Comparison of first trimester dating methods for gestational age estimation and their implication on preterm birth classification in a North Indian cohort

Ramya Vijayram[1,2,*], Nikhita Damaraju[1,2,*], Ashley Xavier[1,2,*], Koundinya Desiraju[3,4], Ramachandran Thiruvengadam[3,4], GARBH-Ini[4], Raghunathan Rengasamy[2,5,6], Himanshu Sinha[1,2,6], Shinjini Bhatnagar[3,4]

[1] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India

[2] Initiative for Biological Systems Engineering (IBSE), Indian Institute of Technology Madras, Chennai, India

[3] Maternal and Child Health Program, Translational Health Science and Technology Institute, Faridabad, India

[4] Interdisciplinary Group for Advanced Research on Birth Outcomes - DBT India Initiative, Translational Health Science and Technology Institute, Faridabad, India

[5] Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai, India

[6] Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), Indian Institute of Technology Madras, Chennai, India

[*] These authors contributed to the equally

Correspondence

For GARBH-Ini related queries: Shinjini Bhatnagar, Maternal and Child Health program, Translational Health Science and Technology Institute, 3rd Milestone, Faridabad-Gurgaon Expressway, PO Box #04, Faridabad, Haryana, India. Email: shinjini.bhatnagar@thsti.res.in

For data science-related queries: Himanshu Sinha, Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India. Email: sinha@iitm.ac.in

ABSTRACT

Background

Different methods and formulae have been developed for different populations for estimation of GA in the first trimester of pregnancy. In this study, we develop an Indian population-specific GA dating formula and compare its performance with the previously published formulae. Finally, we evaluate the implications of the choice of dating method on preterm birth (PTB) rate. The data for this study was from the GARBH-Ini cohort, an ongoing (2015-2019) longitudinal pregnancy cohort of North Indian women to study PTB.

Methods

Comparisons between USG (Hadlock) and LMP-based dating methods were made by studying the distribution of their differences by BA analysis. Population-specific dating formula for the first trimester of pregnancy (Garbhini-1 formula) was developed by constructing a regression model for GA as a non-linear function of CRL, which was then compared with published formulae by BA analysis. The PTB rate was estimated by using each of these methods and expressed as prevalence and 95% CI.

Results

LMP-based method overestimates GA by three days as compared to USG (Hadlock) method with limits of agreement between -4.39 and 3.51 weeks (95% CI). CRL is the most critical parameter in estimating GA in the first trimester. No other clinical or socioeconomic parameter enumerated in the GARBH-Ini cohort study contributes to GA estimation. The GA estimated by all the formulae compared showed an agreement within a week with the uppermost and the lowermost limits of agreement (LOA) being -0.46,0.96 weeks. The estimated PTB rate across all the formulae ranged between 12.12 and 16.85% with Garbhini-1 formula estimating the least rate.

Conclusions

Our study reinforces the fact that CRL-based USG method is best for estimation of GA in the first trimester and addition of clinical and demographic features does not improve its accuracy. Garbhini-1 formula developed from our population data performs at par with the existing formulae but estimates the lowest PTB rate with better precision than other formulae. The applicability of Garbhini-1 formulae for the rest of the Indian population needs to be validated in subsequent studies.

SYNOPSIS

Study question

- Is there a need for an Indian population-specific GA estimation model?
- Which clinical and socioeconomic features affect the estimation of GA in an Indian population?
- Does the choice of a GA estimating model affect the classification of PTB?

What is already known

- Several first trimester GA estimation formulae have been published based on different population studies.
- In India, Hadlock's formula, based on a US population, is primarily used for GA estimation.
- Reliable GA is required for accurately estimating the PTB rate in a population.

What this study adds

- We have developed Indian population-specific Garbhini-1 formula for GA estimation in the first trimester.
- Garbhini-1 formula performs comparably to other published formulae in estimating GA and in classifying PTB.
- In the first trimester, CRL is the only feature that affects GA estimation.

KEYWORDS

Gestational age; Crown-rump length; CRL; Preterm birth; Last menstrual period; GARBH-Ini; Machine learning

Word count: 3511

1. BACKGROUND

Preterm birth (PTB) is conventionally defined as a birth that occurs before 37 completed weeks of gestation[1]. It is a unique disease in the way it is defined by the duration of gestation and not by a pathological process. The duration of gestation is the period between the date of conception and date of delivery. While the date of delivery can be documented with fair accuracy, ascertainment of date of conception is challenging. The estimation of gestational age (GA) during the antenatal period also called as the dating of pregnancy has been conventionally done using the first day of the recall-based last menstrual period (LMP) or measurement of foetal biometry by ultrasonography (USG). Each of these methods poses a unique set of challenges. The accuracy of dating by LMP method is dependent on accurate recall, and regularity of menstrual cycle which, in turn, is affected by numerous physiological and pathological conditions such as obesity[2], polycystic ovarian syndrome[3], breast feeding[4] and use of contraceptive methods[5].

The USG method is based on foetal biometry using crown-rump length (CRL) in the first trimester. Several formulae exist to estimate GA using CRL, including Hadlock's formula[6], based on a US population-based study, which is widely used in India. However, the choice of dating formula might influence the accuracy of dating, as these formulae have been developed from studies that differed both in the study population and study design. The error and bias due to the choice of a dating formula need to be quantitatively studied to estimate the rate of PTB in a specific population accurately. In addition to its public health importance, accurate dating is essential for clinical decision making during the antenatal period, such as for scheduling monitoring visits and recommending appropriate antenatal care.

In this study, we first quantified the discrepancy between LMP and USG-based (Hadlock) dating methods during the first trimester in an Indian population. We characterised how each method could contribute to the discrepancy in calculating the GA. We then built our population-specific model from the GARBH-Ini cohort (Interdisciplinary Group for Advanced Research on BirtH outcomes - DBT India Initiative), Garbhini-1, and compared its performance with the published 'high quality' formulae for the first trimester dating[7]- McLennan and Schluter[8], Robinson and Fleming[9], Sahota[10] and Verburg[11], INTERGROWTH-

21[12], and Hadlock's formula[6] (Supplementary Table 1). Finally, we quantified the implications that the choice of dating methods based on different formulae would have on the PTB rate in our study population.

