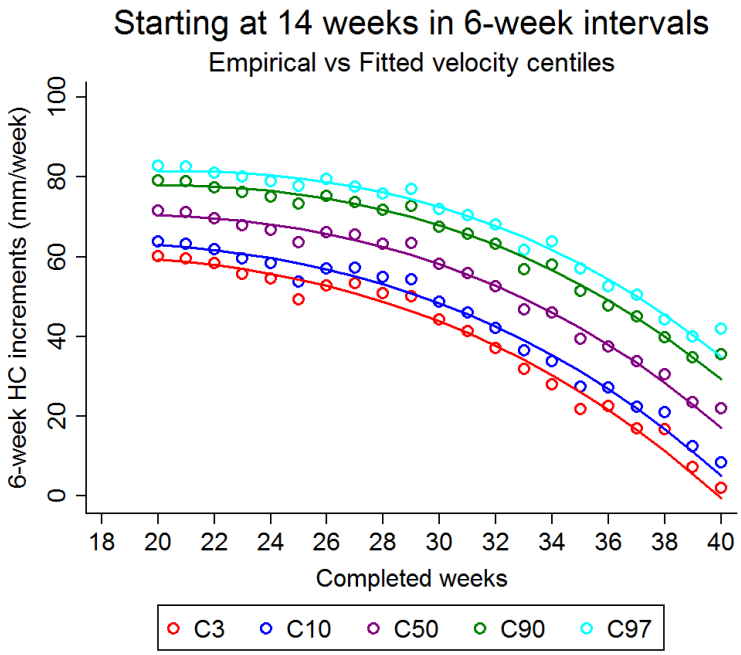


Figure 7.15: Comparison of fitted (solid lines) and empirical (open circles) 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles for the 6-week increments in fetal head circumference (mm) according to gestational age (weeks).

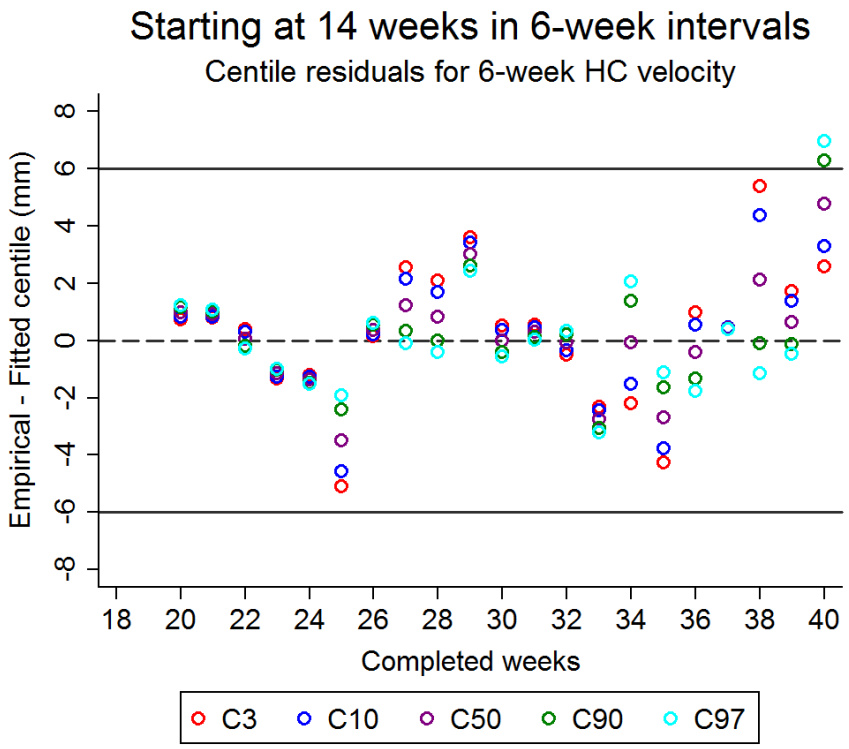


Figure 7.16: Differences in fitted and empirical centiles (residuals) for 6-week increments in fetal head circumference (mm) at 3^{rd}, 10^{th}, 50^{th}, 90^{th}, and 97^{th} centiles according to gestational age (weeks).

Gestational age (completed weeks)	Total (n)	3^{rd} centile	10^{th} centile	50^{th} centile	90^{th} centile	97^{th} centile
14–20	90	59.45	62.97	70.48	78.00	81.51
15–21	114	58.78	62.41	70.15	77.89	81.52
16–22	209	57.95	61.68	69.65	77.63	81.35
17–23	252	56.94	60.78	68.98	77.18	81.02
18–24	119	55.74	59.68	68.11	76.55	80.49
19–25	65	54.34	58.39	67.05	75.71	79.75
20–26	123	52.72	56.87	65.76	74.65	78.80
21–27	147	50.86	55.12	64.24	73.36	77.62
22–28	183	48.76	53.13	62.48	71.82	76.19
23–29	166	46.40	50.88	60.45	70.03	74.50
24–30	102	43.77	48.36	58.16	67.96	72.54
25–31	112	40.85	45.55	55.58	65.61	70.30
26–32	122	37.64	42.43	52.69	62.95	67.75
27–33	132	34.11	39.01	49.50	59.99	64.89
28–34	113	30.25	35.26	45.97	56.69	61.70
29–35	94	26.05	31.16	42.11	53.06	58.17
30–36	86	21.49	26.72	37.89	49.07	54.29
31–37	87	16.57	21.90	33.30	44.71	50.04
32–38	72	11.26	16.70	28.33	39.97	45.40
33–39	16	5.56	11.11	22.97	34.83	40.37
34–40	7	-0.55	5.10	17.19	29.28	34.93
Total	2,411					

Table 7.7: Fitted 6-week increments in fetal head circumference (mm) at the 3^{rd}, 10^{th}, 50^{th}, and 97^{th} centiles according to gestational age (weeks).

7.5.2 Velocity gain z-scores

Figure 7.17 shows a plot of FHC according to GA in completed weeks. It clearly illustrates how measurements were taken at every point of gestation from 14 to 40 weeks. A histogram and scatter plots of the untransformed empirical correlations and Fisher's transformed correlation are shown in Figures 7.18 and 7.19. Plots of empirical and fitted correlations starting at 14 weeks versus time, t_2 connected by points with the same t_1, are shown in Figures 7.20 and 7.21. Figure 7.22 compares empirical and fitted correlations for 4-, 5-, and 6-week correlations starting at 14 weeks GA. Figure 7.23 compares empirical and smoothed fitted correlations.

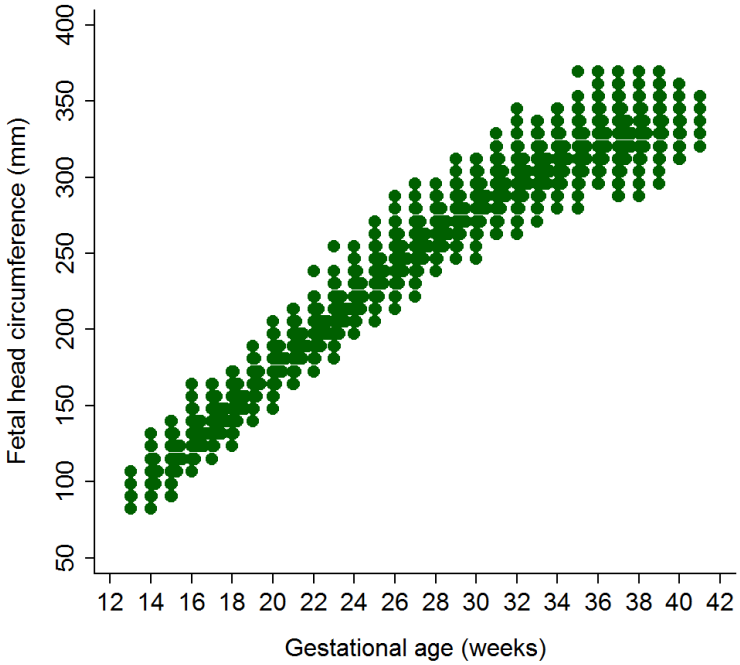


Figure 7.17: Scatter plot of fetal head circumference (mm) according to gestational age in completed weeks.

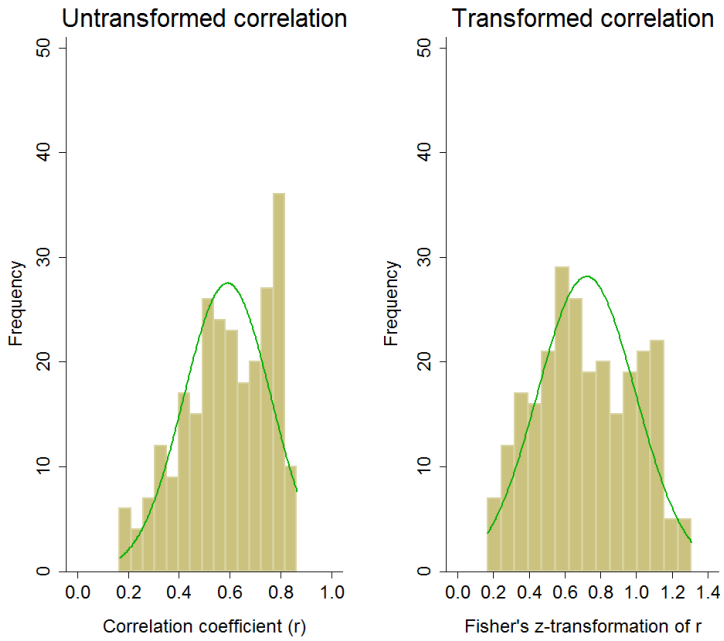


Figure 7.18: Histograms of fetal head circumference empirical correlations (untransformed, left) and Fisher's transformed correlations (right) for sets of paired fetal head circumference data

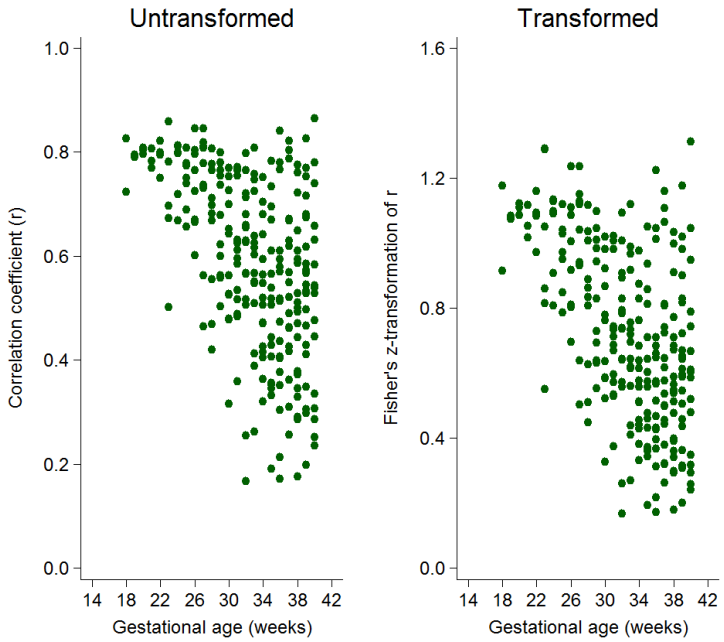


Figure 7.19: Scatter plot of fetal head circumference empirical correlations (untransformed, left) and Fisher's transformed correlations (right) for sets of paired fetal head circumference data by gestational age.

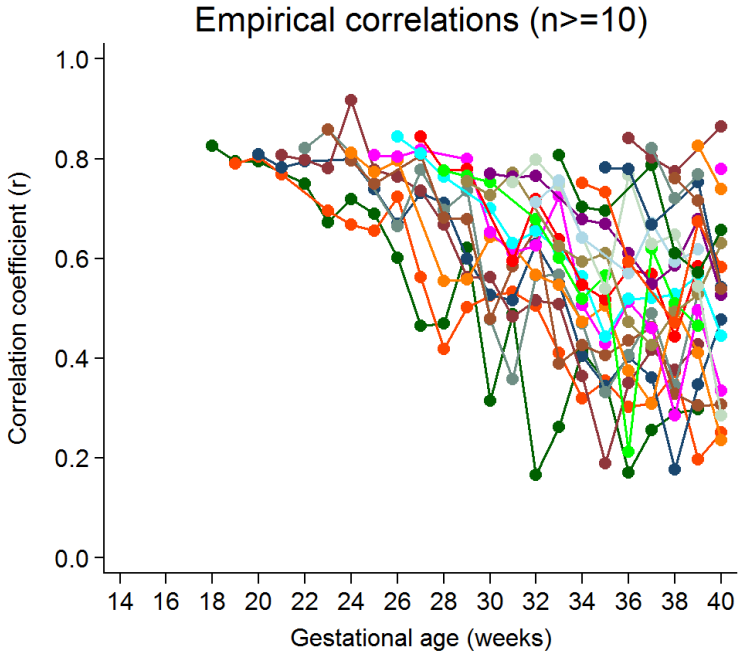


Figure 7.20: Plot of fetal head circumference empirical correlations starting at 14 weeks versus time, t_2. Points with the same t_1 are connected. Each colour represents a separate starting point, for example a starting point of 14 weeks is shown in green.

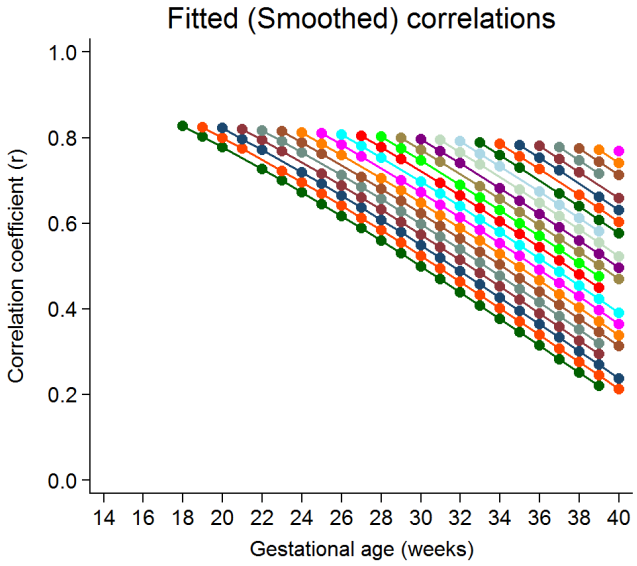


Figure 7.21: Fitted correlations from Table 7.10 starting at 14 weeks versus time, t_2 (i.e., gestational age). Points with the same $t_1$are connected. Each colour represents a separate starting point, for example a starting point of 14 weeks is shown in green.

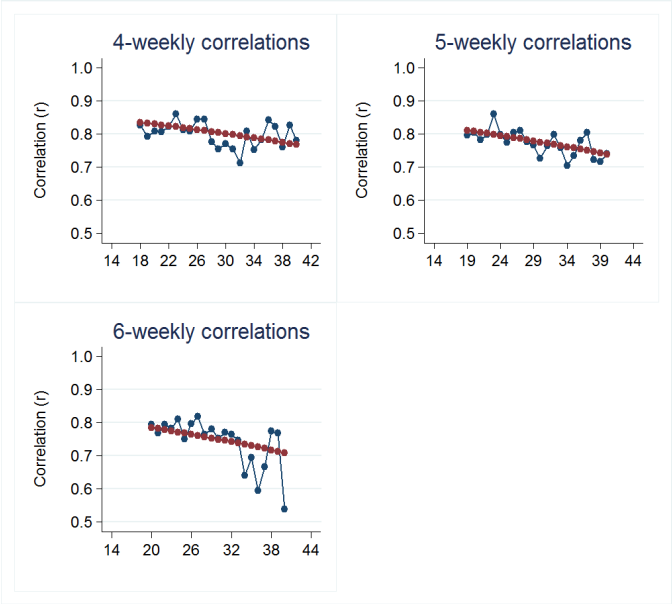


Figure 7.22: Comparing empirical (blue closed circles) with fitted (red closed circles) correlations for sets of paired fetal head circumference data for 4-, 5-, and 6-week correlations starting at 14 weeks gestational age. Correlations obtained from the 4-week increments refer to gestational ages 14–18, 15–19, . . . , 36–40 weeks. Correlations obtained from the 5-week increments refer to gestational ages 14–19, 15–20, . . . , 35–40 weeks. Correlations obtained from the 6-week increments refer to gestational ages 14–20, 15–21, . . . , 34–40 weeks.

