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

Running title: Garbhini-1, gestational age formula for India

Ramya Vijayram[1,2,*], Nikhita Damaraju[1,2,*], Ashley Xavier[1,2,*], Bapu Koundinya Desiraju[3,4], Ramachandran Thiruvengadam[3,4], Sumit Misra[3,4], Shilpa Chopra[3,4], Ashok Khurana[5], Nitya Wadhwa[3,4], GARBH-Ini[4], Raghunathan Rengaswamy[2,6,7], Himanshu Sinha[1,2,7], 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] The Ultrasound Lab, Defence Colony, New Delhi, India

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

[7] 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 cohort 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

SYNOPSIS

Study question

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

What is already known

- Several first trimester GA estimation formulae have been published based on different population globally.
- In India, CRL-based Hadlock's formula, based on a US population, is primarily used for GA estimation.

What this study adds

- We have developed an Indian population-specific formula (Garbhini-1) for GA estimation in the first trimester.
- Garbhini-1 performs comparably to other published formulae in estimating GA with the highest balanced accuracy in classifying PTB.
- Other clinical and socioeconomic features do not improve the accuracy of the first trimester dating.
- Our results reinforce the need to develop population-specific GA formulae.

ABSTRACT

Background

Different formulae have been developed globally for estimation of gestational age (GA) by ultrasonography in the first trimester of pregnancy. In this study, we develop an Indian population-specific dating formula and compare its performance with 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 GARBH-Ini, an ongoing pregnancy cohort of North Indian women to study PTB.

Methods

Comparisons between ultrasonography-Hadlock and last menstrual period (LMP) based dating methods were made by studying the distribution of their differences by Bland-Altman analysis. Using data driven approaches, we removed data outliers more efficiently than by applying clinical parameters. We applied advanced machine learning algorithms to identify relevant features for GA estimation and developed an Indian population-specific formula (Garbhini-1) for the first trimester. PTB rates of Garbhini-1 and other formulae were compared by estimating sensitivity and accuracy.

Results

Performance of Garbhini-1 formula, a non-linear function of crown-rump length (CRL), was equivalent to published formulae for estimation of first trimester GA (limits of agreement, -0.46,0.96 weeks). We found that CRL was the most important parameter in estimating GA and no other clinical or socioeconomic covariates contributed to GA estimation. The estimated PTB rate across all the formulae including LMP ranged 11.54-16.50% with Garbhini-1 estimating the least rate with highest sensitivity and accuracy. While LMP-based method overestimated GA by three days compared to USG-Hadlock formula; at an individual level, these methods had less than 50% agreement in classification of PTB.

Conclusions

An accurate estimation of GA is crucial for management of PTB. Garbhini-1, the first such formula developed in an Indian setting, estimates PTB rates with higher accuracy especially when compared to commonly used Hadlock formula. Our results reinforce the need to develop population-specific gestational age formulae.

KEYWORDS

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

Word count: 3430

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, ascertaining the 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 pose 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 get an accurate estimation of the rate of PTB in a specific population. 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] (eTable 1). Finally, we quantified the implications of the choice of dating methods on PTB rates in our study population.

2. METHODS

2.1. Study design

GARBH-Ini is a collaborative program, initiated by Translational Health Science and Technology Institute, Faridabad with partners from Regional Centre of Biotechnology, Faridabad; National Institute of Biomedical Genomics, Kalyani; Civil Hospital, Gurugram; Safdarjung hospital, New Delhi. The 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 objective of the cohort study is to develop an effective risk stratification that facilitates timely referral for women at high risk of PTB, 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 clinical and socioeconomic variables are collected across pregnancy to help stratify women into defined risk groups for PTB.

2.2. 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.

2.3. Sampling strategy and participant datasets derived for the study

The samples for this analysis 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 and singleton pregnancy which advanced beyond 20 weeks of gestation, i.e. the pregnancy did not end in a spontaneous abortion. If more than one scan was performed <14 weeks, data from both the scans were included as unique observations (N_o). Therefore, 1,721 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* (Figure 1). This 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 in order 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 999 participants enrolled beyond 3,499 in this cohort (Figure 1). By applying identical processing steps as described for the *MAIN DATASET*, the *TEST DATASET* was obtained (N_o=808 from N_p=559; Figure 1).