2. METHODS

2.1. Study design

The GARBH-Ini cohort is a prospective observational cohort of pregnant women initiated in May 2015 at the District Civil Hospital that serves a large rural and semi-urban population in the Gurugram district, Haryana, India. The main aim of the cohort study is to develop an effective risk stratification that facilitates timely referral for a critical health care intervention for high-risk women, particularly in low- and middle-income countries. Women in the GARBH-Ini cohort are enrolled within 20 weeks of gestation and are followed three times during pregnancy till delivery and one visit postpartum[13]. After a verbal consent to be interviewed, informed consent to screen is obtained for women who are at < 20-weeks period of gestation (POG) calculated by the last menstrual period. A dating ultrasound is performed within the week to confirm a viable intrauterine pregnancy with < 20-weeks POG using standard foetal biometric parameters. A time-series data on a large set of variables- including clinical, environmental, genomic, epigenomic, metabolomic, and proteomic is collected across pregnancy to help in stratifying women into defined risk groups for PTB.

2.2. Sampling strategy and participant datasets derived for the study

The samples of this study were derived from the first 3,499 participants enrolled in the GARBH-Ini study. We included 1,721 participants (N_p=1,721) who had POG <14 weeks and had information on the LMP, CRL on the ultrasound and singleton pregnancy which advanced beyond 20 weeks of gestation, i.e. the pregnancy did not end in a spontaneous abortion. For 1,721 participants, more than one scan performed during this period, and data from both the scans were included as unique observations (N_o). These participants contributed a total of 2,562 observations (N_o=2,562) that was used for further analyses, and this dataset of observations was termed as the *main dataset* (N_o=2,562 observations from N_p=1,721 participants; Figure 1). The *main dataset* was used to develop a population-based dating model named Garbhini-1, for the first trimester.

It is essential to independently evaluate models on data that was not used for building the model. This is to eliminate any biases that may have been incorporated due to the iterative learning process of the model building dataset and estimate the expected performance when applying the model on new data in the real world. We used an unseen *test dataset* created from the 1,500 participants enrolled further in this cohort (Figure 1). By applying identical processing steps as described for the *main dataset*, the *test dataset* (N_o=808 observations from N_p=559 participants) was obtained (Figure 1).

2.3. Assessment of LMP, CRL and CRL-based GA

The date of LMP was ascertained from the participant's recall of the first day of the last menstrual period. CRL from an ultrasound image (GE Voluson E8 Expert, General Electric Healthcare, Chicago, Illinois) was captured in midline sagittal section of the whole foetus by placing the callipers on the outer margin of skin borders of the foetal crown and rump. The CRL measurement was done thrice on three different ultrasound images, and the average of the three measurements was considered for estimation of GA. Under the supervision of medically qualified researchers, nurses documented the clinical and socio-demographic characteristics[13].

2.4. Development of population-specific gestational dating model

We created two subsets from the *main dataset* for developing the first trimester population-based dating formula and its comparison with the existing published models, based on two approaches. The first approach excluded participants with potentially unreliable LMP or high risk of foetal growth restriction, giving us the *clinically-filtered* dataset (N_o=980 observations from N_p=650 participants, Figure 1, Supplementary Table 2).

The second approach used Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method to remove outliers based on noise in the data points. DBSCAN identifies 'noise' by classifying points into clusters if there are a sufficient number of neighbours that lie within a specified Euclidean distance or if the point is adjacent to another data point meeting the criteria[14]. DBSCAN was used to identify and remove outliers in the *main* dataset using the parameters for distance cut-off (epsilon, *eps*) 0.5 and the minimum number of neighbours (*minpoints*) 20. A range of values for *eps* and *minpoints* did not markedly change

the clustering result (Supplementary Table 3). The resulting dataset that retained reliable data points for the analysis was termed as the *dbscan* dataset (N_o=2,156 observations from N_p=1,476 participants, Figure 1).

Development of a first trimester dating formula based on the GARBH-Ini cohort data was done by fitting models on the above two datasets. Linear, quadratic and cubic regression formulae were developed for GA (weeks) as a function of CRL (cm) on the *main* dataset. The performance of these formulae was tested in the unseen *test* dataset.

In addition to CRL as a primary indicator, to identify other variables which may be predictive of GA during the first trimester, a list of 282 candidate variables were explored by feature selection methods on the *dbscan* dataset. These methods helped to find uncorrelated, non-redundant features that might improve the accuracy of GA prediction (Supplementary Table 4). First, the feature selection was done using Boruta[15], a random forest classifier, which identified six features and second, by implementing Generalised Linear Modelling (GLM) that identified two features as candidate features to predict GA. A union of features from these two methods (Supplementary Table 5), gave a list of six candidate predictors. Equations were generated using all combinations of these predictors in the form of linear, logarithmic, polynomial and fractional power equations. The best fit model was termed Garbhini-1 formula and was validated for its performance in a *test* dataset.

2.5. Comparison of LMP- and USG-based dating methods during the first trimester

To estimate the effect of factors that contribute to the accuracy of LMP and factors that contribute to foetal growth restriction, two sub-datasets were derived from the *main* dataset (Figure 1). The first sub-dataset, *dataset1* (N_o=1,261 observations from N_p=791 participants), excluded the *main* dataset participants with unreliable LMP recall (Figure 1). These factors included the use of contraceptives a month prior to the pregnancy, assisted conception, enrolment BMI outside the normal range, and breastfeeding in the two months before conception. Therefore, this *dataset1* was considered as a dataset of participants with more reliable LMP recall.