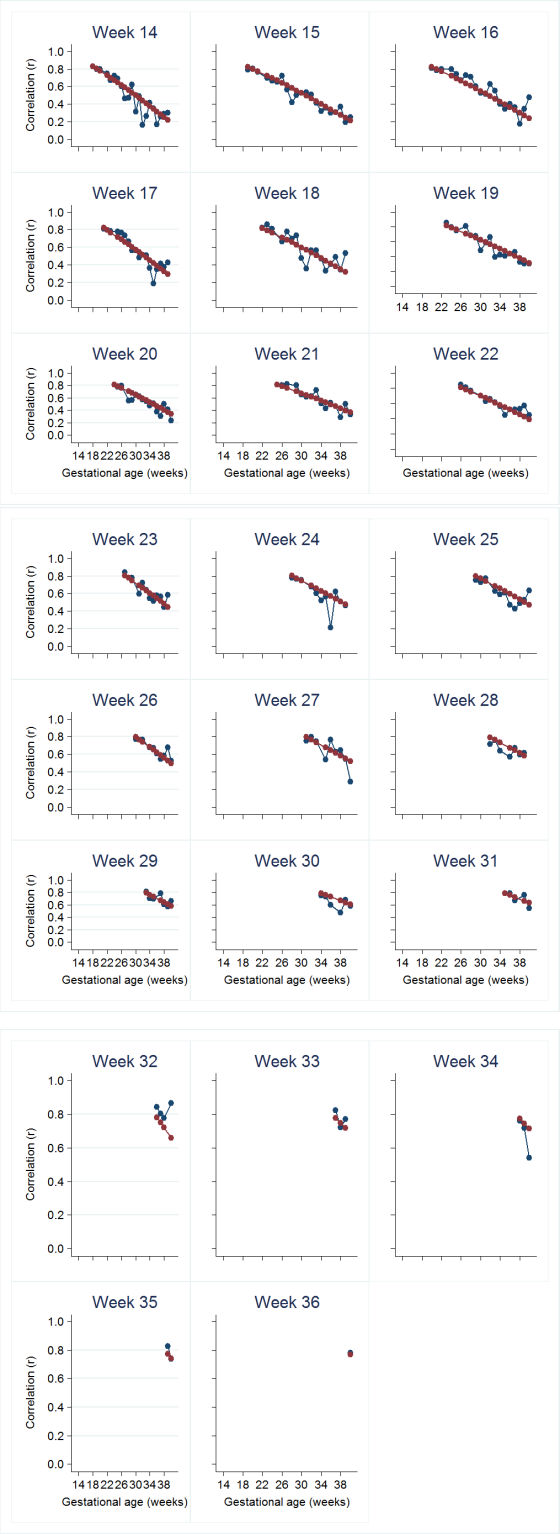


Figure 7.23: Plots comparing empirical (blue closed circles) versus smoothed fitted (red closed circles) correlations for each completed week of gestation for head circumference z-scores.

7.5.3 Evidence of regression to the mean

The correlations below the diagonal in Tables 7.8 and 7.9 measure regressions to the mean from one GA week to the next. For example, between 14 and 18 weeks, the estimated correlation was 0.827. A typical fetus with an FHC Z_1 = -2 at 14 weeks would regress to the mean by this amount, so the expected Z_2 at 18 weeks is $-2 \times 0.827 = -1.654$. Between 18 and 22 weeks, the estimated correlation was 0.817. The same fetus would then be expected to move from $Z_1 = -1.654$ to $Z_2 = -1.654 \times 0.817 = -1.351$.

In general, the amount of regression to the mean over a series of measurements taken every four weeks is obtained by multiplying the correlations between adjacent measurements together. If measurements are taken every four weeks from 14 to 40 weeks, the amount of regression to the mean can be obtained by multiplying the six correlations below the diagonal together, giving 0.264. An average fetus with an FHC Z_1 of -2 at 14 weeks would end up with Z_2 near $-2 \times 0.264 = -0.528$ at 40 weeks, which corresponds to approximately the 30^{th} centile. Similarly, by symmetry, fetuses starting at an FHC Z_1 of +2 will tend to shift down to the same extent. Table 7.10 shows the FP regression analysis equation of the transformed correlation between successive FHC z-scores as a function of the time interval between measurements and the fetus's mean GA (both measured in completed weeks).

In the INTERGROWTH-21^{st} study, 359, 786, and 218 fetuses were seen every 4, 5, or 6 weeks, respectively, from 14 weeks of gestation. The results were consistent with expected regression to the mean across all three intervals (4-, 5-, and 6-week intervals), as the FHC drifted towards the median. The majority of the fetuses with z-scores between +2.5 and +1.5 drifted downwards towards the median for all three intervals (74.6%, 72.3%, and 69.3% of fetuses seen every 4, 5, and 6 weeks, respectively). Similarly, 70.7%, 76.8%, and 76.2% of fetuses with z-scores between

-2.5 and -1.5 seen every 4, 5, and 6 weeks, respectively, drifted upwards towards the median (Tables 7.11–7.13). An example of observed FHC z-score measurements, expected z-scores accounting for previous measurements, and velocity gain z-scores for a fetus in the INTERGROWTH-21st study is shown in Figure 7.24.

weeks	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
14	1.00																										
15		1.00																									
16			1.00																								
17				1.00																							
18	0.83				1.00																						
19	0.79	0.79				1.00																					
20	0.80	0.80	0.81	0.77			1.00																				
21		0.77	0.78	0.81				1.00																			
22	0.75	0.91	0.79	0.80	0.82				1.00																		
23	0.67	0.70	0.94	0.78	0.86	0.86				1.00																	
24	0.72	0.67	0.80	0.92	0.81	0.80	0.81				1.00																
25	0.69	0.66	0.74	0.78	0.43	0.75	0.77	0.81				1.00															
26	0.60	0.72	0.67	0.76	0.67	0.96	0.80	0.80	0.84				1.00														
27	0.46	0.56	0.73	0.74	0.78	0.81	0.85	0.82	0.81	0.84				1.00													
28	0.47	0.42	0.71	0.67	0.70	0.68	0.55	0.55	0.76	0.78	0.78				1.00												
29	0.62	0.50	0.60	0.56	0.74	0.68	0.56	0.80	0.80	0.78	0.77	0.75				1.00											
30	0.32	0.53	0.53	0.56	0.48	0.48	0.64	0.65	0.70	1.00	0.75	0.73	0.77				1.00										
31	0.49	0.53	0.52	0.48	0.36	0.58	0.63	0.61	0.63	0.59	0.51	0.77	0.76	0.75	0.82			1.00									
32	0.17	0.51	0.63	0.52	0.57	0.66	0.57	0.63	0.65	0.72	0.68	0.26	0.76	0.80	0.71				1.00								
33	0.26	0.41	0.55	0.51	0.57	0.39	0.55	0.72	0.62	0.64	0.60	0.63	0.30	0.75	0.76	0.81	0.68			1.00							
34	0.41	0.32	0.40	0.36	0.47	0.43	0.47	0.51	0.56	0.55	0.52	0.59	0.68	0.88	0.64	0.70	0.75	0.78			1.00						
35	0.35	0.36	0.34	0.19	0.33	0.41	0.51	0.43	0.44	0.52	0.57	0.61	0.67	0.54	0.68	0.69	0.73	0.78	0.56			1.00					
36	0.17	0.30	0.40	0.35	0.41	0.44	0.38	0.51	0.52	0.58	0.21	0.47	0.61	0.77	0.57	0.45	0.59	0.78	0.84	0.83			1.00				
37	0.26	0.31	0.36	0.42	0.49	0.46	0.31	0.46	0.52	0.57	0.62	0.43	0.55	0.63	0.67	0.79	0.65	0.67	0.80	0.82	0.70			1.00			
38	0.29	0.37	0.18	0.38	0.35	0.33	0.50	0.28	0.53	0.44	0.51	0.49	0.59	0.65	0.59	0.61	0.47	0.82	0.77	0.72	0.76	0.74			1.00		
39	0.30	0.20	0.35	0.43	0.53	0.31	0.41	0.50	0.57	0.58	0.47	0.53	0.68	0.54	0.62	0.57	0.67	0.75		0.77	0.72	0.83				1.00	
40		0.25	0.48			0.31	0.24	0.33	0.44	0.33		0.63	0.53	0.29		0.66	0.58	0.54	0.86		0.54	0.74	0.78				1.00

Table 7.8: Empirical correlation matrix for head circumference z-scores from 14 to 40 weeks gestation

weeks	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
14	1.00																										
15		1.00																									
16			1.00																								
17				1.00																							
18	0.86				1.00																						
19	0.84	0.85				1.00																					
20	0.83	0.84	0.85	0.85			1.00																				
21		0.82	0.83	0.84				1.00																			
22	0.78	0.80	0.82	0.83	0.84				1.00																		
23	0.75	0.77	0.79	0.81	0.82	0.83				1.00																	
24	0.72	0.74	0.77	0.79	0.81	0.82	0.83				1.00																
25	0.68	0.71	0.74	0.76	0.78	0.80	0.81	0.82				1.00															
26	0.64	0.67	0.70	0.73	0.76	0.78	0.79	0.81	0.82				1.00														
27	0.60	0.63	0.67	0.70	0.73	0.75	0.77	0.79	0.80	0.81				1.00													
28	0.55	0.59	0.62	0.66	0.69	0.72	0.74	0.76	0.78	0.79	0.80				1.00												
29	0.51	0.55	0.58	0.62	0.65	0.68	0.71	0.74	0.76	0.78	0.79	0.80				1.00											
30	0.47	0.50	0.54	0.57	0.61	0.64	0.68	0.70	0.73	0.75	0.77	0.78	0.79				1.00										
31	0.44	0.46	0.50	0.53	0.57	0.60	0.64	0.67	0.70	0.72	0.74	0.76	0.78	0.78	0.79			1.00									
32	0.41	0.43	0.46	0.49	0.52	0.56	0.60	0.63	0.66	0.69	0.72	0.74	0.75	0.77	0.78				1.00								
33	0.39	0.40	0.42	0.45	0.48	0.52	0.55	0.59	0.62	0.65	0.68	0.71	0.73	0.75	0.76	0.77	0.78			1.00							
34	0.38	0.38	0.40	0.42	0.44	0.47	0.51	0.54	0.58	0.61	0.65	0.68	0.70	0.72	0.74	0.75	0.76	0.77			1.00						
35	0.39	0.38	0.38	0.39	0.41	0.44	0.47	0.50	0.54	0.57	0.61	0.64	0.67	0.69	0.72	0.73	0.75	0.76	0.76			1.00					
36	0.40	0.38	0.37	0.37	0.38	0.40	0.43	0.46	0.49	0.53	0.56	0.60	0.63	0.66	0.69	0.71	0.73	0.74	0.75	0.75			1.00				
37	0.44	0.40	0.38	0.37	0.37	0.38	0.40	0.42	0.45	0.49	0.52	0.56	0.59	0.62	0.65	0.68	0.70	0.72	0.73	0.74	0.74			1.00			
38	0.48	0.43	0.40	0.37	0.36	0.36	0.37	0.39	0.42	0.45	0.48	0.51	0.55	0.58	0.62	0.64	0.67	0.69	0.71	0.72	0.73	0.74			1.00		
39	0.54	0.48	0.43	0.39	0.37	0.36	0.36	0.37	0.38	0.41	0.44	0.47	0.51	0.54	0.57	0.61	0.64	0.66		0.70	0.72	0.72				1.00	
40		0.54	0.48			0.37	0.35	0.35	0.36	0.38		0.43	0.46	0.50		0.57	0.60	0.63	0.65		0.69	0.71	0.72				1.00

Table 7.9: Fitted (smoothed) correlation matrix for head circumference z-scores from 14 to 40 weeks gestation.

Variable	Coefficient	Standard error	t-statistic	P-value
Constant	1.945	0.00812	85.16	<0.001
Time gap$^{1/2}$	-0.302	0.0310	-29.75	< 0.001
Mean gestational age	-0.00864	0.00168	-4.34	< 0.001

The dependent variable is Fisher's transformed correlation. To obtain the corresponding correlation, the inverse of the transformation must be calculated:

$$\mathrm{r} = \frac{\exp(2\mathrm{Z}) - 1}{\exp(2\mathrm{Z}) + 1}. \tag{7.11}$$

For example, the correlation between 14 and 18 weeks covers a gap of 4 weeks and a mean GA of 16 weeks. Substituting these values into the regression equation in Table 7.10:
Z = 1.945-0.302 $\times(4^{1/2}) - 0.00864 \times 16 = 1.245$

r = $\frac{\exp(2\mathrm{Z})-1}{\exp(2\mathrm{Z})+1,}$

$$\mathrm{r} = \frac{\exp(2 \times 1.245) - 1}{\exp\ (2 \times 1.245) + 1} = 0.834$$

Table 7.10: Regression analysis equation of the transformed correlation between successive fetal head circumference z-scores, as a function of the time interval between measurements and the fetus's mean gestational age (both measured in completed weeks).

4-week intervals	Starting at between 1.5 and 2.5 z-scores					Starting at between -2.5 and -1.5 z-scores				
Gestational age (complete weeks)	Sample	No regression to the mean (n = 92)	Proportion	Average change in z-score overall (n = 359)	Average change in z-score for those that don't show regression to the mean (n = 92)	Sample	No regression to the mean (n = 68)	Proportion	Average change in z-score overall (n = 232)	Average change in z-score for those that don't show regression to the mean (n = 68)
14–18	10	2	20.00	-0.47	0.19	2	1	50.00	0.10	-0.44
15–19	12	6	50.00	-0.36	0.28	9	2	22.22	0.56	-0.31
16–20	27	3	11.11	-0.73	0.08	9	2	22.22	0.32	-0.25
17–21	31	5	16.13	-0.67	0.28	17	5	29.41	0.32	-0.32
18–22	15	3	20.00	-0.53	0.44	14	3	21.43	0.36	-0.35
19–23	4	0	0.00	-0.21	NA	4	2	50.00	0.04	-0.43
20–24	7	2	28.57	-0.19	0.56	7	3	42.86	0.21	-0.28
21–25	13	4	30.77	-0.42	0.39	9	2	22.22	0.41	-0.19
22–26	12	5	41.67	-0.20	0.52	11	3	27.27	0.34	-0.12
23–27	7	3	42.86	0.13	0.63	13	3	23.08	0.32	-0.23
24–28	4	2	50.00	-0.14	0.31	5	1	20.00	0.73	-0.37
25–29	5	3	60.00	-0.17	0.26	9	3	33.33	0.34	-0.37
26–30	16	6	37.50	-0.31	0.49	13	3	23.08	0.68	-0.51
27–31	11	6	54.55	-0.05	0.52	5	3	60.00	0.09	-0.51
28–32	18	3	16.67	-0.66	0.61	9	3	33.33	0.06	-0.41
29–33	17	3	17.65	-045	0.52	8	3	37.50	0.14	-0.49
30–34	19	3	15.79	-0.73	0.19	12	5	41.67	0.33	-0.35
31–35	33	5	15.15	-0.57	0.84	17	3	17.65	0.44	-0.44
32–36	34	10	29.41	-0.44	0.38	16	5	31.25	0.25	-0.34
33–37	36	7	19.44	-0.56	0.30	24	8	33.33	0.22	-0.29
34–38	18	6	33.33	-0.49	0.31	12	2	16.67	0.74	-0.62
35–39	10	5	50.00	-0.20	0.26	7	3	42.86	0.06	-0.48
Total	359	92	25.63	-0.48	0.40	232	68	29.31	0.34	-0.36

Table 7.11: The number and percentage of fetuses not falling below or rising above their starting z-scores after 4 weeks (i.e., not showing an indication of regression to the mean).