2.4. 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, USA) 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, study nurses documented the clinical and socio-demographic characteristics[13].

2.5. 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 from N_p=650; Figure 1, eTable 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 (eTable 3). The resulting dataset that retained reliable data points for the analysis was termed as the *DBSCAN DATASET* (N_o=2,156 from N_p=1,476; Figure 1). Similarly, from the *TEST DATASET*, clinically-filtered and dbscan datasets were derived using identical filtering steps as described for the *MAIN DATASET* except that in this case, epsilon value was 0.6.

Development of a first trimester dating formula was done by fitting linear, quadratic and cubic regression models of GA (weeks) as a function of CRL (cm) on *CLINICALLY-FILTERED* and *DBSCAN* datasets. The performance of the chosen formula was validated in the *TEST DATASET*.

In addition to CRL as a primary indicator, a list of 282 candidate variables were explored by feature selection methods on the *DBSCAN DATASET* to identify other variables which may be predictive of GA during the first trimester. These methods helped to find uncorrelated, non-redundant features that might improve the accuracy of GA prediction (eTable 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 predictors of GA. A union of these features (eTable 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 the *TEST DATASET*.

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

As an additional objective, we estimated the effect of factors that could contribute to the discrepancy between assessment of gestation by LMP and ultrasound. This may be because of an unreliable LMP or foetal growth restriction. In order to quantify the contribution of unreliable LMP, a sub-dataset was derived (labelled as *Dataset1;* N_o=1,261 from N_p=791) from the *MAIN DATASET* (Figure 1) by excluding participants with use of contraceptives a month prior to the pregnancy, assisted conception, enrolment BMI beyond the normal range, and breastfeeding in the two months before conception.

Another analysis was done to estimate the contribution of foetal growth restriction to this discrepancy by creating *Dataset2* (N_o=1,281 from N_p=820), filtered from the *MAIN DATASET* by excluding participants with a known risk of foetal growth restriction. The factors that are known to affect CRL measurements included active and passive smoking, consumption of tobacco and alcohol, and enrolment BMI beyond the normal range.

We calculated the difference between LMP- and USG-based GA for each participant at 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. This analysis was repeated on *Dataset1* and *Dataset2*. The PTB rates with LMP and USG-based methods were reported per 100 live births with 95%CI. We compared different USG-based formulae using correlation analysis. The bias between different formulae was evaluated by BA analysis, and pairwise mean difference and LOA were reported.

2.7. Development of 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 wherever appropriate.

3. RESULTS

3.1. Description of participants included in the study

The median age of pregnant women enrolled in the cohort was 23.0 years (IQR 21.0,26.0), with 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 (98.20%) participants were from the middle or lower socioeconomic strata, and 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) indicating that the LMP-based method overestimated GA by nearly three days. The LOA determined by BA analysis was -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 marginally when tested on *Dataset1* (LOA -4.22, 3.28) or on *Dataset2*, (LOA -4.13,3.21). The wide LOA that persisted (despite ensuring reliable LMP (*Dataset1*) and standardised CRL measurements (*Dataset2*) represents the residual imprecision due to unknown factors in the estimation of GA.

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

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 (Figure 1) 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 with CRL as the only predictor was done using linear, quadratic and cubic regression to identify the best predictive model (eFigure 1). The *CLINICALLY-FILTERED DATASET* could not be used for model building because the estimated GA declined as the CRL increased beyond a specific limit which is biologically implausible. The DBSCAN approach provided a more accurate dataset (i.e. no artefacts as observed in the *CLINICALLY-FILTERED DATASET*) with lesser outliers. We, therefore, used *DBSCAN DATASET* for building dating models. Comparison among various dating models showed that the best regression coefficient (R^2) was for quadratic regression (R^2=0.86) with no further improvement when tested for cubic regression (eTable 6). This provided the basis for using the following quadratic formula 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 (eFigure 2, eTable 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. We tested the performance of the published formulae (eTable 1) and Garbhini-1 formula independently on the *TEST DATASET* (Figure 1). It was observed that Garbhini-1 had a R^2 value of 0.58 (eTable 7, eFigure 3). When tested on clinically-filtered and dbscan datasets derived from *TEST DATASET* (see Methods), the R^2 value were 0.58 and 0.90, respectively. All other formulae performed identically to Garbhini-1 on the *TEST DATASET* (eTable 7). Furthermore, all possible pairwise BA analysis of these formulae (including Garbhini-1) 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.