The second sub-dataset, *dataset2* (N_o=1,281 observations from N_p=820 participants), was formed by excluding the *main* dataset participants with a known risk of foetal growth restriction (Figure 1). The factors that affect CRL measurements included active and passive smoking, consumption of tobacco and alcohol, and enrolment BMI outside the normal range. This *dataset2* was considered as a dataset of participants with low foetal growth restriction risk.

We calculated the difference between LMP- and USG-based GA for each participant at the time of enrolment in the cohort and studied the distribution of the differences by Bland-Altman (BA) analysis[16]. The mean difference between the methods and the limits of agreement (LOA) for 95% CI were reported. To quantify the contribution of LMP and USG-based dating methods to this discrepancy, BA analysis was performed on datasets with reliable LMP (*dataset1*) and low-risk of foetal growth restriction (*dataset2*). The PTB rates with LMP and USG-based methods were reported per 100 live births with 95%CI. Next, different USG-based formulae were compared by calculating correlations between them. The bias between different formulae was evaluated by BA analysis, and pairwise mean difference and LOA were reported.

2.6. Data analysis, machine learning and statistical analysis for population specific first trimester dating model

The data analyses were carried out in R versions 3.6.1 and 3.5.0. DBSCAN was implemented using the package `dbscan`, and the random forests feature selection was done using the `Boruta` package[15]. Statistical analysis for comparison of PTB rate as estimated using different dating formulae was carried out using standard *t*-test with or without Bonferroni multiple testing correction or using Fisher's exact test.

2.7. Ethics approval

Ethics approvals were obtained from the Institutional Ethics Committees of Translational Health Science and Technology Institute; District Civil Hospital, Gurugram; Safdarjung Hospital, New Delhi (ETHICS/GHG/2014/1.43); and Indian Institute of Technology Madras (IEC/2019-03/HS/01/07). Written informed consent was obtained from all study participants

enrolled in the GARBH-Ini cohort. For an illiterate woman, details of the study were explained in the presence of a literate family member or a neighbour who acted as the witness; a verbal consent and a thumb impression were taken from her along with the signature of the witness.

3. RESULTS

3.1. Description of participants included in the study

The median age of pregnant women enrolled in the cohort was 23.0 (IQR 21.0,26.0) years, and about half of them were primigravida. The median weight and height of these participants were 47.0 kg (IQR 42.5,53.3) and 153.0 cm (IQR 149.2,156.8), respectively. The median first trimester BMI of the participants was 20.09 (IQR 18.27,22.59), of which 59.93% were in the normal range. Almost all (98.2%) participants were from the middle or lower socioeconomic strata, and interestingly nearly half (56.25%) of the women were from a nuclear family. The participants selected for this analysis had a median GA of 11.71 weeks (IQR 9.29,13.0). The detailed baseline characteristics are given in Table 1.

3.2. Comparison of USG-Hadlock and LMP-based methods for estimation of GA in the first trimester

The mean difference between USG-Hadlock and LMP-based dating at the time of enrolment was found to be -0.44±2.02 weeks (Figure 2a). This difference means that the LMP-based method overestimated GA by nearly three days as compared to USG (Hadlock)-based estimation. The LOA determined by BA analysis was wide -4.39,3.51 weeks, with 8.82% of participants falling beyond these limits (Figure 2b) suggesting a high imprecision in both the methods. The LOA between USG-Hadlock and LMP-based dating narrowed when tested on *dataset1* (dataset with reliable LMP). Further, the LOA narrowed marginally in the population that excluded those with risk factors of foetal growth restriction (*dataset2*, mean difference -0.34±1.22; LOA -4.13,3.21). Such wide LOA even after ensuring reliable LMP by classically used clinical criteria and standardised CRL measurements in the cohort represents the residual imprecision due to unknown factors in the estimation of GA.

3.3. Development of Garbhini-1 formula for first trimester dating

The discrepancy between in CRL- and LMP-based methods of GA estimation warranted an attempt to develop population specific first trimester dating model. In order to remove noise from the *main* dataset for building population-specific first trimester dating models, two methods were used- clinical criteria-based filtering and DBSCAN (Figure 1). When clinical criteria were used, more than two-third observations (68.46%) were excluded (Figure 3a). However, when DBSCAN was implemented, less than one-fifth observations (15.85%) were removed (Figure 3b). Models for first trimester dating using *clinically-filtered* and *dbscan* datasets using CRL as the only predictor was done using linear, quadratic and cubic regression to identify what fit gives the best predictive response (Supplementary Figure 1). Regression lines of the models based on *clinically-filtered* dataset showed a decline estimated GA as the CRL increased beyond a specific limit which is biologically implausible. So, our *clinically-filtered* dataset was not used for building any dating model. As the DBSCAN approach provided a more accurate dataset (which meant that there was no artefact as observed in the *clinically-filtered* dataset) with lesser outliers, the dating models using CRL were developed using *dbscan* dataset. Comparison among various dating models showed that the best correlation value (R^2) was for quadratic regression (R^2=0.86) with no further improvement when tested for cubic regression (R^2=0.86). Therefore, the following quadratic formula was decided as the final model for estimating GA in the first trimester and was termed as Garbhini-1 formula, where GA is in weeks, and CRL is in cm.

$$GA = -0.063706(CRL^2) + 1.420584(CRL) + 6.455422$$

A multivariate dating model including CRL and the six additional predictors identified by data-driven approaches (GLM and Random forests): resident state, weight, BMI, abdominal girth, age, and maternal education did not improve the performance of the CRL-based dating model (Supplementary Figure 2, Supplementary Table 6).

3.4. Comparison of published formulae and Garbhini-1 formula for estimation of GA

The actual test of the validity of a formula is to estimate GA reliably in an unseen sample population. So, to test the performance of the published formulae (Supplementary Table 1) and Garbhini-1 formula independently, GA was estimated for the *test dataset* (N_o=808

observations from N_p=559 participants, see Methods, Figure 1). It was observed that the correlation values of all the formulae were identical to each other (R^2=0.58; Supplementary Table 7, Supplementary Figure 3). Furthermore, all possible pairwise, BA analysis of these formulae showed that the mean difference of estimated GA varied from -0.17 to 0.50 weeks (Table 2). This result shows that Garbhini-1 performs equally well as other formulae. Since it is developed using an India-population cohort, Garbhini-1 might validate well across different populations in our settings.