5-week intervals	Starting at between 1.5 and 2.5 z-scores					Starting at between -2.5 and -1.5 z-scores				
Gestational age (complete weeks)	Sample	No regression to the mean (n = 218)	Proportion	Average change in z-score overall (n = 786)	Average change in z-score for those that don't show regression to the mean (n = 218)	Sample	No regression to the mean (n = 133)	Proportion	Average change in z-score overall (n = 574)	Average change in z-score for those that don't show regression to the mean (n = 133)
14–19	29	4	13.79	-0.62	0.25	10	1	10.00	0.76	-0.14
15–20	43	13	30.23	-0.42	0.29	27	3	11.11	0.66	-0.22
16–21	63	14	22.22	-0.54	0.44	30	8	26.67	0.40	-0.44
17–22	44	5	11.36	-0.80	0.34	35	6	17.14	0.30	-0.41
18–23	43	8	18.60	-0.46	0.33	40	14	35.00	0.24	-0.29
19–24	23	3	13.04	-0.67	0.35	11	4	36.36	0.26	-0.35
20–25	36	9	25.00	-0.51	0.35	20	6	30.00	0.52	-0.27
21–26	35	19	54.29	0.01	0.43	39	11	28.21	0.35	-0.26
22–27	32	16	50.00	0.03	0.51	50	12	24.00	0.40	-0.32
23–28	37	15	40.54	-0.21	0.41	57	4	7.02	0.77	-0.50
24–29	23	15	65.22	0.25	0.75	32	5	15.63	0.65	-0.76
25–30	25	7	28.00	-0.38	0.37	24	3	12.50	0.70	-0.55
26–31	50	17	34.00	-0.35	0.47	28	6	21.43	0.81	-0.11
27–32	50	18	36.00	-0.38	0.20	37	10	27.03	0.38	-0.39
28–33	58	15	25.86	-0.49	0.40	29	14	48.28	0.19	-0.40
29–34	32	5	15.63	-0.82	0.63	19	7	36.84	0.37	-0.31
30–35	39	7	17.95	-0.84	0.49	19	3	15.79	0.66	-0.32
31–36	38	11	28.95	-0.45	0.36	14	5	35.71	0.36	-0.18
32–37	50	11	22.00	-0.59	0.36	25	10	40.00	0.17	-0.44
33–38	22	4	18.18	-0.53	0.37	17	1	5.88	0.74	-0.98
34–39	13	2	15.38	-0.63	0.32	10	NA	NA	0.96	NA
35–40	1	0	0.00	-0.90	NA	1	NA	NA	1.31	NA
Total	786	218	27.74	-0.46	0.41	574	133	23.17	0.49	-0.36

Table 7.12: The number and percentage of fetuses not falling below or rising above their starting z-scores after 5 weeks (i.e., not showing an indication of regression to the mean).

6-week intervals	Starting at between 1.5 and 2.5 z-scores					Starting at between -2.5 and -1.5 z-scores				
Gestational age (complete weeks)	Sample	No regression to the mean (n = 67)	Proportion	Average change in z-score overall (n = 218)	Average change in z-score for those that don't show regression to the mean (n = 67)	Sample	No regression to the mean (n = 39)	Proportion	Average change in z-score overall (n = 164)	Average change in z-score for those that don't show regression to the mean (n = 39)
14–20	12	2	16.67	-0.74	0.20	4	1	25.00	0.94	-0.33
15–21	14	4	28.57	-0.62	0.23	4	1	25.00	0.45	-0.19
16–22	13	3	23.08	-0.50	0.17	22	9	40.91	0.28	-0.40
17–23	21	4	19.05	-0.49	0.70	16	5	31.25	0.26	-0.24
18–24	7	1	14.29	-0.53	0.42	11	7	63.64	0.05	-0.28
19–25	5	NA	NA	-0.58	NA	4	2	50.00	0.30	-0.05
20–26	19	10	52.63	-0.03	0.42	6	NA	NA	0.54	NA
21–27	10	7	70.00	0.26	0.65	14	1	7.14	0.52	-0.90
22–28	10	5	50.00	-0.14	0.30	15	NA	NA	1.05	NA
23–29	6	5	83.33	0.46	0.58	19	2	10.53	0.81	-0.31
24–30	8	2	25.00	-0.50	0.32	10	4	40.00	0.61	-0.17
25–31	12	7	58.33	0.22	0.77	6	NA	NA	0.95	NA
26–32	20	5	25.00	-0.27	1.15	6	1	16.67	0.42	-0.45
27–33	13	3	23.08	-0.56	0.63	8	3	37.50	0.32	-0.23
28–34	10	2	20.00	-0.89	0.22	5	1	20.00	0.36	-0.31
29–35	6	2	33.33	-0.33	0.65	5	1	20.00	0.41	-0.12
30–36	7	1	14.29	-0.37	0.07	2	NA	NA	0.85	NA
31–37	12	2	16.67	-0.51	0.21	3	1	33.33	0.52	-0.36
32–38	10	2	20.00	-0.52	0.13	2	NA	NA	0.97	NA
33–39	2	NA	NA	-0.29	NA	1	NA	NA	1.82	NA
34–40	1	NA	NA	-1.55	NA	1	NA	NA	0.64	NA
Total	218	67	30.73	-0.36	0.51	164	39	23.78	0.54	-0.29

Table 7.13: The number and percentage of fetuses not falling below or rising above their starting z-scores after 6 weeks (i.e., not showing an indication of regression to the mean).

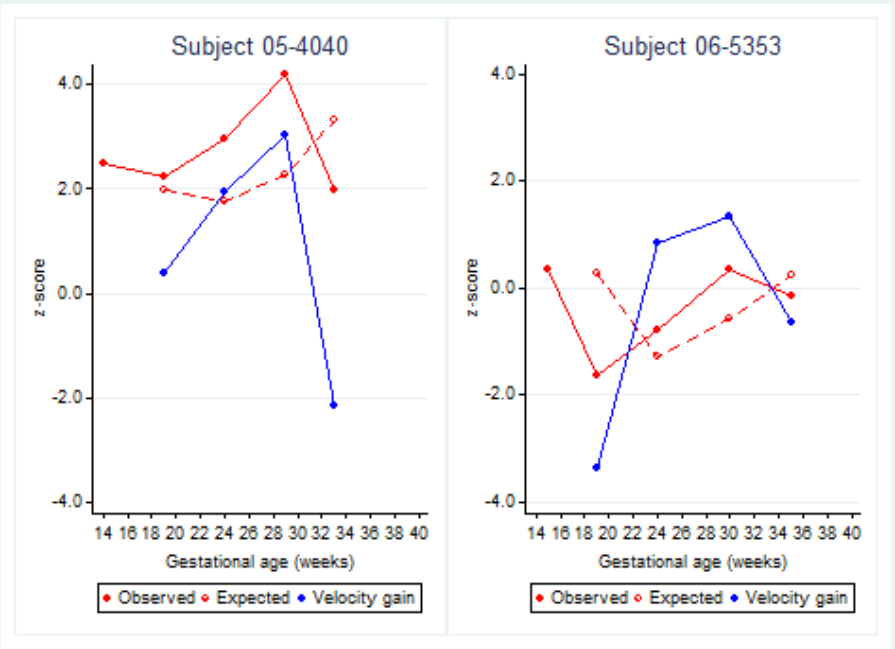


Figure 7.24: INTERGROWTH-21st fetus series of head circumference z-score measurements (red solid circles and solid red line), expected z-scores accounting for previous measurements (red open circles and red dashed line), and velocity gain z-scores (solid blue circles and solid blue line) for two representative INTERGROWTH-21st fetuses.

7.5.4 Comparing velocity or increment reference values and velocity gain z-score approaches

The density distribution plots of the z-scores of the two approaches were generally similar (Figure 7.25). Ninety percent of the differences between velocity gain and velocity increment z-scores were between -0.93 and +0.71 z-scores for 4-weekly correlations, between -0.76 and +0.87 z-scores for 5-weekly correlations, and between -0.72 and +1.07 z-scores for 6-weekly correlations with a start point of 14 weeks (Figure 7.26 and Table 7.14).

The WHO-MGRS reported differences observed between the two methods to be between -0.33 and 0.34 z-scores at the 5^{th} and 95^{th} centiles respectively (90% of the data were within these limits) in children aged up to 2 years [273]. When applied to the INTERGROWTH-21^{st} data,these limits corresponded to the 28^{th} and 81^{st} centiles for 4-weekly correlations (including only 53% of the INTERGROWTH-21^{st} population), 21^{st} and 76^{th} centiles for 5-weekly correlations (including only 55% of the INTERGROWTH-21^{st} population), and 22^{nd} and 68^{th} centiles for 6-weekly correlations (including only 46% of the INTERGROWTH-21^{st} population).

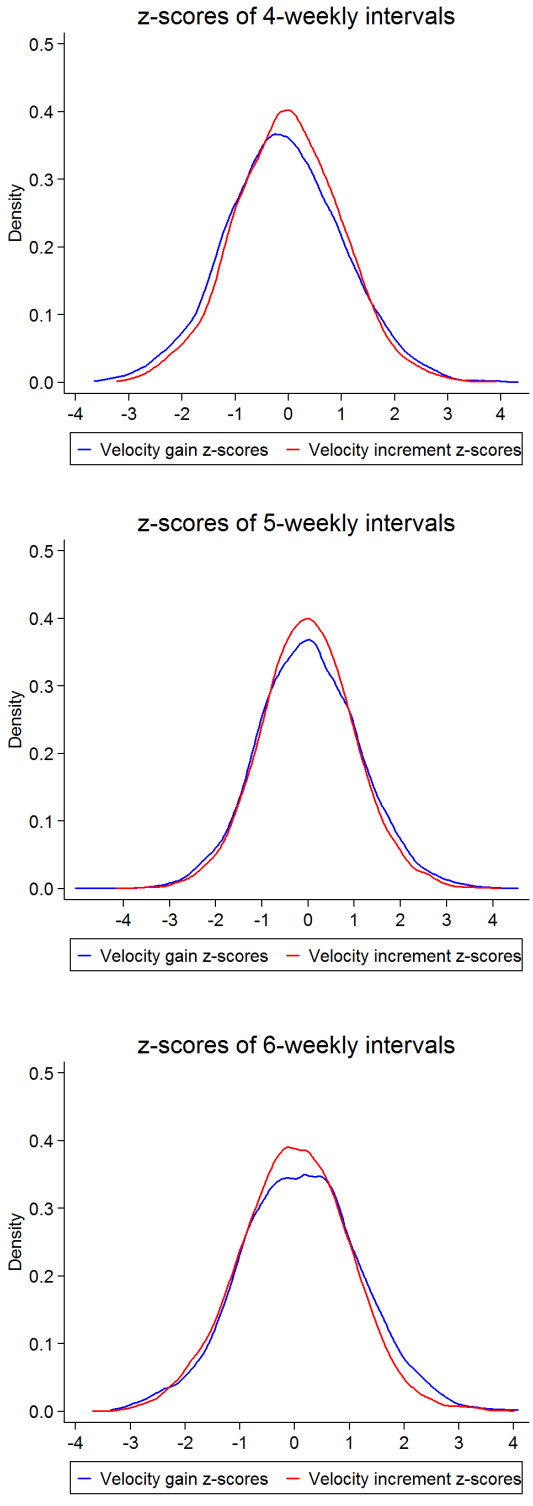


Figure 7.25: Comparison between velocity gain z-scores and velocity increments using (A) 4-week intervals, (B) 5-week intervals, and (C) 6-week intervals.

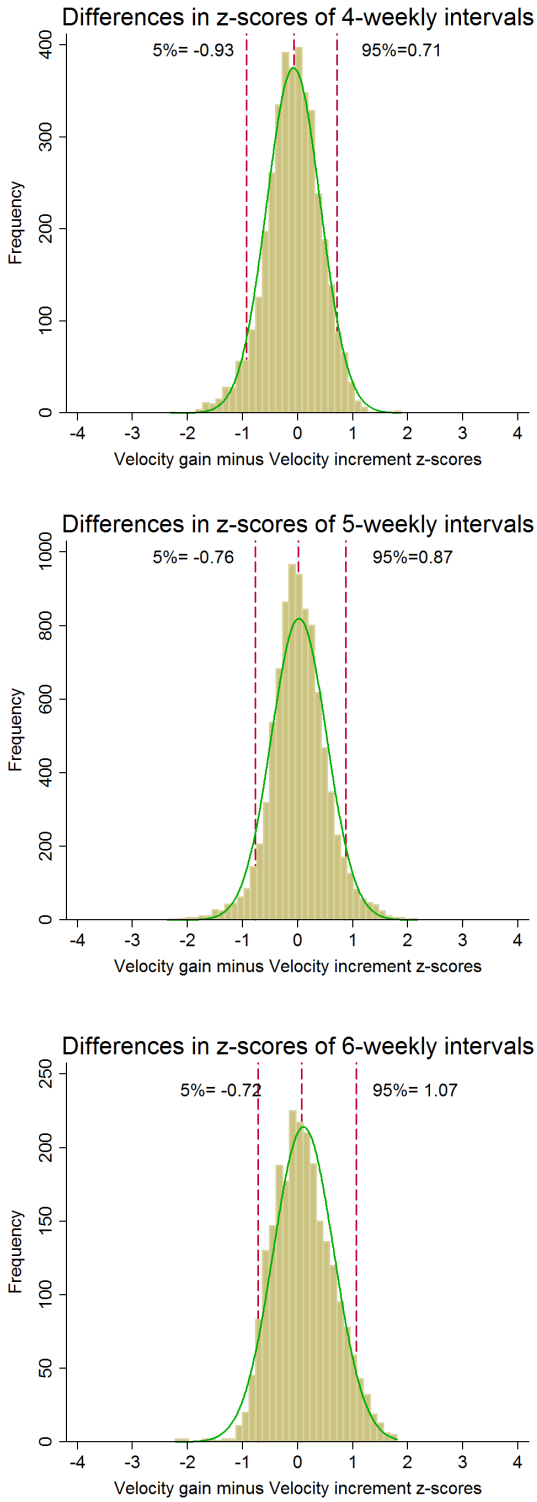


Figure 7.26: Differences between velocity gain z-scores and velocity increments using (A) 4-week intervals, (B) 5-week intervals, and (C) 6-week intervals.

Interval lengths	Distribution of pairwise differences between the velocity gain z-scores and velocity increment z-scores			
	5%	Median	95%	Average
4–weeks	-0.93	-0.06	0.71	0.82
5–weeks	-0.76	0.01	0.87	0.82
6–weeks	-0.72	0.07	1.07	0.90
7–weeks	-0.98	0.37	1.40	1.19
8–weeks	-0.51	0.39	1.27	0.89
9–weeks	-0.91	-0.10	0.58	0.75
10–weeks	-0.55	0.12	0.84	0.70
11–weeks	-0.99	0.08	0.94	0.965
12–weeks	-1.84	0.70	2.64	2.24

Table 7.14: Summary of the pairwise differences between the velocity gain approach and velocity increment approach, starting at 14 weeks gestational age.

7.6 Discussion

Though remarkable progress has been made in improving fetal, neonatal, and child health since the turn of the century, there is still considerable improvement to be achieved. Every day an estimated 16,000 children <5 years of age die, amassing to 5.9 million deaths annually. The vast majority of these deaths could and should have been avoided [295]. Most of them can be attributed, either directly or indirectly, to undernutrition [296].

The study by Schwinger *et al.* [297] assessed whether growth velocity was a good indicator for predicting child mortality. Their growth assessments were based on the WHO child growth velocity standards [273] and were applied to a historical cohort data in the Democratic Republic of Congo(1989–1991) to predict adverse health outcomes. They found that weight and length velocity z-scores had better predictive abilities (area under the curve = 0.67 and 0.69, respectively) than static measures of weight-for-age (area under the curve = 0.57) and length-for-age (area under the curve = 0.52) z-scores. They concluded that assessing a child's recent growth trajectory provided a more accurate prognosis of likely death than the use of static measures

of nutritional status. Although repeated growth measures are slightly more complex to implement, their ability to predict mortality suggests that they could be used for identifying children at increased risk of death. Velocity standards would be valuable for improving the prediction of poor perinatal outcomes, assessing individuals' progress, and identifying those at risk of requiring medical intervention based on changes in growth observed from previous ultrasound measurement [298, 299, 80].

This chapter focused on fetal growth velocity and how best to calculate it. The methods described assess change in fetal size measurements using centiles or z-scores. Two approaches for calculating velocity were discussed and applied to the FGLS data from the INTERGROWTH-21st Project [24, 11]. The methodologies discussed have previously been applied to child growth data [242, 243, 273], but not to fetal data, to the best of my knowledge. The primary objective was to construct, for the first time, fetal growth velocity standards. Such standards were not currently available largely due to lack of appropriate populations studied following the WHO recommended prescriptive approach and good quality longitudinal fetal data. Growth velocity has been shown to be a more sensitive indicator of growth faltering than attained size and therefore has the potential to contribute to effective pregnancy management when used alongside attained size charts [79].

Velocity increments are limited by their rigidity, as they require estimates to be developed within specified intervals, are only applicable in the specified fixed time intervals, are subject to measurement error, and do not account for regression to the mean [241]. Their use is not practical in clinical settings, as measurements are usually taken at unstructured times. If an individual has a measurement time outside the intervals covered in the tables, clinicians have to rely on estimates that closely approximate the interval of interest, with the risk of biased estimates.