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

The PTB rates estimated using different methods ranged between 11.5 and 16.5% with Garbhini-1 estimating the least (11.5%; CI 9.95, 13.28), LMP (14.0%; CI 12.25, 15.86), Hadlock (14.5%; CI 12.77, 16.43), and Robinson-Fleming formula the highest (16.5%; CI 14.64, 18.49). Amongst all pairwise comparisons performed, the differences in PTB rates estimated by Garbhini-1 in comparison with Robinson-Fleming or McLennan-Schluter were statistically significant (Fisher's exact test with Bonferroni correction for $p<0.05$, eTable 8). Furthermore, among all the formulae tested, Garbhini-1 formula had the highest sensitivity and balanced accuracy (eTable 9).

When these methods were used to determine PTB at an individual level, the Jaccard similarity coefficient (a statistic used for gauging the similarity and diversity of sample sets) ranged between 0.49-0.98 (Table 3). Interestingly, even though the two most used methods of dating, LMP and USG-Hadlock had similar PTB rates (14.0 and 14.5%, respectively) at the population-level, the Jaccard similarity coefficient was only 0.49 suggesting a poor agreement between the methods at an individual-level (Figure 2C, Table 3).

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, develop a population-specific dating model for the first trimester and study the differences in PTB rate estimation using these formulae. Our findings show that the LMP-based method overestimates GA by three days compared to the USG (Hadlock) method. While this bias does not have impact at the population level with similar overall PTB rates determined by both methods, interestingly, there is less than 50% agreement between these methods on who are classified as preterm at an individual level. This is consistent with the pattern observed in a recent study from a Zambian cohort[20]. The Hadlock formula for USG-based estimation of GA was developed on a Caucasian population and has been used for several decades globally. We developed and tested population-specific dating formula to estimate GA in an Indian setting. The CRL-based Garbhini-1 formula performed the best and addition of other clinical and sociodemographic predictors identified from machine learning tools did not improve the performance of CRL-based

Garbhini-1 formula. While most of the dating formulae estimated similar PTB rates, Garbhini-1 formula estimated the lowest PTB rate and had the best sensitivity to determine preterm birth.

4.2. Strengths of the study

The Garbhini-1 formula developed from Indian population overcomes the poor representativeness of existing dating formulae. Using advanced data-driven approaches we evaluated multiple combinations of various clinical and sociodemographic parameters for estimation of gestational age. We conclusively show that CRL is the sufficient parameter for first trimester dating of pregnancy and addition of other clinical or social parameters do not improve the performance of the dating model. Further, to build Garbhini-1 formula, we used a data-driven approach to remove outliers which retained more observations for building the model than would have been possible if clinical criteria-based method had been used for identifying the reference standard. Another important strength of our study is the standardised measurement of CRL. This reduces the imprecision to the minimum and makes USG-based estimation of gestational age 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 relatively incomplete as we had residual imprecision, which was not accounted for by the clinical criteria. We tried to overcome this limitation by using data-driven approaches to improve precision.

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 overestimation of GA by LMP-based method as seen in our cohort has been reported in other populations from Africa and North America[20,21]. But the magnitude of overestimation varies as seen in studies done earlier[20-22]. These differences could be attributed to the precision and accuracy with which

the LMP was recalled by the participants of these cohorts. In our study, the bias in LMP-based dating was not reflected in the population-level PTB rates; however, at an individual level, LMP and USG-Hadlock had less than 50% agreement in the classification of PTB. Such a large discordance is concerning as the clinical decisions during the early neonatal period largely depends on the GA at birth. Further, any clinical and epidemiological research studying the risk factors and complications of PTB will be influenced by the choice of dating method.