3.5. Impact of the choice of USG dating formula on the estimation of the rate of PTB

When the dating of pregnancy was done using LMP, the estimated PTB rate was 13.30% (95%CI: 11.94, 14.77%), while using USG-Hadlock formula, it was 14.79% (95%CI: 13.36, 16.31%, Table 3). This underestimation of PTB rate by LMP-based method was found to be non-significant (Fisher's exact test, p=0.16). Furthermore, comparisons of various published formulae and Garbhini-1 formula showed that the estimated rate of PTB varied between 12.12 (10.81,13.53) and 16.85 (15.34,18.45) %. Garbhini-1 formula estimated the least PTB rate while Robinson-Fleming formula estimated the highest. Most differences between the PTB rates were statistically not significant except between Robinson-Fleming and Garbhini-1, McLennan-Schluter and Garbhini-1 dating (using Fisher's exact test with Bonferroni correction for p<0.05, Table 3). The estimation of PTB rate when using Garbhini-1 with a narrower 95%CI best represents the PTB rate in the population would make a strong argument in favour of using this formula in the Indian population.

4. COMMENT

4.1. Principal findings

The main objectives of this study were to compare different methods and formulae used for GA estimation during the first trimester and develop a population-specific dating model for the first trimester. Our findings show that there is a wide discrepancy between GA estimated from LMP and USG (Hadlock) formula. This discrepancy could not be accounted for by classical factors of LMP reliability or foetal growth restriction. This shows that there is a significant degree of noise in both LMP- and USG-based methods that should be considered when estimating GA as well as the PTB rate. To better represent and estimate GA and PTB rates in an Indian population, Garbhini-1 formula was developed. Garbhini-1

formula performs similar to previously published formulae in estimating GA in the first trimester using CRL as the sole feature. No other predictors identified from machine learning tools helped in improving the performance of the Garbhini-1 formula. However, Garbhini-1 formula estimated the lowest PTB rate amongst all the formulae that we tested in our population and this estimate was significantly different from two formulae commonly used in the West: Robinson-Fleming and McLennan-Schluter.

4.2. Strengths of the study

Using data-driven approaches, we have developed an India population-specific first trimester GA estimation formula, Garbhini-1, which uses CRL as the sole parameter and performs equally well as other published formulae for the first trimester. Using machine learning techniques, several parameters were identified and tested, but no other feature other than CRL was found to relevant for estimating GA in the first trimester. Further, we used a data-driven approach to remove outliers while building the GA estimation model. Such an approach retained a larger proportion of points than clinical criteria-based filtering for the reference standard. Another strength of our study is the standardised measurement of CRL. This reduces the imprecision to the minimum possible and makes USG–based estimation of gestational age reasonably accurate.

4.3. Limitations of the data

For the development of Garbhini-1 model, it would have been ideal to have used documented LMP collected pre-conceptionally. Since our GARBH-Ini cohort enrols participants in the first trimester of pregnancy, clinical criteria based on data collected using a questionnaire was used to derive a subset of participants with reliable LMP. This was incomplete as we had residual imprecision, which was not accounted for by the clinical criteria.

4.4. Interpretation

The LMP-based dating is prone to errors from recall and irregularity of menstrual cycles due to both physiological causes and pathological conditions. The errors in LMP-based dating methods have been found to be systematic and associated with biases in the GA estimation:

inaccurate recall of LMP which tended to overestimate the time since LMP has been seen in studies from Africa and other Western populations[20,21].

In our cohort GARBH-Ini representing a North Indian urban and semi-urban population, predominantly from middle and lower socioeconomic strata, the discrepancy in LMP and USG (Hadlock) based dating was found to be three days during the first trimester. While this is not significant, this discrepancy can result in PTB false positive rates. An Indian population based first trimester CRL-based dating formula, Garbhini-1 was developed, and it predicted GA as well as the most commonly used Hadlock formula. The lower false positive rate of PTB estimation by Garbhini-1 formula has significant clinical consequences and financial ramifications, particularly in low- and middle-income countries. Therefore, the development of population-specific GA dating formulae accounts not only for population-based differences in foetal development but also to better classify PTB in populations.

5. CONCLUSIONS

CRL-based USG method is the best for estimation of GA in the first trimester and addition of clinical and demographic features does not improve its accuracy. Garbhini-1 formula is an Indian-population based formula for estimating the GA in the first trimester based on CRL as a parameter. It has better precision than the most commonly used Hadlock formula in estimating the rate of PTB. These results need to be further validated in subsequent multi-ethnic cohorts before it can be applied for wider use.

Disclosure of interests

All listed authors declare that they have no conflicts of interest.

Contribution to authorship

RT, HS, SB conceived this study, RV, ND, AX, performed data and statistical analyses, KD, RT performed data exports and contributed to data analysis, RR provided critical feedback on data analysis, KD, RT, HS, SB interpreted the results, RV, ND, AX, HS wrote the first draft of the manuscript and all listed authors critically revised subsequent manuscript drafts and contributed essential discussion points. All authors approved the final draft of the manuscript.

Funding

This study was funded by an intramural grant from Initiative for Biological Systems Engineering, IIT Madras (BIO/18-19/304/ALUM/KARH). GARBH-Ini cohort study is funded by Department of Biotechnology, Government of India (BT/PR9983/MED/97/194/2013) and for some components of the biorepository by the Grand Challenges India-All Children Thriving Program (supported by the Programme Management Unit), Biotechnology Industry Research Assistance Council, Department of Biotechnology, Government of India (BIRAC/GCI/0114/03/14-ACT).