An alternative approach is the velocity gain first proposed by Wright [242] and later used by Cole using routinely collected data [243]. The velocity gain approach

is flexible and does not require fixed time intervals. However, large variations are expected over short periods mainly due to measurement error [243]. This is a practical problem in antenatal care as repeated measures are done in fetuses suspected to be at risk to decide appropriate clinical interventions. Determining the appropriate time interval between successive measurements is a trade-off between the noise-to-signal ratio with the goal of minimising measurement error, to avoid biased estimates.

The calculation of a velocity gain depends on (a) the fetal measurement at t_1, (b) the time interval between t_1 and t_2, and (c) the correlation of the fetal measurements between t_1 and t_2. The fetal measurements must be converted to z-scores, which was done here using the INTERGROWTH-21st international fetal standards [24] previously published using the same data set. The correlation between any specified two time points must also be calculated. Fisher's transformed correlation of fetal measures from 14 to 40 weeks was modelled here using FPs. An equation for estimating the correlation between any two measurements within this GA range was presented. The calculation of growth velocity and changes in velocity expressed as z-scores requires two measurements, and therefore combines the imprecision of the two measurements [287, 300, 301]. Single estimates of velocity are not useful as they only identify fetuses or children whose growth is so extreme that their calculated centile is outside the cut-off point chosen. In growth monitoring, multiple measurements are better than just two measurements for identifying, measuring and plotting errors, and for recognising truly abnormal patterns.

Various operational definitions of 'abnormal growth velocity' have been proposed. A change of more than plus or minus one centile band (0.67 SD) has been proposed as abnormal [302, 303], as it represents the width of centile bands on growth charts (the distance between the 3^{rd}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 97^{th} centiles). It thus corresponds to the clinically salient event of an individual changing by one major

growth reference centile, e.g., from the 25^{th} centile to 10^{th} centile [304].

Ong *et al.* [247] also used a change in reference-based standard scores greater than +/- 0.67 as clinically significant in their study to identify predictors of postnatal catch-up growth from birth to 2 years and its relation to size and obesity at 5 years. They found that children who showed catch-up growth of 0.67 SD or more between 0 and 2 years were fatter and had more central fat distribution at 5 years than other children. Lampl *et al.* [248] considered a higher criterion of 0.73 SD to take into account inherent measurement error when investigating whether an episode of preterm labour compromised fetal growth.

The WHO report on standards for the velocity of child growth compared the velocity increment and velocity gain methods using longitudinal weight data from birth to 2 years. They concluded that recording growth in snapshots of 1- or 2-month intervals was affected by regression to the mean between birth and 1 or 2 months, but that regression to the mean was less evident at later ages [273]. They found a minor difference between the two methods of ±0.34 SD and therefore opted for the velocity increment approach for practical reasons that are also relevant for pregnancy care. They also provided sex-specific centiles for weight increments on birthweight in 1- and 2-week intervals from birth to 2 months, which is the period known to be most affected by regression to the mean [273].

In this chapter, a similar analysis to the WHO report was conducted, comparing the two methods but applying them to fetal data. Ninety percent of the differences between the velocity gain and velocity increments z-scores were between -0.93 and +0.71 SDs for 4-week correlations, between -0.76 and +0.87 SDs for 5-week correlations, and between -0.72 and +1.07 SDs for 6-week correlations with a start point at 14 weeks. The differences observed between the two approaches were two to three times greater using FHC data than those reported by the WHO for child weight. The differences of -0.33 and 0.34 reported by the MGRS corresponded to the 28^{th}

and 81^{st} centiles for 4-week intervals (including only 53% of the INTERGROWTH-21^{st} population), $21^{st}$and 76^{th} centiles for 5-week intervals (including only 55% of the INTERGROWTH-21^{st} population), and 22^{nd} and 68^{th} centiles for 6-week intervals (including only 46% of the INTERGROWTH-21^{st} population).

The main difference between the velocity increments and velocity gain approaches is that the latter adjusts for both GA and regression to the mean [242, 243, 292]. The results in this chapter have demonstrated that fetal data measured in 4-, 5-, and 6-week intervals do indeed demonstrate the well-known phenomenon of regression to the mean. This finding reiterates the need to account for regression to the mean in the calculation of velocity. Failure to so this may lead to bias, especially in fetuses with high or low starting velocities at any given GAs. The literature has suggested that differences of ± 0.67 SD are 'clinically significant' for risk profiling neonatal outcomes [247, 248, 304, 305]. This cut-off is based on statistical convenience and is not a direct measure of clinical relevance.

Despite their potential usefulness, velocity standards have not been widely used by clinicians because they have not been readily available. If they were available, uptake may be low as they often require computerisation, are challenging to use, and clinicians would have to learn how to use them. The uptake of velocity standards remains to be seen if they are clinically useful or how they would be used. Growth velocities pose other technical difficulties, for example, how frequently such measurements ought to be obtained. There are inherent trade-offs: the shorter the interval, the higher the variability in growth and in the measurement error compared with the actual growth. Extending the period loses the benefit of assessing velocity for informing clinical decisions. Frequent measurements also have implications for cost, staff numbers, and workload.

The WHO recommended using velocity increment standards in infants and children rather than velocity gain standards, as they are simple and are presented easily

using charts and tables. They also rightly argued that the velocity gain approach will require computerisation if it is to be clinically useful [273].

The main challenge in computing velocity gain is in the computation of a correlation matrix representing the whole time-space when measurements are taken. I have shown that this is achievable with careful modelling. Empirical correlations of fetal measurements were modelled and showed that fitted correlations offer a good fit to the empirical correlation structure. The correlations were modelled in weekly intervals on the assumption that the correlations were reasonably stable within each week of gestation. Linearity of the fitted correlation structure was assumed as correlation was seen to decrease with increasing GA. Estimates for unobserved measurement times in an interval could therefore be interpolated based on the model fit and assumed to decrease in a linear fashion. The closed form formula representing the correlation structure between 14 and 40 weeks enabled the calculation of velocity gain given any two time points in that age range. To aid the use of the velocity gain approach, an application will be developed and made freely available. It will have to be incorporated into ultrasound machine software to be clinically useful.

Correlation modelling is uncommon in the statistical literature [306]. Correlations are usually modelled when accounting for non-independence of observations, typically in a multi-level framework. Wade *et al.* [208] modelled a dataset exhibiting a strong correlation structure between the CD4 counts of the uninfected children of HIV-1 infected women using five correlation models: zero correlation; constant correlation; continuous time-dependent correlation; discrete time-dependent correlation; and both age- and time-dependent correlations. They explored the effect of incorporating correlations between measurements into estimates of age-related centiles and concluded that there was little effect on model choice, fitted centiles, or precision.

In another study, Wade *et al.* [**?**] re-analysed fetal abdominal circumference and bi-parietal diameter data previously published by Kurmanavicius *et al.* [179, 211].

This was a prospective study of pregnant women examined routinely three times during pregnancy and every 2–3 weeks for high-risk pregnancies. The original analysis used only the first of each series of measurements taken during pregnancy to create each chart. Wade *et al.* incorporated all of the measurements, which led to an 8-fold increase in the number of abdominal circumference and biparietal diameter measurements. The original centiles reported by Kurmanavicius *et al.* [179, 211] were very similar to those based on the re-analysis by Wade *et al.*, which incorporated a correlation structure between repeat measurements of the same individual [**?**]. However, incorporating the correlation structure was preferable to transforming the data to the cross-sectional form, as the latter could severely affect the precision and accuracy of the centiles.

Argyle [306] modelled the correlation for monitoring child growth data. She used the same data used by Wright (the Newcastle infancy data [242]) and Cole (the Cambridge infant study [243]) to define a normal rate of weight gain in infancy and develop reference charts to assess weight gain in British infants, respectively. Argyle modelled the correlation using a simple two-parameter Markov correlation model. As a consequence of the Markov property, conditioning was only based on the most recent previous measurement to assess the development of a child's growth profile. Inferences about the correlation parameters were derived from the likelihood methods based either on observed z-scores or, if raw data were unavailable, on an observed correlation matrix [306]. The Markov model was compared with a previously derived six-parameter correlation model by Cole [243]. Argyle concluded that the predictive performance of the two models was very similar.

8

General conclusions

Human growth has become an increasingly key area of interest over the last century. Establishing the expected attained size at a given age has become the norm for classifying 'normal' versus 'abnormal' growth. Therefore the statistical analysis of growth data has been of interest to many academics, from a wide range of disciplines, over the last century. In this thesis, I have focused on growth in the prenatal period, namely fetal and newborn growth. The thesis assessed approaches for attained size at a given age and the velocity gain (rate or speed of growth) of a fetus.

Methods for attained size have been discussed extensively in the literature. I have demonstrated their application using the INTERGROWTH-21st data to construct international standards for fetal and newborn growth. The methodology for calculating velocity gain is more complex and has never been tested with fetal data. In Chapter 7, I applied an approach based on the conditional velocity gain z-score using fetal data from the FGLS to assess a fetus's current HC z-score given their previous HC z-scores.

8.1 General introduction

In Chapter 1, I presented the results of three systematic reviews conducted by the INTERGROWTH-21st team to identify gaps in the literature. These reviews showed that even basic concepts like study design are still problematic and not well addressed by most studies in this field. The reviews revealed deficiencies in the statistical methodologies used in some studies, even though data analysis issues have been addressed before.

The distinction between a reference and a standard chart is important as it guides how the charts should be used and interpreted. These two terms have been used interchangeably even though the two chart types are based on different population constructs. This is a key concept that requires clarity for researchers when designing similar studies in future. A sample population used to produce a reference chart should consist of women at low risk of developing pathologies in pregnancy. It is therefore reasonable to choose a reference equation for use in a clinical service from a publication with the lowest risk of methodological bias.

The heterogeneity observed in the three systematic reviews was mainly due to considerable methodological heterogeneity between the studies in terms of sample selection, methodology, and statistical modelling methods. Very few studies used a quality control process or standardisation of the sonographers and anthropometrics who were involved in obtaining measurements. The use of standardised instruments, measurement protocols, and trained staff were not common.

I highlighted the potential pitfalls in reporting studies that aim to create charts. I proposed a checklist for evaluating the methodological quality of studies that provides a rough guide of the minimum information that should be reported in these studies. This checklist is not intended to commend or discard studies but rather to act as a consensus guideline to improve consistency and as a guide for

evaluating similar studies for future research in human growth studies.

8.2 Design and methodological considerations for the construction of human fetal and neonatal size and growth charts

The numerous systematic reviews conducted by the INTERGROWTH-21st team revealed that many studies aiming to construct charts are still poorly conducted. This motivated Chapter 2, in which I discussed some of the key issues in study design, statistical methods, and reporting that should be considered when designing such studies and builts upon similar work by Altman and Chitty [67], and Royston and Altman [93].

The aim of this chapter was to stress some of the concepts that have already been addressed in the literature and to address concepts that have not been discussed before, such as the importance of taking repeat measurements and blinding sonographers when taking ultrasound measurements. The issues were largely informed by the experience of conducting the large prospective INTERGROWTH-21st study. There have been improvements over time in the quality of studies, shown by positive correlations between quality scores and year of publication. The methodological quality of these studies is therefore steadily improving, thanks to the earlier efforts by Altman and Chitty, and Royston and Altman.

There is as yet no guidance on sample size calculations for studies aiming to construct references centiles. More work is needed on calculating the sample size required for fetal growth studies. In particular, aspects of longitudinal studies have not yet been considered, such as the effect of correlations between measurements of the same individual at different ages, the number of replicates per measurement, the timing of measurements (in general more measurements are needed in periods of more rapid

growth to accurately capture the pattern of growth), and the number of observations per individual. The majority of the studies considered in the INTERGROWTH-21st reviews did not provide a statement on how their sample size calculations were done. This is crucial as precision of centiles is dependent on an adequate sample size, especially for the outer centiles that are the target for clinical decisions.

8.3 Assessing the combinability of linear growth data

It was important to check whether the data from the eight sites in the INTERGROWTH-21st project were similar enough to be combined before using them to construct a global standard. This challenging problem was addressed in Chapter 3. It is relatively easy to assess how comparable data from multiple sites are when only considering the homogeneity of the populations across the different sites. However, the INTERGROWTH-21st project standardised the populations across its multiple sites. The combinability of the data could only be assessed by looking at the actual measurements, which were continuous measurements that varied by GA. There was no standard methodology available to determine the combinability of such data. There was also no consensus on how best to compare sets of centiles across a range of GAs.

The challenge here was identifying the time points at which to determine the combinability of the data. For example, the data could be compared on the smallest unit of GA, which is a day. As the women were recruited at 9 weeks and term delivery is considered to be 40 weeks, there were 217 days over which the data could be compared. This introduced issues associated with multiple testing. The other challenge was judging what would be considered a significant difference at each time point.

I conducted an extensive literature search and could not find any such evaluations previously done on fetal data. The MGRS experienced a similar problem and used three methods to assess the similarity of child growth data collected from six sites. I successfully applied these three methods (variance component analysis, FP regression, and SSD analysis) and applied them successfully to the INTERGROWTH-21st fetal data. I also considered a fourth, complementary method, meta-analytic assessment using regression.

The results of the four somewhat-related approaches were consistent and confirmed that the data were largely similar enough to be pooled together for the construction of a unified standard. The results were in remarkable agreement with those of the WHO-MGRS, and suggested that the differences in fetal growth and newborn size reported in the literature are more likely due to environmental and socioeconomic differences than genetic variation, as has been demonstrated for infants and children.

8.4 Statistical methodology for cross-sectional studies of human growth: Using the INTERGROWTH-21st Project as a case study

The choice of an appropriate statistical methodology for creating fetal standards was important as inaccurate centiles resulting from inferior methods can lead to mis-judgements about fetal size development and result in sub-optimal clinical care [34]. Choosing the best model from among many is not trivial, especially when dealing with large datasets such as the INTERGROWTH-21st data (N = 20,302). Normally, significance testing and goodness-of-fit statistics like the likelihood ratio test or the AIC are used to discriminate between models. However, these methods tend not to be useful when examining large datasets, as very small differences will be statistically significant even if they are indistinguishable on actual centile plots.

The desire for centiles that are both smooth and precise is a trade-off between the statistically best model and factors such as the aesthetic appeal and complexity of the model. The analysis of the INTERGROWTH-21st Project benefited from past experiences of similar work conducted by the senior statistician on the project (Professor Douglas G. Altman). Statistical input was available from the design stage of the project. This invaluable input helped to ensure good data quality.

The focus of this thesis was the analysis of a single biometric measurement, such as FHC. It is intuitive that biometric measurements are correlated. As several fetal biometry measurements were collected at each visit, a multivariable model that considers all of the measurements collected at each visit may give a better understanding of fetal attained size and growth patterns than a model that considers just one measurement. The added value of considering all of the collected measurements in a multivariable analysis has not yet been evaluated.

8.5 Statistical methodology for longitudinal studies of human growth: Using the INTERGROWTH-21st Project as a case study

Longitudinal studies pose an extra challenge over cross-sectional studies as they require more complex statistical models that take into account the correlation structure of repeat observations. Ignoring multiple measurements per subject is likely to lead to overestimated centile precision, resulting in narrower centiles [**?**]. Many methods exist that can take into account the non-independence of observations. Despite the availability of such methods, the INTERGROWTH-21st reviews revealed that some studies with repeated measures do not use these methods. Instead, they simply select one observation at random, discarding the rest. Discarding data cannot be justified, considering the effort and resources required to obtain it. This

is data wastage and should be highly discouraged.

The analysis of the INTERGROWTH-21st repeat measurements data in Chapter 5 revealed that there was hardly any difference in centile estimates when data dependency was or was not accounted for. This was somewhat expected given the homogeneity of the women included in the FGLS component of the INTERGROWTH-21st study, which resulted in a balanced dataset. Measurements were taken following a unified protocol, 95% of the subjects were measured four to six times, and the frequency of the measurements was independent of previous measurements [**?**]. The homogeneity of this cohort and the small variation in the number of repeat measurements per woman explain the minimal differences observed when using different modelling approaches.

It is important to fully account for repeat measurements as important differences may emerge in other datasets that are not similar to the INTERGROWTH-21st data. It is therefore important to determine the main aim of a study from the outset and choose an appropriate study design based on that aim.