Garbhini-1 formula was developed from the Indian population based on first trimester CRL can be interchangeably used with Hadlock, INTERGROWTH-21st, Verburg and Sahota but not with McLennan-Schluter and Robinson-Fleming formulae in our population. The higher sensitivity of Garbhini-1 formula to detect PTB in our study population is encouraging but should be externally validated in other populations within the country before it can be recommended for application. The comparable performance of Garbhini-1 formula with most of those developed globally may be explained by the negligible differences in the early foetal growth as measured by CRL across populations. It would be useful to evaluate the performance of population-specific formulae for second and third trimesters of gestation as ethnic differences in foetal growth might manifest during this period.

CONCLUSIONS

LMP overestimates GA by three days as compared to USG-Hadlock method, and only half of the preterm birth were classified correctly by both these methods. 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 sensitivity than the most used Hadlock formula in estimating the PTB rate. Our results reinforce the need to develop population-specific GA formulae. These results need to be further validated in subsequent multi-ethnic cohorts before it can be applied for wider use.

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 foetal 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 foetal 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 foetal 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.

Acknowledgements

We thank all the participants of GARBH-Ini study. We thank Karthik Raman and Nirav Bhatt from Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences and Initiative of Biological Systems Engineering, IIT Madras, and Gagandeep Kang from Translational Health Science and Technology Institute for their valuable suggestions.

The members of GARBH-Ini study include the following: Translational Health Science and Technology Institute, NCR Biotech Cluster, Faridabad, Delhi NCR, India (Shinjini Bhatnagar, Vineeta Bal, Bhabatosh Das, Mahadev Dash, Bapu Koundinya Desiraju, Pallavi Kshetrapal, Sumit Misra, Balakrish G Nair, Uma Chandra Mouli Natchu, Satyajit Rath, Kanika Sachdeva, Shailaja Sopory, Amanpreet Singh, Dharmendra Sharma, Ramachandran Thiruvengadam, Nitya Wadhwa); National Institute of Biomedical Genomics, Kalyani, West Bengal, India (Arindam Maitra, Partha P Majumder); Regional Centre for Biotechnology, NCR Biotech Cluster, Faridabad, Delhi NCR, India (Tushar K Maiti, Dinakar M Salunke); Clinical Development Services Agency, Translational Health Science and Technology Institute, NCR-Biotech Cluster, Faridabad, Delhi NCR, India (Monika Bahl, Shubhra Bansal); Gurugram Civil Hospital, Haryana, India (Umesh Mehta, Sunita Sharma, Alka Singh, Brahmdeep Sindhu); Safdarjung Hospital, New Delhi, India (Sugandha Arya, Rekha Bharti, Harish Chellani, Pratima Mittal); Maulana Azad Medical College, New Delhi, India (Siddarth Ramji, Reva Tripathi, Anju Garg), The Ultrasound Lab, Defence Colony, New Delhi, India (Ashok Khurana); Hamdard Institute of Medical Sciences and Research, Jamia Hamdard University, New Delhi, India (Reva Tripathi); All India Institute of Medical Sciences, New Delhi, India (Smriti Hari, Yashdeep Gupta, Nikhil Tandon); Government of Haryana, India (Rakesh Gupta); International Centre For Genetic Engineering and Biotechnology, New Delhi, India (Dinakar M Salunke); Indian Institute of Science Education and Research, Pune, Maharashtra, India (Vineeta Bal).

Funding

This study was funded by an alumni endowment from Dr. Prakash Arunachalam to the 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).

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, AK, NW, SM, SC developed and implemented the clinical data collection methods and data management in GARBH-Ini cohort, 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.

CODE AVAILABILITY

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

FIGURE LEGENDS

Figure 1: Outline of the data selection process for different datasets – **(A)** *MAIN DATASET* and **(B)** *TEST DATASET*. 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. [a] Biologically implausible CRL values (either less than 0 or more than 10 cm) for the first trimester were excluded, [b] Biologically implausible GA values (either less than 0 and more than 45 weeks) were excluded.

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. **(C)** Comparison of individual-level classification of preterm birth by Hadlock- and LMP-based methods. Green (term for both), red (preterm for both), blue (term for LMP but preterm for Hadlock) and purple (term for Hadlock but preterm for LMP).