Acknowledgements

We thank the participants of GARBH-Ini study. We thank Karthik Raman, Nirav Bhatt and other colleagues at Initiative of Biological Systems Engineering, IIT Madras for their valuable suggestions. The work comprising the analyses of data was conducted at the Initiative of Biological Systems Engineering that is in part supported by an alumni endowment from Dr Prakash Arunachalam.

CODE AVAILABILITY

All the codes used for this paper are available at https://github.com/HimanshuLab/GARBH-Ini_1

SUPPORTING INFORMATION

Additional supporting information is available online.

REFERENCES

1. Preterm birth [Internet]. [cited 2019 Nov 25]. Available from: https://www.who.int/news-room/fact-sheets/detail/preterm-birth
2. Wei S, Schmidt MD, Dwyer T, Norman RJ, Venn AJ. Obesity and menstrual irregularity: associations with SHBG, testosterone, and insulin. *Obes Sci Pract.* 2009 May;17(5):1070–6.

3. Lobo RA. What are the key features of importance in polycystic ovary syndrome? *Fertil Steril.* 2003 Aug 1;80(2):259–61.
4. Chowdhury R, Sinha B, Sankar MJ, Taneja S, Bhandari N, Rollins N, et al. Breastfeeding and maternal health outcomes: a systematic review and meta-analysis. *Acta Paediatr.* 2015 Dec;104(467):96–113.
5. Creinin MD, Keverline S, Meyn LA. How regular is regular? An analysis of menstrual cycle regularity. *Contraception*. 2004 Oct;70(4):289–92.
6. Hadlock FP, Shah YP, Kanon DJ, Lindsey JV. Fetal crown-rump length: reevaluation of relation to menstrual age (5-18 weeks) with high-resolution real-time *US Radiology.* 1992 Feb 1;182(2):501–5
7. Napolitano R, Dhami J, Ohuma EO, Ioannou C, Conde-Agudelo A, Kennedy SH, et al. Pregnancy dating by fetal crown-rump length: a systematic review of charts. *BJOG.* 2014 Apr;121(5):556–65.
8. McLennan AC, Schluter PJ. Construction of modern Australian first trimester ultrasound dating and growth charts. *J Med Imag Radiat On.* 2008;52(5):471–9.
9. Robinson HP, Fleming JEE. A Critical Evaluation of Sonar “Crown-Rump Length” Measurements. *BJOG.* 1975;82(9):702–10.
10. 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 Obst Gyn.* 2009;33(2):157–60.
11. Verburg BO, Steegers E a. P, Ridder MD, Snijders RJM, Smith E, Hofman A, et al. New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study. *Ultrasound Obst Gyn*. 2008;31(4):388–96.
12. Papageorghiou AT, Kennedy SH, Salomon LJ, Altman DG, Ohuma EO, Stones W, et al. The INTERGROWTH-21st fetal growth standards: toward the global integration of pregnancy and pediatric care. *Am J Obstet Gynecol*. 2018 Feb 1;218(2, Supplement):S630–40.
13. Bhatnagar S, Majumder PP, Salunke DM. A Pregnancy Cohort to Study Multidimensional Correlates of Preterm Birth in India: Study Design, Implementation, and Baseline Characteristics of the Participants. *Am J Epidemiol*. 2019 Apr 1;188(4):621–31.

14. Ester M, Kriegel H-P, Sander J, Xu X. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining [Internet]. *AAAI Press;* 1996. p. 226–231. (KDD'96). Available from: http://dl.acm.org/citation.cfm?id=3001460.3001507
15. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw.* [Internet]. 2010 [cited 2019 Nov 25];036(i11). Available from: https://ideas.repec.org/a/jss/jstsof/v036i11.html
16. Martin Bland J, Altman Douglas G. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet.* 1986 Feb 8;327(8476):307–10.
17. Obesity and Colorectal Cancer [Internet]. [cited 2019 Dec 15]. Available from: /https://icmr.nic.in/sites/default/files/icmr_bulletins/Bul_July_Sept.pdf
18. Wegienka G, Baird DD. A Comparison of Recalled Date of Last Menstrual Period with Prospectively Recorded Dates. *J Women's Health*. 2005 Apr 1;14(3):248–52.
19. Waller DK, Spears WD, Gu Y, Cunningham GC. Assessing number-specific error in the recall of onset of last menstrual period. *Paediatr Perinat Ep*. 2000;14(3):263–7.
20. Price JT, Winston J, Vwalika B, Cole SR, Stoner MCD, Lubeya MK, et al. Quantifying bias between reported last menstrual period and ultrasonography estimates of gestational age in Lusaka, Zambia*. IJGO.* 2019;144(1):9–15.
21. Savitz DA, Terry JW, Dole N, Thorp JM, Siega-Riz AM, Herring AH. Comparison of pregnancy dating by last menstrual period, ultrasound scanning, and their combination. *Am J Obstet Gynecol*. 2002 Dec 1;187(6):1660–6.
22. Hoffman CS, Messer LC, Mendola P, Savitz DA, Herring AH, Hartmann KE. Comparison of gestational age at birth based on last menstrual period and ultrasound during the first trimester. *Paediatr Perinat Ep*. 2008;22(6):587–96.
23. Oboro V, Bello T, Oyeniran A. First trimester sonographic dating formula for the Nigerian obstetric population. *West Afr J Radiol*. 2012;19(1):1–4.
24. Wani RT. Socioeconomic status scales-modified Kuppuswamy and Udai Pareekh's scale updated for 2019. *J Family Med Prim Care*. 2019 Jun 1;8(6):1846.

TABLES

Table 1. Baseline characteristics of the participants included in the *main* dataset (N_o=2,562) for the comparison of different methods of dating.