8.6 Estimating gestational age from crown-rump length in early pregnancy when gestational age range is truncated

I encountered an unusual problem when modelling the CRL data, as the outcome variable, GA, was truncated at both ends, at 9 and 14 weeks. This feature of the data had the potential to introduce considerable bias, mostly at the extremes of GA, unless analysed carefully. This problem is not commonly encountered in most analyses of medical data. Altman *et al.* [78] addressed a similar problem when estimating GA using FHC by restricting the range of measurements included in the regression analyses. However, their FHC data had a substantial GA range

of 12—42 weeks. In contrast, the INTERGROWTH-21st CRL data spanned only 5 weeks. Using only the GA data unaffected by truncation would have led to a large loss of data and limited clinical usefulness.

In Chapter 6, I created three *ad hoc* approaches to circumvent the problem of truncation. Although these approaches did not follow standard statistical analysis paradigms for modelling, I showed empirically that they dealt sufficiently with the problem of data truncation. I was therefore able to avoid or reduce bias in the modelled estimates as a consequence of data truncation.

Although only examined for CRL, these methods may be able to solve other truncation problems involving similar data. Their applicability to other settings needs to be evaluated. The choice of which approach is best is hard to justify through formal statistical testing, and is likely to depend on the specific data being analysed.

8.7 Fetal growth velocity standards

The work presented on fetal velocity in Chapter 7 is novel as there are currently no fetal velocity standards in existence. Most of the work that has been done on velocity has focused on postnatal growth, i.e., child growth velocity. A few studies have looked at fetal velocity, but not in detail. This is mainly due to lack of appropriate data and perhaps the complex statistical methodology required.

All that is needed to calculate a conditional fetal biometry gain z-score is a growth standard to convert fetal measurements to z-scores and the correlation structure of the fetal measurement z-scores during pregnancy. I modelled the correlation structure of the FHC, abdominal circumference, and femur length z-scores, allowing the correlation for any pair of GAs during pregnancy to be calculated. The two approaches that I used had not been tested with fetal data before.

The correlation matrix of expected fetal velocities for FHC, abdominal circumference,

and femur length will be useful for determining the expected velocity for a particular fetus at any two time points conditional on previous measurements.

8.8 Future work

The approaches used to address the truncation problem in Chapter 6 did not follow the conventional statistical approaches for analysing such data. Although I have demonstrated that the proposed methods were fit for purpose in this scenario, more work is needed to formalise these ideas based on statistical theory. Ideally, appropriate flexible, nonlinear, double-truncated distributions that represent the truncated dataset would need to be identified first. This is important for making statistical inferences on the identified distribution to estimate mean, SD, and any desired centiles.

The work on fetal velocity in Chapter 7 requires extension to better understand how these velocities relate to newborn outcomes. I calculated the expected velocities for a fetus from 14 to 40 weeks. This information can be used to define a fetus that is growth faltered and relate its growth to outcomes. Modelling individual growth trajectories for longitudinal data is another area of interest, as it would highlight the spurts, troughs, and levelling off that are characteristic of the growth process. Such an approach would enable important patterns and sources of variation to be explored. The work on pooling data from multiple sites presented in Chapter 3 requires extension on what clinically meaningful differences should prevent data from being combined. Pertinent questions, such as how to determine an acceptable difference, remain unanswered. Applying the same methodology to datasets that have not been carefully selected and in more heterogeneous populations may be useful in understanding and assessing the differences we expect in the general population and whether these differences can be predictive of newborn outcomes.

Appendices

A

Methodological criteria used to score studies that created pregnancy dating charts.

	Item	Low risk of bias	High risk of bias
		1. STUDY DESIGN	
1.1	Recruitment period	Reported in months	Not reported
1.2	Prospective data collection	Prospective study and ultrasound data were collected specifically for the purpose of constructing charts for dating of pregnancy	Retrospective study or data not collected specifically for the purpose of constructing charts (e.g., use of routinely collected data)
1.3	Population	Women were reported as coming from an unselected population or from a population at low risk of pregnancy complications	Women did not come from an unselected population, or were selected, or were at high risk of pregnancy complications, or the population was not reported
1.4	Spontaneous conception	Pregnancies following spontaneous conception	Pregnancies conceived by assisted reproductive technology
1.5	Sample selection	Women were selected either consecutively or at random	Convenience sampling, arbitrary recruitment, or not reported
1.6	Sample size	A priori determination or calculation of sample size and justification	Lack of a priori sample size determination or calculation and justification

1.7	Design	Clearly either a cross-sectional or longitudinal design	Not reported, or a mixture of cross-sectional and longitudinal data
1.8	Method of selecting the gestational ages at which the fetuses were measured **(only for longitudinal studies)**	Interval of measures prospectively pre-specified and justified	Interval of measures not prospectively pre-specified and justified or not reported
1.9	Number of occasions each fetus was measured **(only for cross-sectional studies)**	Each fetus was measured and included only once	Some fetuses were measured and included more than once
1.10	Exclusion criteria	The study made it clear that women at high risk of pregnancy complications were not included and that women with abnormal outcomes were excluded, i.e. an effort was made to include only "normal" outcomes as best possible	The study population included both low-risk and high-risk pregnancies, or women with abnormal outcomes were not excluded

1.11	Method of dating pregnancy	Clearly described, by LMP	Not by LMP, or not described clearly, or not reported.
1.12	Certainty of the LMP assessed	All of the following criteria were reported: LMP certain; Regular menstrual cycles prior to pregnancy; No recent use of contraceptive (1 month or more); No recent breastfeeding (1 month or more); No recent pregnancy (1 month or more)	Any of the criteria were not assessed
2. STATISTICAL METHODS			
2.1.	Characteristics of the study population	Presented in a table or clearly described and included a minimum dataset of age, weight and height (or BMI), and parity	Not presented in a table, or not clearly described, or does not contain the minimum data set
2.2	GA range	Reported	Not reported
2.3	Ultrasound machine(s) used	Clearly specified	Not clearly specified
2.4	Probe type (transvaginal or transabdominal)	Reported	Not reported

2.5	Ultrasound machine type (static or real-time)	Reported	Not reported
2.6	Description of measurement techniques	The study described sufficient and unambiguous details of the measurement techniques used for fetal CRL	The study did not describe sufficient and unambiguous details of the measurement techniques used
2.7	Number of sonographers that took the measurements	Reported	Not reported
2.8	Contains quality control measures	Should include the following: assessment of intra-observer variability; assessment of inter-observer variability; image review; image storage	Does not contain quality control measures
3. REPORTING METHODS			
3.1	Report of mean and SD of each measurement and the sample size for each week of gestation.	Presented in a table or clearly described	Not presented in a table or not clearly described

3.2	Report of regression equations for the mean (and SD if relevant) for each measurement	Reported	Not reported
3.3	Number of CRL measurements taken	More than one measure per fetus per scan	Single measure or not specified
3.4	Statistical methods	Clearly described and applied	Not clearly described and applied
3.5	Assessment of increasing variability of the data with gestation	Performed	Not performed
3.6	Assessment of the goodness of fit of the models	A test of the goodness-of-fit of the models was reported	The goodness-of-fit of the models was not reported
3.7	Scatter diagram of the data with the fitted median/mean superimposed	Study included scatter diagrams of the data with the median/mean superimposed	Study did not include scatter diagrams of the data with the median/mean superimposed

3.8	Change of the mean or median across GA	Smooth change	Not smooth change
3.9	Change of the SD or centile across GA	Smooth change	Not smooth change

CRL, crown-rump length; LMP, last menstrual period.

Note: For each criterion, one point was given for low risk of bias and zero for high risk of bias.

B

Methodological criteria used to score the studies that created charts of fetal size.

	Item	Low risk of bias	High risk of bias
		1. STUDY DESIGN	
1.1.	Study design	Clearly described as either cross-sectional or longitudinal	Not reported, a mixture of cross-sectional and longitudinal data
1.2	Sample selection	Population-based study with attempts to identify and clearly define populations from a specific geographic area. From this underlying population women are selected either consecutively or at random.	Not population-based, convenience sampling, arbitrary recruitment, or not reported
1.3	Number of occasions each fetus was measured (only for cross-sectional design)	Each fetus was measured and included only once	Some fetuses were measured and included more than once.
1.4.	Method of selecting the gestational ages at which the fetuses were measured (only for longitudinal studies)	Interval of measures prospectively prespecified and justified	Interval of measures not prospectively prespecified and justified or not reported

1.5.	Reason(s) for choosing a particular number of serial measurements (only for longitudinal studies)	Clear documentation of the intended number of serial measurements	No clear documentation of the intended number of serial measurements
1.6.	Inclusion and exclusion criteria	The study made it clear that women at high risk of pregnancy complications were not included and that women with abnormal outcome were excluded, i.e., an effort was made to include normal outcomes as best as possible	The study population included both low-risk and high-risk pregnancies or women with abnormal outcomes were not excluded
1.7.	Sample size	Given the lack of proper guidelines for sample size calculation when constructing references, we consider that any a priori determination / calculation of sample size and justification is acceptable	No a priori sample size determination/calculation and justification

1.8.	Data collection	Prospective and data collected specifically for the purpose of constructing charts of fetal size or fetal growth	Retrospective, or data not collected specifically for the purpose of constructing charts of fetal size or fetal growth, or unclear (e.g. use of routinely collected data)
1.9.	Method of dating pregnancy	Clearly described. Known LMP and regular menstrual cycles prior to pregnancy AND a sonogram before 14 weeks demonstrating a crown-rump length (CRL) that **corroborates the LMP date** (within how many days unspecified). The assessment of GA should be performed < 14 weeks and include measurement of fetal CRL	Not described clearly. GA assessment at >14 weeks, or GA assessment not including ultrasound verification.
1.10.	Collection of data on gestational age at inclusion	The gestational age was calculated precisely to the day	Truncation of the gestational age to the number of completed weeks

2. STATISTICAL METHODS			
2.1.	Number of measurements taken for each biometric variable	More than one measure per fetus per scan	Single measure or not specified
2.2.	Statistical methods	Clearly described and identified	Not clearly described and identified
2.3.	Assessment of increasing variability of the data with gestation	Performed	Not performed
2.4.	Assessment of the goodness of fit of the models	A test of the goodness-of-fit of the models was reported	The goodness-of-fit of the models was not reported
2.5.	Scatter diagram of the data with the fitted percentiles superimposed	Study included scatter diagrams of the data with the percentiles superimposed	Study did not include scatter diagrams of the data with the percentiles superimposed
2.6.	Change of reference centiles across GA	Smooth change	Not smooth change

2.7.	Methods used to estimate age-specific reference intervals for fetal size measurements	Mean and SD model, smoothed crude centiles, or LMS method	Inadequate methodology
3. REPORTING METHODS			
3.1.	Characteristics of study population	Presented in a table or clearly described, and includes the minimum dataset of age, weight and height or BMI, and parity	Not presented in a table, or not clearly described, or does not contain the minimum data set
3.2	Description of number approached / enrolled	Described	Not described
3.3.	Ultrasound machine (s) used	Clearly specified	Not clearly specified
3.4.	Number of sonographers that took the measurements	Reported	Unreported
3.5.	Description of measurement techniques	The study described sufficient and unambiguous details of the measurement techniques used for fetal size parameters	The study did not described sufficient and unambiguous details of the measurement techniques used for fetal size parameters

3.6.	Contains quality control measures	including: assessment of intra-observer variability; assessment of inter-observer variability; image review; image scoring; image storage	Does not contain quality control measures
3.7.	Report of mean and SD of each measurement and the sample size for each week of gestation.	Presented in a table or clearly described	Not presented in a table or not clearly described
3.8.	Report of regression equations for the mean (and SD if relevant) for each measurement	Reported	Unreported

Note: For each criterion, one point was given for low risk of bias and zero for high risk of bias.

C

Methodological criteria used to score studies that created charts of neonatal size.

	Item	Low risk of bias	High risk of bias
1. STUDY DESIGN			
1.1	Aim of the study	To create neonatal size charts (for weight, length, head circumference, or a combination of weight and length)	Not to create neonatal size charts, to assess the postnatal growth of premature babies, or aim not clearly defined
1.2	Definition of target population	Target population was clearly defined (e.g. geographical area, ethnic group, single or multiple pregnancy, among others)	Not clearly defined

1.3	Distinction between reference and standard	The authors clearly stated whether the chart was a reference or a standard, or this information could be discerned from the methods section ***Reference*** *refers to anthropometry of a given population at a particular time and place, such as a hospital, region, or country. It includes an unselected group of women with minimal exclusion criteria regarding risk factors for optimal health.* ***Standard*** *refers to anthropometry of a population considered to be of optimal health, with good education, socioeconomic status, adequate nutritional status, and at low risk of abnormal growth. It shows how humans should grow independent of time and place.*	Not clearly stated or could not be discerned

1.4	Sample selection	The sample was part of the target population and the inclusion/exclusion criteria were clearly reported. Data were collected prospectively enrolling the neonates consecutively	Inclusion/exclusion criteria were not clearly reported. The study was not planned or the data were not collected prospectively (e.g. data set collected for different study or for a hospital or national registry)
1.5.	Gestational age evaluation	Gestational age (GA) was determined by both last menstrual period (LMP) and ultrasound assessment (US). Reliability of the GA was based on the difference between the US and LMP assessment (less than 2 weeks). Data with unreliable GA were excluded	The assessment of GA was not reported or unclear. Only LMP or US GA evaluation. No reliability of GA, or methods different from that described. Neonates with uncertain GA were not excluded

1.6.	Anthropometric evaluation	Anthropometric traits were measured using standardised instruments. Instruments were calibrated at least fortnightly. Anthropometric traits were measured using standardised protocols. Time at which the measures were taken after birth was reported. Measures were taken by at least two different operators. Operators were trained. Operators were standardised	Standardised instruments or protocols were not used or reported. The instruments were not periodically calibrated. The measures were taken by only one operator. The operators were not trained/standardised. The time when the measures were taken after birth was not reported or it was too long
1.7.	Number of neonates at each GA	Reported and adequate sample size for creating reliable charts	Not reported or small sample size (non-adequate for creating reliable charts)

2. STATISTICAL METHODS			
2.1.	Assessment of outliers	Method used to detect the outliers was appropriate, well described, and justified. Outlier values were corrected (if it was possible) or excluded, and number of outliers detected was reported	The method used was not appropriate (e.g., low or high centiles), not reported, or not justified. Outliers were not evaluated, or not corrected, or excluded
2.2.	Gender as covariate	The charts were presented by gender, except for weight-for-length, body mass index and ponderal index (which can be produced without covariates)	No gender specific charts were produced (except for weight-for-length, body mass index and ponderal index)
2.3.	Statistical models	The model used to create the charts was clearly described and identified. There was agreement between the model and the underlying distribution of anthropometric measures	The statistical method was not reported, unclear, or not appropriate
2.4.	Assessment of goodness of fit	Method used for assessment of goodness of fit was reported	Not reported or not done

2.5.	Lack of precision of the estimates	Standard errors (or confidence limits) of extreme centiles were reported and were small	Not done, not reported, or non-small standard errors
2.6.	Smooth centiles	Smoothed centiles	Raw centiles
3. REPORTING OF RESULTS			
3.1.	Characteristics of the study population	Baseline characteristics (pregnancy and neonatal morbidity and mortality) were presented in tables or in text	Baseline characteristics not presented in tables or described in the text
3.2.	Gestational age expression	It was clearly stated whether GA was expressed as completed weeks, at the nearest week, or weeks + days	Not stated or unclear
3.3.	Chart presentation	Values of (at least) 10^{th}, 50^{th}, and 90^{th} centiles, or parameters that allow them to be computed were reported. Z-scores were directly presented or charts that allowed them to be computed were presented	Values of 10^{th}, 50^{th} or 90^{th} centiles were not reported or not computable. Z-scores were not presented or not computable

Note: For each criterion, one point was given for low risk of bias and zero for high risk of bias.

Bibliography

[1] Villar J, Smeriglio V, Martorell R, Brown CH, Klein RE. Heterogeneous growth and mental development of intrauterine growth-retarded infants during the first 3 years of life. Pediatrics. 1984;74(5):783–791.

[2] Belizan JM, Villar J, Nardin JC, Malamud J, De Vicurna L. Diagnosis of intrauterine growth retardation by a simple clinical method: measurement of uterine height. Am J Obstet Gynecol. 1978;131(6):643–6.

[3] Villar J, Belizan JM. The Tinning factor in the pathophysiology of the intrauterine growth retardation syndrome. Obstetrical Gynecological Survey. 1982;37(8):499–506.

[4] Villar J, Belizan JM, Spalding J, Klein RE. Postnatal growth of intrauterine growth retarded infants. Early Human Development. 1982;6(3):265–271.

[5] Belizan JM, Villar J, Bergel E, del Pino A, Di Fulvio S, Galliano SV, et al. Long-term effect of calcium supplementation during pregnancy on the blood pressure of offspring: follow up of a randomised controlled trial. BMJ. 1997;315(7103):281–5.

[6] Villar J, Belizan JM. The relative contribution of prematurity and fetal growth retardation to low birth weight in developing and developed societies. Am J Obstet Gynecol. 1982;143(7):793–8.

[7] Barker DJ, Osmond C. Infant mortality, childhood nutrition and ischaemic heart disease in England and Wales. Lancet. 1986;1(8489):1077–81.

[8] Barker DJ, Winter PD, Osmond C, Margetts B, Simmonds SJ. Weight in infancy and death from ischaemic heart disease. Lancet. 1989;2(8663):577–80.

[9] Barker DJ, Gluckman PD, Godfrey KM, Harding JE, Owens JA, Robinson JS. Fetal nutrition and cardiovascular disease in adult life. Lancet. 1993;341(8850):938–41.

[10] Barker DJ. The origins of the developmental origins theory. J Intern Med. 2007;261(5):412–7.

[11] Villar J, Altman DG, Purwar M, et al. The objectives, design and implementation of the INTERGROWTH-21st Project. BJOG. 2013;120(Suppl 2):9–26.

[12] Villar J, Papageorghiou AT, Pang R, et al. The likeness of fetal growth and newborn size across non-isolated populations in the INTERGROWTH-21st Project: the Fetal Growth Longitudinal Study and Newborn Cross-Sectional Study. The Lancet Diabetes and Endocrinology. 2014;2(10):781–92.

[13] Raman S, Teoh T, Nagaraj S. Growth patterns of the humeral and femur length in a multiethnic population. Int J Gynaecol Obstet. 1996;54(2):143–7.

[14] Schwarzler P, Bland JM, Holden D, Campbell S, Ville Y. Sex-specific antenatal reference growth charts for uncomplicated singleton pregnancies at 15-40 weeks of gestation. Ultrasound Obstet Gynecol. 2004;23(1):23–9.

[15] Krampl E, Lees C, Bland JM, Espinoza Dorado J, Moscoso G, Campbell S. Fetal biometry at 4300 m compared to sea level in Peru. Ultrasound Obstet Gynecol. 2000;16(1):9–18.

[16] Giuliani F, Ohuma E, Spada E, Bertino E, Al Dhaheri AS, Altman DG, et al. Systematic review of the methodological quality of studies designed to create neonatal anthropometric charts. Acta Paediatrica. 2015;104(10):987–996.

[17] Bertino E, Milani S, Fabris C, De-Curtis M. Neonatal anthropometric charts: what they are, what they are not. Archives of Disease in Childhood Fetal and Neonatal Edition. 2007;92(1):F7–F10.

[18] de Onis M, Garza C, Onyango AW, Martorell R. WHO child growth standards. Acta Paediatr. 2006;450:1–101.

[19] de Onis M. Assessment of differences in linear growth among populations in the WHO Multicentre Growth Reference Study. Acta Paediatrica. 2006;95(S450):56–65.

[20] Villar J, Altman D, Purwar M, Noble J, Knight H, Ruyan P, et al. The objectives, design and implementation of the INTERGROWTH-21st Project. BJOG. 2013;120(s2):9–26.

[21] Papageorghiou AT, Kennedy SH, Salomon LJ, et al. International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown-rump length in the first trimester of pregnancy. Ultrasound in Obstetrics and Gynecology. 2014;44(6):641–8.

[22] Altman DG, Ohuma EO. Statistical considerations for the development of prescriptive fetal and newborn growth standards in the INTERGROWTH-21st Project. BJOG. 2013;120(s2):71–76.

[23] Ohuma EO, Hoch L, Cosgrove C, Knight HE, Cheikh Ismail L, Juodvirsiene L, et al. Managing data for the international, multicentre INTERGROWTH-21st Project. BJOG. 2013;120(s2):64–70.

[24] Papageorghiou AT, Ohuma EO, Altman DG. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21^s Project. Lancet. 2014;384(9946):869–79.

[25] Villar J, Ismail LC, Victora CG, et al. International standards for newborn weight, length and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. The Lancet. 2014;384(9946):857–68.

[26] Villar J, Giuliani F, Bhutta ZA, Bertino E, Ohuma EO, Ismail LC, et al. Postnatal growth standards for preterm infants: the Preterm Postnatal Follow-up Study of the INTERGROWTH-21st Project. The Lancet Global Health. 2015;3(11):e681–e691.

[27] Napolitano R, Dhami J, Ohuma E, et al. Pregnancy dating by fetal crown-rump length: a systematic review of charts. BJOG. 2014;121(5):556–565.

[28] Ioannou C, Talbot K, Ohuma E, et al. Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size. BJOG. 2012;119(12):1425–1439.

[29] Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283(15):2008–12.

[30] Tanner J. Foetus into Man: Physical Growth from Conception to Maturity. 2nd ed. Cambridge, MA: Ware: Castlemead Publications; 1989.

[31] The Institute of Obstetricians and Gynaecologists; Royal College of Physicians of Ireland and Directorate of Quality and Clinical Care. Ultrasound diagnosis of early pregnancy miscarriage. Clinical Practice guideline. 2010;Version 1(Guideline No. 1):4–7.

[32] Campbell S, et al. Routine ultrasound screening for the prediction of gestational age. Obstetrics and Gynecology. 1985;65(5):613–620.

[33] Waldenstrom U, Axelsson O, Nilsson S. A comparison of the ability of a sonographically measured biparietal diameter and the last menstrual period to predict the spontaneous onset of labor. Obstetrics and Gynecology. 1990;76(3 Pt 1):336–338.

[34] Treloar A, Behn BG, Cowan DW. Analysis of gestational interval. American journal of Obstetrics and Gynecology. 1967;99(1):34–45.

[35] ISUOG Practice Guidelines: Performance of first-trimester fetal ultrasound scan. Ultrasound in Obstetrics and Gynecology. 2013;41(1):102–113.

[36] Verburg BO, Steegers EAP, De-Ridder M, et al. New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study. Ultrasound in Obstetrics and Gynecology. 2008;31(4):388–96.

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

[38] Blondel B, Morin I, Platt RW, Kramer MS, Usher R, Breart G. Algorithms for combining menstrual and ultrasound estimates of gestational age: consequences for rates of preterm and postterm birth. BJOG. 2002;109(6):718–20.

[39] Taipale P, Hiilesmaa V. Predicting delivery date by ultrasound and last menstrual period in early gestation. Obstetrics and Gynecology. 2001;97(2):189–194.

[40] Kalish RB, et al. First- and second-trimester ultrasound assessment of gestational age. American journal of Obstetrics and Gynecology. 2004;191(3):975–978.

[41] Caughey AB, Nicholson JM, Washington AE. First- vs second-trimester ultrasound: the effect on pregnancy dating and perinatal outcomes. American journal of Obstetrics and Gynecology. 2008;198(6):703.e1–703.e6.

[42] Bennett KA, et al. First trimester ultrasound screening is effective in reducing postterm labor induction rates: A randomized controlled trial. American

journal of Obstetrics and Gynecology. 2004;190(4):1077–1081.

[43] Robinson HP, Fleming JEE. A critical evaluation of sonar "crown-rump length" measurements. BJOG. 1975;82(9):702–10.

[44] Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller AB, Narwal R, 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. The Lancet. 2012;379(9832):2162–2172.

[45] de Onis M, Blossner M, Villar J. Levels and patterns of intrauterine growth retardation in developing countries. Eur J Clin Nutr. 1998;52(Suppl 1):S5–15.

[46] McLennan AC, Schluter PJ. Construction of modern Australian first trimester ultrasound dating and growth charts. journal of Medical Imaging and Radiation Oncology. 2008;52(5):471–9.

[47] Sahota DS, Leung TY, Leung TN, Chan OK, Lau TK. Fetal crown-rump length and estimation of gestational age in an ethnic Chinese population. Ultrasound in Obstetrics and Gynecology. 2009;32(2):157–60.

[48] Pedersen JF. Fetal crown-rump length measurement by ultrasound in normal pregnancy. Br J Obstet Gynaecol. 1982;89(11):926–30.

[49] Piantelli G, Sacchini C, Coltri A, Ludovici G, Paita Y, Gramellini D. Ultrasound dating-curve analysis in the assessment of gestational age. Clin Exp Obstet Gynecol. 1994;21(2):108–18.

[50] Kurjak A, Cecuk S, Breyer B. Prediction of maturity in first trimester of pregnancy by ultrasonic measurement of fetal crown-rump length. J Clin Ultrasound. 1976;4(2):83–4.

[51] Bovicelli L, Orsini LF, Rizzo N, Calderoni P, Pazzaglia FL, Michelacci L. Estimation of gestational age during the first trimester by real-time

measurement of fetal crown-rump length and biparietal diameter. J Clin Ultrasound. 1981;9(2):71–5.

[52] Selbing A. Gestational age and ultrasonic measurement of gestational sac, crown-rump length and biparietal diameter during first 15 Weeks of pregnancy. Acta Obstetricia et Gynecologica Scandinavica. 1982;61(3):233–235.

[53] Rossavik IK, Torjusen GO, Gibbons WE. Conceptual age and ultrasound measurements of gestational sac and crown-rump length in in vitro fertilization pregnancies. Fertil Steril. 1988;49(6):1012–7.

[54] van de Velde EHE, Broeders GHB, Horbach JGM, Esser-Rath MVWCJ. Estimation of pregnancy duration by means of ultrasonic measurements of the fetal crownrump length. European Journal of Obstetrics Gynecology and Reproductive Biology. 1980;10(4):225–230.

[55] Conde-Agudelo A, Papageorghiou AT, Kennedy SH, Villar J. Novel biomarkers for predicting intrauterine growth restriction: a systematic review and meta-analysis. BJOG. 2013;120(6):681–94.

[56] Campbell S. An improved method of fetal cephalometry by ultrasound. BJOG. 1968;75(5):568–576.

[57] Bricker L, Neilson JP, Dowswell T. Routine ultrasound in late pregnancy (after 24 weeks' gestation). Cochrane Database Syst Rev. 2008;4(4):Cd001451.

[58] Sylvan K, Ryding EL, Rydhstroem H. Routine ultrasound screening in the third trimester: a population-based study. Acta Obstet Gynecol Scand. 2005;84(12):1154–8.

[59] Lubchenco LO, et al. Intrauterine growth as estimated from liveborn birth-weight data at 24 to 42 weeks of gestation. Pediatrics. 1963;32(5):793–800.

[60] Battaglia FC, Lubchenco LO. A practical classification of newborn infants by weight and gestational age. The journal of Pediatrics. 1967;71(2):159–163.

[61] Platt RW, Abrahamowicz M, Kramer MS, Joseph KS, Mery L, Blondel B, et al. Detecting and eliminating erroneous gestational ages: a normal mixture model. Stat Med. 2001;20(23):3491–503.

[62] Tinggaard J, Aksglaede L, Sorensen K, Mouritsen A, Wohlfahrt-Veje C, Hagen CP, et al. The 2014 Danish references from birth to 20 years for height, weight and body mass index. Acta Paediatr. 2014;103(2):214–24.

[63] Michaelsen KF. Are the new Danish 2014 growth references really more appropriate than the World Health Organization standards? Acta Paediatrica. 2014;103(5):464–465.

[64] Bertino E, Spada E, Occhi L, Coscia A, Giuliani F, Gagliardi L, et al. Neonatal anthropometric charts: the Italian neonatal study compared with other European studies. J Pediatr Gastroenterol Nutr. 2010;51(3):353–61.

[65] Gardosi J. Dating of pregnancy: time to forget the last menstrual period. Ultrasound in Obstetrics and Gynecology. 1997;9(6):367–368.

[66] Mongelli M, Wilcox M, Gardosi J. Estimating the date of confinement: ultrasonographic biometry versus certain menstrual dates. Am J Obstet Gynecol. 1996;174(1 Pt 1):278–81.

[67] Altman DG, Chitty LS. Design and analysis of studies to derive charts of fetal size. Ultrasound in Obstetrics and Gynecology. 1993;3(6):378–384.

[68] DiFranza JR, Lew RA. Effect of maternal cigarette smoking on pregnancy complications and sudden infant death syndrome. J Fam Pract. 1995;40(4):385–94.

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

[70] Hendrix N, Berghella V. Non-placental causes of intrauterine growth restriction. Semin Perinatol. 2008;32(3):161–5.

[71] Zeitlin J, Ancel PY, Saurel-Cubizolles MJ, Papiernik E. The relationship between intrauterine growth restriction and preterm delivery: an empirical approach using data from a European case-control study. BJOG. 2000;107(6):750–8.

[72] Stacey T, Thompson JM, Mitchell EA, Ekeroma AJ, Zuccollo JM, McCowan LM. The Auckland Stillbirth study, a case-control study exploring modifiable risk factors for third trimester stillbirth: methods and rationale. Aust N Z J Obstet Gynaecol. 2011;51(1):3–8.

[73] Salomon LJ. Early fetal growth: concepts and pitfalls. Ultrasound in Obstetrics and Gynecology. 2010;35(4):385–9.

[74] Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343(6172):747–751.

[75] Witherspoon DJ, Wooding S, Rogers AR, Marchani EE, Watkins WS, Batzer MA, et al. Genetic similarities within and between human populations. Genetics. 2007;176(1):351–359.

[76] Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505(7481):43–49.

[77] WHO Physical status: the use and interpretation of anthropometry. WHO Expert Committee and others; WHO technical report series. 1995;854(121):55.

[78] Altman DG, Chitty LS. New charts for ultrasound dating of pregnancy. Ultrasound in Obstetrics and Gynecology. 1997;10(3):174–191.

[79] Bertino E, Di-Battista E, Bossi A, et al. Fetal growth velocity: kinetic, clinical and biological aspects. Archives of Disease in Childhood - Fetal and Neonatal Edition. 1996;74(1):F10–F5.

[80] Owen P, Donnet ML, Ogston SA, Christie AD, Howie PW, Patel NB. Standards for ultrasound fetal growth velocity. BJOG. 1996;103(1):60–9.

[81] Simonton DK. Francis Galton's Hereditary Genius: Its place in the history and psychology of science. American Psychological Association; 2003.

[82] Jensen AR. Galton's legacy to research on intelligence. J Biosoc Sci. 2002;34(2):145–72.

[83] Bowditch HP. The growth of children, studied by Galton's method of percentile grades. Boston; 1891.

[84] Evans T, Farrant P, Gowland M, McNay M, Richards B. Clinical applications of ultrasonic fetal measurements. In: The British Medical Ultrasound Society Fetal Measurements Working Party Report. British Institute of Radiology London; 1990. .

[85] Grummer-Strawn LM, Garza C, Johnson CL. Childhood Growth Charts. Pediatrics. 2002;109(1):141–142.

[86] Habicht JP, Martorell R, Yarbrough C, Malina RM, Klein RE. Height and weight standards for preschool children. How relevant are ethnic differences in growth potential? Lancet. 1974;1(7858):611–4.

[87] de Onis M, Onyango A, Borghi E, Siyam A, Blssner M, Lutter C. Worldwide implementation of the WHO Child Growth Standards. Public Health Nutrition. 2012;15(09):1603–1610.

[88] de Onis M, Onyango AW, Broeck JV, Chumlea WC, Martorell R. Measurement and standardization protocols for anthropometry used in the construction of a new international growth reference. Food Nutr Bull. 2004;25((1 Suppl)):S27–36.

[89] Deter RL, Harrist RB, Hadlock FP, Carpenter RJ. The use of ultrasound in the assessment of normal fetal growth: A review. journal of Clinical

Ultrasound 1981. 1981;9(9):481–93.

[90] Williams RL, Creasy RK, Cunningham GC, Hawes WE, Norris FD, Tashiro M. Fetal growth and perinatal viability in California. Obstetrics and Gynecology. 1982;59(5):624–32.

[91] Altman DG, Hytten FE. Intrauterine growth retardation: let's be clear about it. BJOG. 1989;96(10):1127–32.

[92] Altman DG, Chitty LS. Charts of fetal size: 1. Methodology. BJOG. 1994;101(1):29–34.

[93] Royston P, Altman DG. Design and analysis of longitudinal studies of fetal size. Ultrasound in Obstetrics and Gynecology. 1995;6(5):307–312.

[94] Royston P, Altman DG. Using fractional polynomials to model curved regression relationships. Stata Technical Bulletin. 1994;21(21).