Socio-demographic characteristics	**Median (IQR) or N (%) or Mean ± SD**
Age (year)	23 (21-26)
GA at enrolment by LMP (weeks)	11.31±2.67
GA at enrolment by USG-Hadlock (weeks)	10.87±2.28
BMI at enrolment into the cohort[a]	
Underweight	*27.20%*
Normal weight	*59.93%*
Obese	*9.09%*
Overweight	*1.66%*
Haemoglobin (g/dL)	8.8 (8.2-9.2)
Height (cm)	153 (149.2-156.8)
Socioeconomic status[b]	
Upper class	*0.66%*
Upper middle class	*15.40%*
Lower middle class	*33.98%*
Upper lower class	*48.96%*
Lower class	*0.43%*
Undetermined	*0.57%*
Parity (number)	
0	*49.53%*
1	*33.55%*
2	*12.60%*
3	*3.34%*
4	*0.74%*
5	*0.14%*
Level of education	

Illiterate	*21.58%*
Literate or primary school	*8.63%*
Middle school	*15.09%*
High school	*18.61%*
Post high school diploma	*20.89%*
Graduate	*12.23%*
Post-graduate	*2.94%*
Occupation	
Unemployed	*93.48%*
Unskilled worker	*3.34%*
Semi-skilled worker	*0.97%*
Skilled worker	*1.40%*
Clerk, shop, farm owner	*0.17%*
Semi-professional	*0.26%*
Professional	*0.34%*
Religion	
Hindu	*92.14%*
Muslim	*6.60%*
Sikh	*0.40%*
Christian	*0.74%*
Buddhist	*0.0 %*
More than one religion	*0.09%*
Fuel used for cooking[c]	
Biomass fuel	*7.86%*
Clean fuel[d]	*92.14%*
Source of drinking water	
Safe water[e]	*49.80%*
Unsafe water	*50.20%*
Second-hand tobacco smoke	

Exposed	*19.23%*
Unexposed	*80.57%*
Undetermined	*0.20%*
History of any chronic illnesses[f]	
Absent	*99.03%*
Present	*0.97%*
History of hypertensive disease of pregnancy	
Absent	*99.57%*
Present	*0.43%*

Abbreviations: BMI, body mass index; IQR, interquartile range; GA, Gestational age; LMP, Last mensural period

[a] Pre-pregnancy BMI was calculated as weight (kg)/height (m)^2 from participants' weight and height measured at enrolment. ICMR-based categories[17] were defined as: underweight (< 18.5); normal (18.5-24.9); overweight (25.0-29.9); obese (≥ 30.0).

[b] Socioeconomic status was assessed using Modified Kuppuswamy's socioeconomic scale[24], calculated using education and occupation of the head of the family and monthly family income.

[c] Indoor air pollution: use of biomass fuel for cooking or presence of a smoker in the residential compound, as reported by the participant.

[d] Clean fuel includes liquefied petroleum gas and electricity.

[e] Safe water includes bottled water or piped water into the residence.

[f] Chronic illnesses include a history of hypertension, diabetes, cardiac disease and thyroid disorders.

Table 2: Pairwise comparison of mean difference and LOA between different first trimester dating formulae (Difference: Column formula - Row formula). Values shown in white are for the *main* dataset and values shown in grey are for the *test* dataset (see Methods for details).

Formula	**Hadlock**	**McLennan-Schluter**	**Robinson-Fleming**	**Sahota**	**Verburg**	**INTERGROWTH-21**	**Garbhini-1**
Hadlock		-0.159 (-0.397,0.079)	-0.173 (-0.363,0.016)	0.034 (-0.224,0.292)	0.037 (-0.408,0.481)	0.079 (-0.543,0.700)	0.296 (-0.229,0.822)
McLennan-Schluter	0.141 (-0.032,0.314)		-0.015 (-0.161,0.132)	0.193 (0.050,0.335)	0.195 (-0.104,0.495)	0.237 (-0.359,0.833)	0.455 (0.042,0.867)
Robinson-Fleming	0.165 (-0.019,0.349)	0.024 (-0.095,0.144)		0.207 (0.082,0.333)	0.210 (-0.097,0.516)	0.252 (-0.347,0.851)	0.469 (-0.021,0.959)
Sahota	-0.052 (-0.298,0.194)	-0.193 (-0.328,-0.057)	-0.217 (-0.346,-0.088)		0.002 (-0.196,0.201)	0.044 (-0.457,0.546)	0.262 (-0.124,0.648)
Verburg	-0.065 (-0.508,0.378)	-0.206 (-0.518,0.106)	-0.230 (-0.540,0.080)	-0.013 (-0.217,0.190)		0.042 (-0.446,0.530)	0.259 (-0.136,0.655)
INTERGROWTH-21	-0.118 (-0.786,0.550)	-0.259 (-0.901,0.383)	-0.283 (-0.943,0.377)	-0.066 (-0.619,0.487)	-0.053 (-0.591,0.485)		0.217 (-0.098,0.533)
Garbhini-1	-0.342 (-0.840,0.157)	-0.483 (-0.932,-0.034)	-0.507 (-1.015,0.001)	-0.29 (-0.692,0.112)	-0.277 (-0.704,0.150)	-0.224 (-0.494,0.046)	

Table 3: PTB rate estimated by various first trimester dating formulae used in this study.

Formula	**PTB rate (%)**	**95% CI**
LMP-based	13.30	(11.94, 14.77)
Hadlock	14.79	(13.36, 16.31)
McLennan-Schluter	16.72	(15.21, 18.31)
Robinson-Fleming	16.85	(15.34, 18.45)
Sahota	14.31	(12.90, 15.81)
Verburg	14.40	(12.98, 15.91)
INTERGROWTH-21	14.09	(12.69, 15.59)
Garbhini-1	12.12	(10.81, 13.53)

FIGURES

Figure 1: Outline of the data selection process for different datasets for the *main* and *test* datasets. Coloured boxes indicate the datasets used in the analysis. The names of each of the dataset are indicated below the box. Exclusion criteria for each step are indicated. N_p indicates the number of participants included or excluded by that particular criterion and N_o indicates the number of unique observations derived from the participants in a dataset.