[95] Bellera CA, Hanley JA. A method is presented to plan the required sample size when estimating regression-based reference limits. journal of Clinical Epidemiology. 2007;60(6):journal of Clinical Epidemiology.

[96] Royston P. Constructing time-specific reference ranges. Statistics in Medicine. 1991;10(5):675–90.

[97] Jennen-Steinmetz C. Sample size determination for studies designed to estimate covariate-dependent reference quantile curves. Stat Med. 2014;33(8):1336–48.

[98] Harris EK. Statistical bases of reference values in laboratory medicine. New York. 1995;.

[99] Linnet K. Nonparametric Estimation of Reference Intervals by Simple and Bootstrap-based Procedures. Clin Chem. 2000;46(6):867–9.

[100] Healy MJR, Rasbash J, Yang M. Distribution-free estimation of age-related centiles. Annals of Human Biology. 1988;15(1):17–22.

[101] Wright EM, Royston P. Calculating reference intervals for laboratory measurements. Statistical Methods in Medical Research. 1999;8(2):93–112.

[102] Linnet K. Two-stage transformation systems for normalization of reference distributions evaluated. Clin Chem. 1987;33(3):381–6.

[103] Royston P, Matthews JNS. Estimation of reference ranges from normal samples. Statistics in Medicine. 1991;10(5):691–5.

[104] Royston P. Calculation of unconditional and conditional reference intervals for foetal size and growth from longitudinal measurements. Statistics in Medicine. 1995;14(13):1417–36.

[105] Silverwood R, Cole TJ. Statistical methods for constructing gestational age related reference intervals and centile charts for fetal size. Ultrasound in Obstetrics and Gynecology. 2007;29(1):6–13.

[106] Healy MJ. Notes on the statistics of growth standards. Annals of Human Biology. 1974;1(1):41–6.

[107] Bland M. An introduction to medical statistics. OUP Oxford; 2015.

[108] Virtanen A, Kairisto V, Uusipaikka E. Regression-based reference limits: determination of sufficient sample size. Clin Chem. 1998;44(11):2353–8.

[109] Virtanen A, Kairisto V, Irjala K, Rajamaki A, Uusipaikka E. Regression-based reference limits and their reliability: example on hemoglobin during the first year of life. Clin Chem. 1998;44(2):327–35.

[110] Elveback LR, Taylor WF. Statistical methods of estimating percentiles. Annals of the New York Academy of Sciences. 1969;161(2):538–48.

[111] Altman DG. Construction of age-related reference centiles using absolute residuals. Statistics in Medicine. 1993;12(10):917–24.

[112] Tsang PKS, Larew JSA, Larew LA, Miyakawa TW, Hofer JD. Statistical approaches to determine analytical variability and specifications: application of experimental design and variance component analysis1. journal of Pharmaceutical and Biomedical Analysis. 1998;16(7):1125–41.

[113] Royston P. Estimating departure from normality. Stat Med. 1991;10(8):1283–93.

[114] Berwick DM. Continuous improvement as an ideal in health care. New England journal of Medicine. 1989;320(1):53–6.

[115] Williams SM, Parry BR, Schlup MM. Quality control: an application of the cusum. BMJ. 1992;304(6838):1359–61.

[116] Savulescu J. Beyond Bristol: taking responsibility. journal of Medical Ethics. 2002;28(5):281–2.

[117] Treasure T. Lessons from the Bristol case. BMJ. 1998;316(7146):1685–6.

[118] Wisheart JD, Dhasmana JP. The Bristol Affair: lessons to be learned. The Annals of Thoracic Surgery. 2001;71(4):1403–4.

[119] Salomon LJ, Bernard JP, Duyme M, B Doris aNM, Ville Y. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. Ultrasound in Obstetrics and Gynecology. 2006;27(1):34–40.

[120] Salomon LJ, Bernard JP, Ville Y. Analysis of Z-score distribution for the quality control of fetal ultrasound measurements at 20-24 weeks. Ultrasound in Obstetrics and Gynecology. 2005;26(7):750–4.

[121] Sarris I, Ioannou C, Dighe M, Mitidieri A, Oberto M, Qingqing W, et al. Standardization of fetal ultrasound biometry measurements: improving

the quality and consistency of measurements. Ultrasound Obstet Gynecol. 2011;38(6):681–7.

[122] Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research. 1999;8(2):135–60.

[123] Malina R, Hamill P, Lemeshow S. Selected measurements of children 6-11 years. Vital Health and Statistics Series. 1973;11.

[124] Kestin IG. A statistical approach to measuring the competence of anaesthetic trainees at practical procedures. Br J Anaesth. 1995;75(6):805–9.

[125] Runcie CJ. Assessing the performance of a consultant anaesthetist by control chart methodology. Anaesthesia. 2009;65(3):293–6.

[126] Kemp SV, Batrawy SHE, Harrison RN, et al. Learning curves for endobronchial ultrasound using cusum analysis. Thorax. 2010;65(6):534–8.

[127] Ioannou C, Hoch L, Ohuma EO, Altamn DG, Chamberlain P, Salomon L, et al. Standardisation and quality control of ultrasound measurements of the INTERGROWTH-21st Project. BJOG. 2013;120(s2):33–37.

[128] Sarris I, Ioannou C, Ohuma EO, et al. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. BJOG. 2013;120(Suppl 2):33–7.

[129] Biau DJ, Rigon MR, Petit GG, Nizard RS, Porcher R. Quality control of surgical and interventional procedures: a review of the CUSUM. Quality and Safety in Health Care. 2007;16(3):203–7.

[130] de Oliveira, Filho GR. The construction of learning curves for basic skills in anesthetic procedures: an application for the cumulative sum method. Anesthesia and Analgesia. 2002;95(2):411–6.

[131] Okrainec A, Ferri LE, Feldman LS, Fried GM. Defining the learning curve in laparoscopic paraesophageal hernia repair: a CUSUM analysis. Surgical Endoscopy. 2011;25(4):1083–7.

[132] Codman EA. The classic: A study in hospital efficiency: as demonstrated by the case report of first five years of private hospital. Clin Orthop Relat Res. 2013;471(6):1778–83.

[133] de Lusignan S, Metsemakers JF, Houwink P, Gunnarsdottir V, van der Lei J. Routinely collected general practice data: goldmines for research? A report of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCIWG) from MIE2006, Maastricht, The Netherlands. Inform Prim Care. 2006;14(3):203–9.

[134] Etheredge LM. A Rapid-Learning Health System. Health Affairs. 2007;26(2):w107–w118.

[135] Jansen AC, van Aalst-Cohen ES, Hutten BA, Buller HR, Kastelein JJ, Prins MH. Guidelines were developed for data collection from medical records for use in retrospective analyses. J Clin Epidemiol. 2005;58(3):269–74.

[136] Wu L, Ashton CM. Chart review. A need for reappraisal. Eval Health Prof. 1997;20(2):146–63.

[137] Hess DR. Retrospective studies and chart reviews. Respir Care. 2004;49(10):1171–4.

[138] Cardo S, Agabiti N, Picconi O, Scarinci M, Papini P, Guasticchi G, et al. [The quality of medical records: a retrospective study in Lazio Region, Italy]. Ann Ig. 2003;15(5):433–42.

[139] Bush TL, Miller SR, Golden AL, Hale WE. Self-report and medical record report agreement of selected medical conditions in the elderly. Am J Public Health. 1989;79(11):1554–6.

[140] Westbrook JI, McIntosh JH, Rushworth RL, Berry G, Duggan JM. Agreement between medical record data and patients' accounts of their medical history and treatment for dyspepsia. J Clin Epidemiol. 1998;51(3):237–44.

[141] Jordan K, Jinks C, Croft P. Health care utilization: measurement using primary care records and patient recall both showed bias. J Clin Epidemiol. 2006;59(8):791–797.

[142] Tilley BC, Barnes AB, Bergstralh E, Labarthe D, Noller KL, Colton T, et al. A comparison of pregnancy history recall and medical records. Implications for retrospective studies. Am J Epidemiol. 1985;121(2):269–81.

[143] Casey R, Rieckhoff M, Beebe SA, Pinto-Martin J. Obstetric and perinatal events: the accuracy of maternal report. Clin Pediatr (Phila). 1992;31(4):200–4.

[144] O'Sullivan JJ, Pearce MS, Parker L. Parental recall of birth weight: how accurate is it? Archives of Disease in Childhood. 2000;82(3):202–203.

[145] Tate AR, Dezateux C, Cole TJ, Davidson L, Group tMCSCH. Factors affecting a mother's recall of her baby's birth weight. International journal of Epidemiology. 2005;34(3):688–695.

[146] Villar J, Dorgan J, Menendez R, Bolanos L, Pareja G, Kestler E. Perinatal data reliability in a large teaching obstetric unit. Br J Obstet Gynaecol. 1988;95(9):841–8.

[147] Borghi E, de Onis M, Garza C, et al. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. Statistics in Medicine. 2006;25(2):247–65.

[148] Stewart GB, Altman DG, Askie LM, Duley L, Simmonds MC, Stewart LA. Statistical Analysis of Individual Participant Data Meta-Analyses: A

Comparison of Methods and Recommendations for Practice. PLoS ONE. 2012;7(10):e46042.

[149] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58.

[150] Kahan BC, Morris TP. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects? Statistics in Medicine. 2013;32(7):1136–1149.

[151] Sauerbrei W, Royston P. A new strategy for meta-analysis of continuous covariates in observational studies. Statistics in Medicine. 2011;30(28):3341–3360.

[152] Verma V. Comparability in international survey statistics. In: International Conference on Improving Surveys, Copenhagen; 2002. p. 25–28.

[153] Cook DJ, Guyatt GH, andreas L, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. CHEST journal. 1992;102(Suppl 4):305S–311S.

[154] Deeks JJ. Systematic reviews of published evidence: Miracles or minefields? Annals of Oncology. 1998;9(7):703–709.

[155] Stewart GB, Altman DG, Askie LM, Duley L, Simmonds MC, Stewart LA. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. PloS ONE. 2012;7(10):e46042.

[156] Kahan BC, Morris TP. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects? Stat Med. 2013;32(7):1136–49.

[157] Sauerbrei W, Royston P. A new strategy for meta-analysis of continuous covariates in observational studies. Stat Med. 2011;30(28):3341–60.

[158] de Onis M. WHO Child Growth Standards based on length/height, weight and age. Acta Paediatrica. 2006;95(S450):76–85.

[159] de Onis M, Garza C, Victora CG, Onyango AW, Frongillo EA, Martines J. The WHO Multicentre Growth Reference Study: Planning, study design and methodology. Food and Nutrition Bulletin. 2004;21(1):15S–26S.

[160] Pickering RM, Weatherall M. The analysis of continuous outcomes in multicentre trials with small centre sizes. Statistics in Medicine. 2007;26(30):5445–56.

[161] Tangri N, Kitsios GD, Su SH, Kent DM. Accounting for center effects in multicenter trials. Epidemiology. 2010;21(6):912–3.

[162] Localio AR, Berlin JA, Have TRT, Kimmel SE. Adjustments for center in multicenter studies: an overview. Annals of Internal Medicine. 2001;135(2):112–23.

[163] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. journal of the Royal Statistical Society Series C (Applied Statistics). 1994;43(3):429–467.

[164] Bhandari N, Bahl R, Taneja S, de Onis M, Bhan MK. Growth performance of affluent Indian children is similar to that in developed countries. Bull World Health Organ. 2002;80(3):189–95.

[165] Owusu WB, Lartey A, de Onis M, Onyango AW, Frongillo EA. Factors associated with unconstrained growth among affluent Ghanaian children. Acta Paediatr. 2004;93(8):1115–9.

[166] Garza C, de Onis M. Rationale for developing a new international growth reference. Food Nutr Bull. 2004;25(1 Suppl):S5–14.

[167] de Onis M. Update on the implementation of the WHO child growth standards. World Review of Nutrition and Dietetics. 2013;106:75–82.

[168] Eskenazi B, Bradman A, Finkton D, Purwar M, Noble JA, Pang R, et al. A rapid questionnaire assessment of environmental exposures to pregnant women in the INTERGROWTH-21st Project. BJOG. 2013;120(s2):129–138.

[169] Villar J, Ismail LC, Victora CG, Ohuma EO, Bertino E, Altman DG, et al. International standards for newborn weight, length and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. The Lancet. 2014;384(9946):857–868.

[170] Cole TJ, Green PJ. Smoothing reference centile curves: The lms method and penalized likelihood. Statistics in Medicine. 1992;11(10):1305–19.

[171] Rigby RA, Stasinopoulos DM. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. Statistical Modelling. 2006;6(3):209–29.

[172] Rigby RA, Stasinopoulos DM. Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. Statistics in Medicine. 2004;23(19):3053–76.

[173] Ismail LC, Knight H, Bhutta Z, Chumlea W. International F, Century NGCfts. Anthropometric protocols for the construction of new international fetal and newborn growth standards: the INTERGROWTH-21st Project. BJOG. 2003;120(Suppl 2):42–7.

[174] Wright EM, Royston P. Comparison of statistical methods for age-related reference intervals. Journal of the Royal Statistical Society: Series A (Statistics in Society). 1997;160(1):47–69.

[175] Hynek M. Approaches for constructing age-related reference intervals and centile charts for fetal size. European journal for Biomedical informatics. 2010;6(1):51–60.

[176] Bland JM, Altman DG. Statistics notes: Transformations, means and confidence intervals. BMJ. 1996;312(738):1079.

[177] Nelson DB. Conditional Heteroskedasticity in Asset Returns: A New Approach. Econometrica. 1991;59(2):347–70.

[178] Royston P. A method for estimating age-specific reference intervals ('normal ranges') based on fractional polynomials and exponential transformation. journal of the Royal Statistical Society: Series A (Statistics in Society). 1998;161(1):79–101.

[179] Kurmanavicius J, Wright EM, Royston P, et al. Fetal ultrasound biometry: 1. Head reference values. BJOG. 1999;106(2):126–35.

[180] Hof M, Wit JM, Roede MJ. A method to construct age references for skewed skinfold data, using Box-Cox transformations to normality. Human Biology. 1985;57(1):131–9.

[181] Box GEP, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society Series B (Methodological). 1964;26(2):211–52.

[182] Cole TJ. Fitting smoothed centile curves to reference data. Journal of the Royal Statistical Society Series A (Statistics in Society). 1988;151(3):385–418.

[183] Cole TJ. Using the LMS method to measure skewness in the NCHS and Dutch National height standards. Annals of Human Biology. 1989;16(5):407–19.

[184] Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19(6):716–23.

[185] van Buuren S, Fredriks M. Worm plot: a simple diagnostic device for modelling growth reference curves. Statistics in Medicine. 2001;20(8):1259–77.

[186] Royston P, Wright EM. Goodness-of-fit statistics for age-specific reference intervals. Statistics in Medicine. 2000;19(21):2943–62.

[187] D'Agostino RB, Belanger A, D'Agostino J. A suggestion for using powerful and informative tests of normality. The American Statistician. 1990;44(4):316–21.

[188] Robert AR, Stasinopoulos DM. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. journal of Statistical Software. 2007;23((i07)):1–46.

[189] R Development Core Team (2010): R, a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2010;.

[190] Royston P, Wright EM. How to construct 'normal ranges' for fetal variables. Ultrasound in Obstetrics and Gynecology. 1998;11(1):30–8.

[191] Jones MC, Faddy MJ. A skew extension of the t-distribution, with applications. journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003;65(1):159–74.

[192] Wade A, Kurmanavicius J. Creating unbiased cross-sectional covariate-related reference ranges from serial correlated measurements. Biostatistics. 2009;10(1):147–154.

[193] Wei Y, Pere A, Koenker R, He X. Quantile regression methods for reference growth charts. Statistics in Medicine. 2006;25(8):1369–82.

[194] Koenker R. Quantile regresssion. Encyclopedia of Environmetrics. 2006;.

[195] Koenker R. Quantile regression for longitudinal data. journal of Multivariate Analysis. 2004;91(1):74–89.

[196] Muggeo VM, Sciandra M, Tomasello A, Calvo S. Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. Environmental and ecological statistics. 2013;20(4):519–531.

[197] Noufaily A, Jones MC. Parametric quantile regression based on the generalized gamma distribution. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2013;62(5):723–40.

[198] Hauspie RC, Cameron N, Molinari L. Methods in human growth research. vol. 39. Cambridge University Press; 2004.

[199] Rasbash J, Charlton C, Browne WJ, Healy M, Cameron B. MLwiN Version 2.1. Centre for multilevel modelling, University of Bristol. 2009;.

[200] Leckie G, Charlton C. A program to run the MLwiN multilevel modeling software from within STATA. journal of Statistical Software. 2013;52(11):1–40.

[201] Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika. 1986;73(1):43–56.

[202] Rutter CM, Elashoff RM. Analysis of longitudinal data: Random coefficient regression modelling. Statistics in Medicine. 1994;13(12):1211–31.

[203] Goldstein H. Efficient statistical modelling of longitudinal data. Annals of Human Biology. 1996;13(2):129–141.

[204] Rabe-Hesketh S, anders S. Multilevel and longitudinal modeling using Stata. STATA press; 2008.

[205] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38(4):963–74.

[206] Pan H, Goldstein H. Multilevel models for longitudinal growth norms. Statistics in Medicine. 1997;16(23):2665–78.

[207] Pan H, Goldstein H. Multilevel repeated measures growth modelling using extended spline functions. Statistics in Medicine. 1998;17(23):2755–70.

[208] Wade AM, Ades AE. Incorporating correlations between measurements into the estimation of age-related reference ranges. Statistics in Medicine. 1998;17(17):1989–2002.

[209] Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet. 1986;327(8476):307–10.

[210] Goldstein H. Multilevel statistical models. vol. 922. John Wiley and Sons; 2011.

[211] Kurmanavicius J, et al. Fetal ultrasound biometry: 2. Abdomen and femur length reference values. BJOG. 1999;106(2):136–143.

[212] Wade AM, Ades AE. Age-related reference ranges: significance tests for models and confidence intervals for centiles. Stat Med. 1994;13(22):2359–67.

[213] Salomon LJ, et al. The impact of choice of reference charts and equations on the assessment of fetal biometry. Ultrasound in Obstetrics and Gynecology. 2005;25(6):559–565.

[214] MacGregor SN, et al. Underestimation of gestational age by conventional crownrump length dating curves. Obstetrics and Gynecology. 1987;70((3 Pt 1)):344–348.

[215] Savitz DA, et al. Comparison of pregnancy dating by last menstrual period, ultrasound scanning and their combination. American journal of Obstetrics and Gynecology. 2002;187(6):1660–1666.

[216] Geirsson RT. Ultrasound instead of last menstrual period as the basis of gestational age assignment. Ultrasound in Obstetrics and Gynecology. 1991;1(3):212–219.

[217] Hall MH, et al. The extent and antecedents of uncertain gestation. Br J Obstet Gynaecol. 1985;92(5):445–51.

[218] Salpou D, et al. Fetal age assessment based on 2nd trimester ultrasound in Africa and the effect of ethnicity. BMC Pregnancy Childbirth. 2008;8(1):48.

[219] Campbell S. The prediction of fetal maturity by ultrasonic measurement of the biparietal diameter. BJOG. 1969;76(7):603–609.

[220] Robinson HP. Sonar measurement of fetal crown-rump length as means of assessing maturity in first trimester of pregnancy. BMJ. 1973;4(5883):28–31.

[221] Blaas HG, Eik-Nes SH, Bremnes JB. The growth of the human embryo. A longitudinal biometric assessment from 7 to 12 weeks of gestation. Ultrasound Obstet Gynecol. 1998;12(5):346–54.

[222] Hadlock FP, et al. Fetal crown-rump length: reevaluation of relation to menstrual age (5-18 weeks) with high-resolution real-time US. Radiology. 1992;182(2):501–5.

[223] Adam AH, Robinson HP, Dunlop C. A comparison of crown-rump length measurements using a real-time scanner in an antenatal clinic and a conventional Bscanner. Br J Obstet Gynaecol. 1979;86(7):521–4.

[224] Haglund B. Birthweight distributions by gestational age: comparison of LMP-based and ultrasound-based estimates of gestational age using data from the Swedish Birth Registry. Paediatric and Perinatal Epidemiology. 2007;21(s2):72–78.

[225] Grange G, et al. Dating biometry during the first trimester: accuracy of an everyday practice. European journal of Obstetrics and Gynecology and Reproductive Biology. 2008;88(1):61–64.

[226] Dietz PM, et al. A comparison of LMP-based and ultrasound-based estimates of gestational age using linked California livebirth and prenatal screening records. Paediatric and Perinatal Epidemiology. 2007;21(s2):62–71.

[227] Sladkevicius P, et al. Ultrasound dating at 12-14 weeks of gestation. A prospective cross-validation of established dating formulae in in-vitro fertilized pregnancies. Ultrasound in Obstetrics and Gynecology. 2005;26(5):504–511.

[228] Scott DW. Multivariate density estimation: theory, practice and visualization. John Wiley and Sons; 2015.

[229] Papageorghiou AT, Kennedy SH, Salomon LJ, Ohuma EO, Cheikh Ismail L, Barros FC, et al. International standards for early fetal size and pregnancy

dating based on ultrasound measurement of crownrump length in the first trimester of pregnancy. Ultrasound Obstet Gynecol. 2014;44(6):641–8.

[230] Harville EW, et al. Vaginal bleeding in very early pregnancy. Human Reproduction. 2003;18(9):1944–1947.

[231] Bottomley C, Bourne T. Dating and growth in the first trimester. Best Practice and Research Clinical Obstetrics and Gynaecology. 2009;23(4):439–452.

[232] BMJ. Vaginal bleeding in early pregnancy. BMJ. 1980;281(6238):470–470.

[233] Falco P, et al. Sonography of pregnancies with first-trimester bleeding and a viable embryo: a study of prognostic indicators by logistic regression analysis. Ultrasound in Obstetrics and Gynecology. 1996;7(3):165–169.

[234] Jacobs J, Katzur T, Tassenaar V. On estimators for truncated height samples. Economics and Human Biology. 2008;6(1):43–56.

[235] Komlos J. How to (and how not to) analyze deficient height samples. Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2004;37(4):160–173.

[236] A'Hearn B. A restricted maximum likelihood estimator for truncated height samples. Econ Hum Biol. 2004;2(1):5–19.

[237] Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2005;54(3):507–554.

[238] Cole TJ. Some questions about how growth standards are used. Horm Res. 1996;45(Suppl 2):18–23.

[239] Milani S. Invited review: Kinetic models for normal and impaired growth. Annals of Human Biology. 2000;27(1):1–18.

[240] Gasser T, et al. A method for determining the dynamics and intensity of average growth. Ann Hum Biol. 1990;17(6):459–74.

[241] Argyle J. Approaches to detecting growth faltering in infancy and childhood. Ann Hum Biol. 2003;30(5):499–519.

[242] Wright CM, et al. What is a normal rate of weight gain in infancy? Acta Paediatrica. 1994;83(4):351–356.

[243] Cole TJ. Conditional reference charts to assess weight gain in British infants. Archives of Disease in Childhood. 1995;73(1):8–16.

[244] Smith DW, et al. Shifting linear growth during infancy: illustration of genetic factors in growth from fetal life through infancy. Pediatr. 1976;89(2):225–30.

[245] Mei Z, et al. Shifts in percentiles of growth during early childhood: analysis of longitudinal data from the California Child Health and Development Study. Pediatrics. 2004;113(6):e617–27.

[246] Volkl TM, et al. Catch-down growth during infancy of children born small (SGA) or appropriate (AGA) for gestational age with short-statured parents. J Pediatr. 2006;148(6):747–52.

[247] Ong KKL, et al. Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. BMJ. 2000;320(7240):967–971.

[248] Lampl M, et al. Downward percentile-crossing as an indicator of an adverse prenatal environment. Annals of Human Biology. 2008;35(5):462–474.

[249] Tanner JM, Goldstein H, Whitehouse RH. Clinical longitudinal standards for height, weight, height velocity, weight velocity and stages of puberty. Archives of Disease in Childhood. 1976;51(3):170–179.

[250] Cameron N. Conditional standards for growth in height of British children from 5.0 to 15.99 years of age. Ann Hum Biol. 1980;7(4):331–7.

[251] Wilcox AJ. Birth weight, gestation and the fetal growth curve. Am J Obstet Gynecol. 1981;139(8):863–7.

[252] Marconi AM, et al. Comparison of fetal and neonatal growth curves in detecting growth restriction. Obstet Gynecol. 2008;112(6):1227–34.

[253] Emery JL, et al. Apnoea monitors compared with weighing scales for siblings after cot death. Arch Dis Child. 1985;60(11):1055–60.

[254] Chang TC, Robson SC, Spencer JA, Gallivan S. Prediction of perinatal morbidity at term in small fetuses: comparison of fetal growth and Doppler ultrasound. Br J Obstet Gynaecol. 1994;101(5):422–7.

[255] de Jong CLD, et al. Fetal growth rate and adverse perinatal events. Ultrasound in Obstetrics and Gynecology. 1999;13(2):86–89.

[256] Hemachandra AH, Klebanoff MA. Use of serial ultrasound to identify periods of fetal growth restriction in relation to neonatal anthropometry. Am J Hum Biol. 2006;18(6):791–7.

[257] Verkauskiene R, et al. Impact of fetal growth restriction on body composition and hormonal status at birth in infants of small and appropriate weight for gestational age. Eur J Endocrinol. 2007;157(5):605–12.

[258] Cameron N, Preece MA, Cole TJ. Catch-up growth or regression to the mean? Recovery from stunting revisited. Am J Hum Biol. 2005;17(4):412–7.

[259] Bishop L, Horton L. Managing and sharing data: Best practice for researchers. UK Data Archive, University of Essex. 2011;.

[260] Waterlow JC, et al. The presentation and use of height and weight data for comparing the nutritional status of groups of children under the age of 10 years. Bulletin of the World Health Organization. 1977;55(4):489–498.

[261] Sovio U, White IR, Dacey A, Pasupathy D, Smith GC. Screening for fetal growth restriction with universal third trimester ultrasonography in nulliparous

women in the Pregnancy Outcome Prediction (POP) study: a prospective cohort study. The Lancet. 2015;386(10008):2089–2097.

[262] Fowden AL, Giussani DA, Forhead AJ. Intrauterine programming of physiological systems: causes and consequences. Physiology (Bethesda). 2006;21(1):29–37.

[263] NICE. National Collaborating Centre for Women's and Children's Health. Royal College of Obstetricians and Gynaecologists Press: London. 2008;.

[264] ACOG. Practice Bulletin No. 101: Ultrasonography in pregnancy. Obstet Gynecol. 2009;113(2 Pt 1):451–61.

[265] Gluckman PD, et al. Towards a new developmental synthesis: adaptive developmental plasticity and human disease. The Lancet. 2009;373(9675):1654–1657.

[266] Gluckman PD, Hanson MA, Beedle AS. Early life events and their consequences for later disease: a life history and evolutionary perspective. Hum Biol. 2007;19(1):1–19.

[267] Barker DJ. The origins of the developmental origins theory. J Intern Med. 2007;261(5):412–7.

[268] Gluckman PD, Hanson MA. The developmental origins of health and disease. Springer; 2006.

[269] RCOG. Green-top guideline no. 31: The investigation and management of the small-for-gestational-age fetus. Royal College of Obstetricians and Gynaecologists Press. 2013;.

[270] Roche A, Himes JH. Incremental growth charts. Am J Clin Nutr. 1980;33(9):2041–52.

[271] Baumgartner RN, Roche AF, Himes JH. Incremental growth tables: supplementary to previously published charts. Am J Clin Nutr. 1986;43(5):711–22.

[272] Roche AF, Guo S, Moore WM. Weight and recumbent length from 1 to 12 mo of age: reference data for 1-mo increments. J Clin Nutr. 1989;49(4):599–607.

[273] WHO Child Growth Standards: Growth velocity based on weight, length and head circumference: Methods and development. Geneva: World Health Organization. 2009;242.

[274] Healy MJR, et al. The use of short-term increments in length to monitor growth in infancy. Nestle nutrition workshop series. 1988;14:41–55.

[275] Boryslawski K. Structure of monthly increments of length, weight and head circumference in the first year: a pure longitudinal study of 200 Wroclaw infants. Ann Hum Biol. 1988;15(3):205–12.

[276] Thompson H. Data on the growth of children during the first year after birth. Human Biology. 1951;23(2):75–92.

[277] Cole TJ. The use and construction of anthropometric growth reference standards. Nutr Res Rev. 1993;6(1):19–50.

[278] Bairagi R. On Components of Variation of Estimated Weight Velocity of Children. journal of the Royal Statistical Society. 1986;35(2):178–182.

[279] Lampl M, Veldhuis JD, Johnson ML. Saltation and stasis: a model of human growth. Science. 1992;258(5083):801–3.

[280] Marshall WA. Evaluation of growth rate in height over periods of less than one year. Archives of Disease in Childhood. 1971;46(248):414–420.

[281] Tanner JM, Whitehouse RH, Takaishi M. Standards from birth to maturity for height, weight, height velocity and weight velocity: British children, 1965. II. Arch Dis Child. 1966;41(220):613–35.

[282] Campbell S, Newman GB. Growth of the fetal biparietal diameter during normal pregnancy. BJOG. 1971;78(6):513–519.

[283] Cole TJ. Presenting information on growth distance and conditional velocity in one chart: practical issues of chart design. Stat Med. 1998;17(23):2697–707.

[284] Galton F. Regression towards mediocrity in hereditary stature. journal of the Anthropological Institute of Great Britain and Ireland. 1886;15:246–263.

[285] Healy MJ, Goldstein H. Regression to the mean. Ann Hum Biol. 1978;5(3):277–80.

[286] Bland JM, Altman DG. Statistics Notes: Some examples of regression towards the mean. BMJ. 1994;309(6957):780.

[287] Cole TJ. Growth charts for both cross-sectional and longitudinal data. Stat Med. 1994;13(23-24):2477–92.

[288] Prader A, Tanner JM, von HG. Catch-up growth following illness or starvation. An example of developmental canalization in man. J Pediatr. 1963;62(5):646–59.

[289] Stratton JF, et al. Are babies of normal birth weight who fail to reach their growth potential as diagnosed by ultrasound at increased risk? Ultrasound Obstet Gynecol. 1995;5(4):114–8.

[290] Foulkes MA, Davis CE. An Index of Tracking for Longitudinal Data. Biometrics. 1981;37(3):439–446.

[291] Tanner JM. The assessment of growth and development in children. Archives of Disease in Childhood. 1952;27(131):10–33.

[292] Cole TJ. 3-in-1 weight-monitoring chart. The Lancet. 1997;349(9045):102–103.

[293] Papageorghiou A, et al. Ultrasound methodology used to construct the fetal growth standards in the INTERGROWTH-21st Project. BJOG. 2013;120(Suppl 2):27–32.

[294] Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915;10(4):507–521.

[295] Prentice AM, Nabwera H, Unger S, Moore SE. Growth monitoring and the prognosis of mortality in low-income settings. The American Journal of Clinical Nutrition. 2016;103(3):681–682.

[296] Black RE, Victora CG, Walker SP, Bhutta ZA, Christian P, de Onis M, et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. Lancet. 2013;382(9890):427–51.

[297] Schwinger C, Fadnes LT, Van den Broeck J. Using growth velocity to predict child mortality. The American Journal of Clinical Nutrition. 2016;103(3):801–807.

[298] Chang TC, Robson SC, Spencer JA, Gallivan S. Prediction of perinatal morbidity at term in small fetuses: comparison of fetal growth and Doppler ultrasound. Br J Obstet Gynaecol. 1994;101(5):422–7.

[299] de Jong CL, Francis A, van Geijn HP, Gardosi J. Fetal growth rate and adverse perinatal events. Ultrasound Obstet Gynecol. 1999;13(2):86–9.

[300] Bailey BJ. Monitoring the heights of prepubertal children. Ann Hum Biol. 1994;21(1):1–11.

[301] Voss LD, et al. The reliability of height and height velocity in the assessment of growth (the Wessex Growth Study). Archives of Disease in Childhood. 1991;66(7):833–837.

[302] Cole TJ, Hall DM. Screening for growth: towards 2000. Archives of Disease in Childhood. 1996;74(2):183–183.

[303] Lyon AJ, Preece MA, Grant DB. Growth curve for girls with Turner syndrome. Archives of Disease in Childhood. 1985;60(10):932–935.

[304] Cole TJ. Do growth chart centiles need a face lift? BMJ. 1994;308(6929):641–2.

[305] Wright CM, et al. New chart to evaluate weight faltering. Archives of Disease in Childhood. 1998;78(1):40–43.

[306] Argyle J, Seheult AH, Wooff DA. Correlation models for monitoring child growth. Statistics in Medicine. 2008;27(6):888–904.