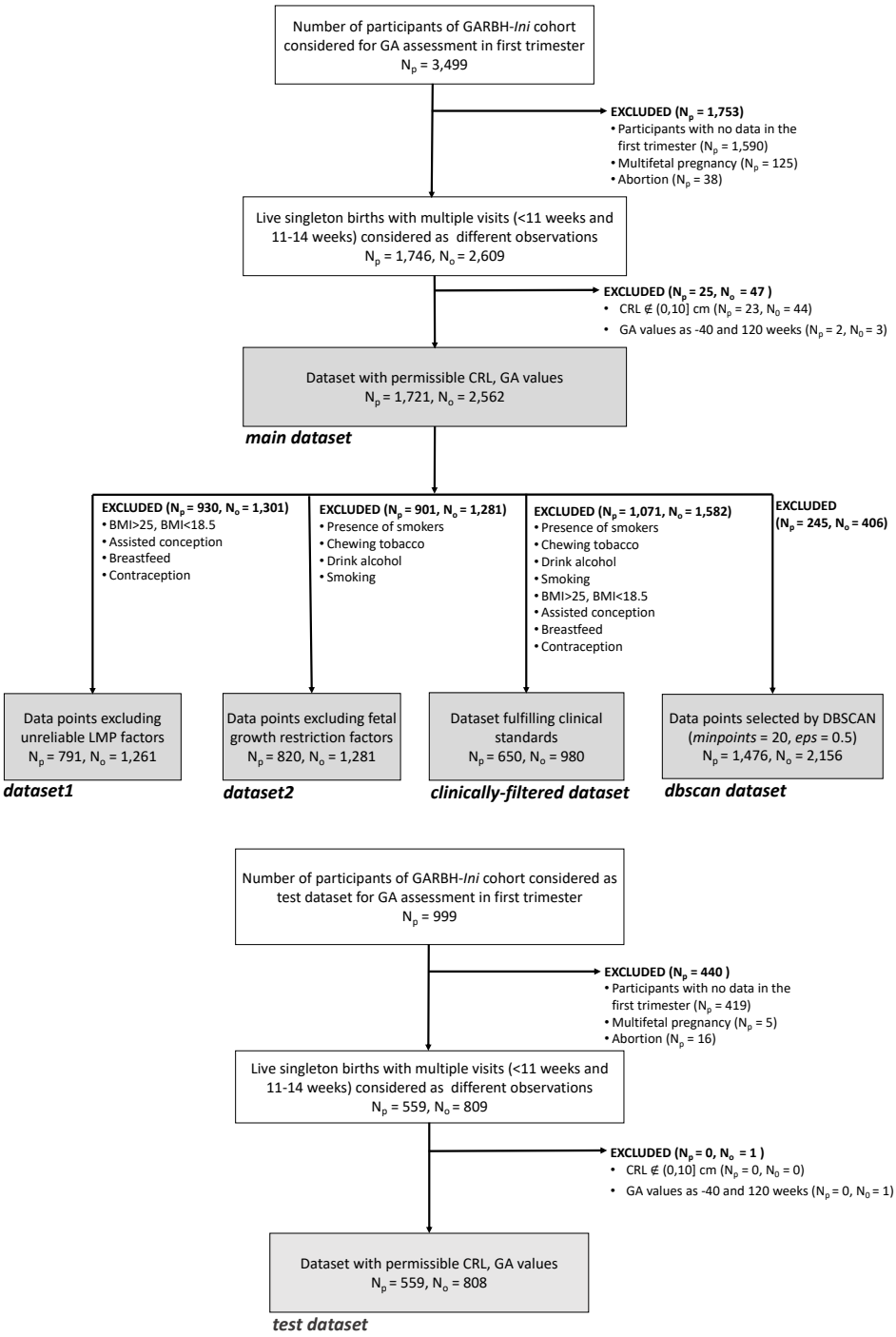

Figure 2: (A) Distribution of the difference between USG- and LMP-based GA. The x-axis is the difference between USG and LMP-based GA in weeks, and the y-axis is the number of observations. **(B)** BA analysis to evaluate the bias between USG and LMP-based GA. The x-axis is mean of Hadlock and LMP-based GA in weeks, and the y-axis is the difference between Hadlock and LMP-based GA in weeks. Regression line with 95% CI is shown.

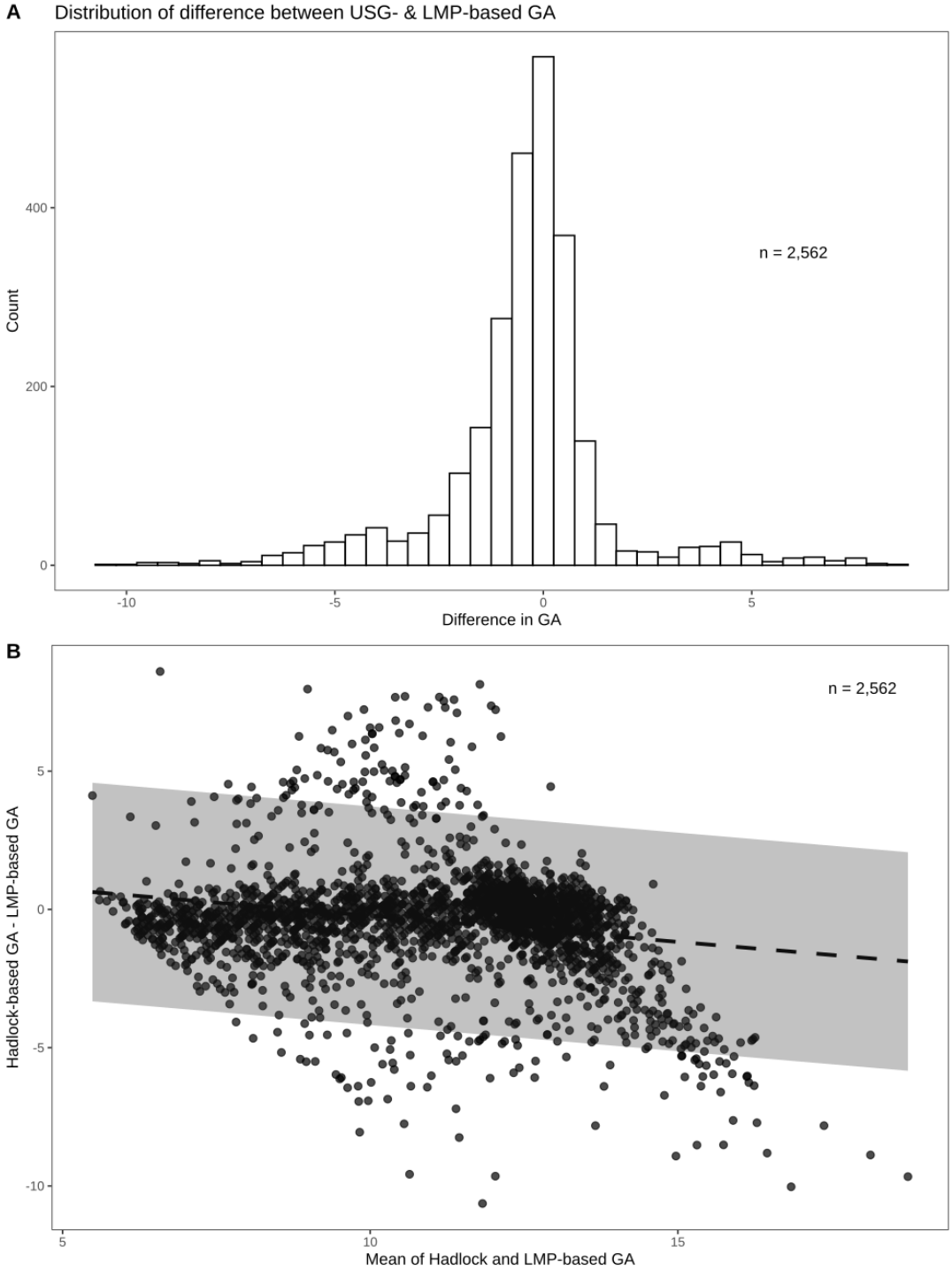
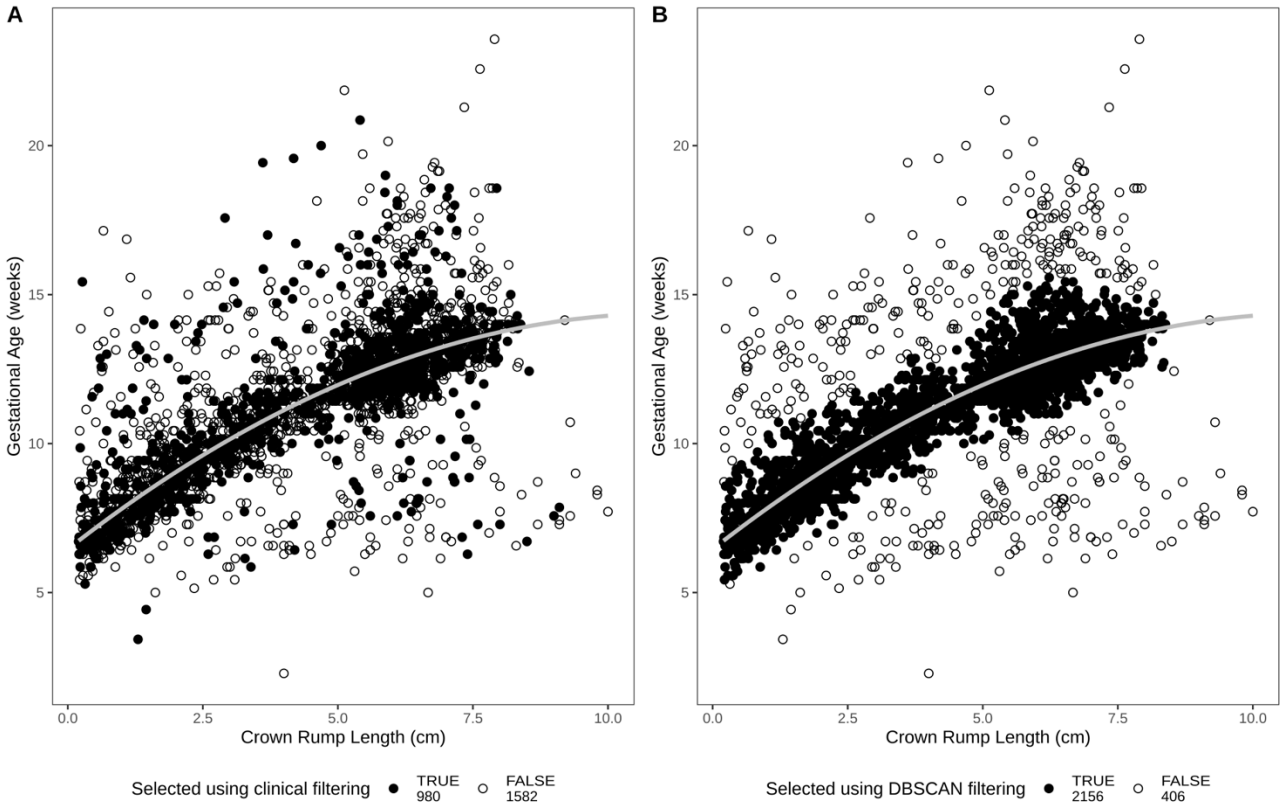
Figure 3: Comparison of data chosen to be reference data for the development of dating formula by **(A)** clinical and **(B)** data-driven (DBSCAN) approaches. The x-axis is CRL in cm, and the y-axis is GA in weeks (LMP-based are datapoints, Garbhini-1 is regression line). The data points selected (TRUE) after filtering are coloured black and points not selected (FALSE) are white.