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.

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.

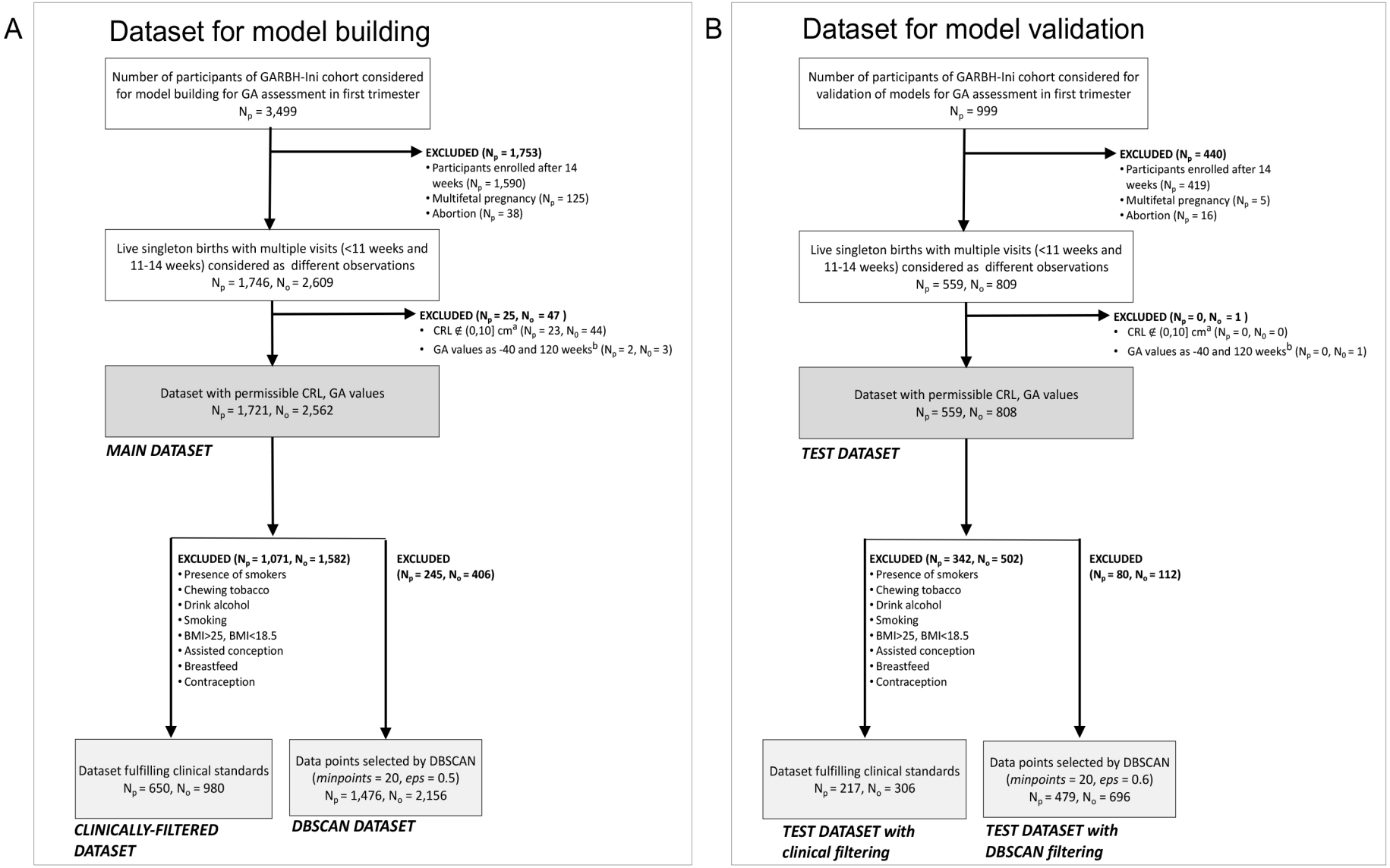
[f] Chronic illnesses include a history of hypertension, diabetes, cardiac disease and thyroid disorders.

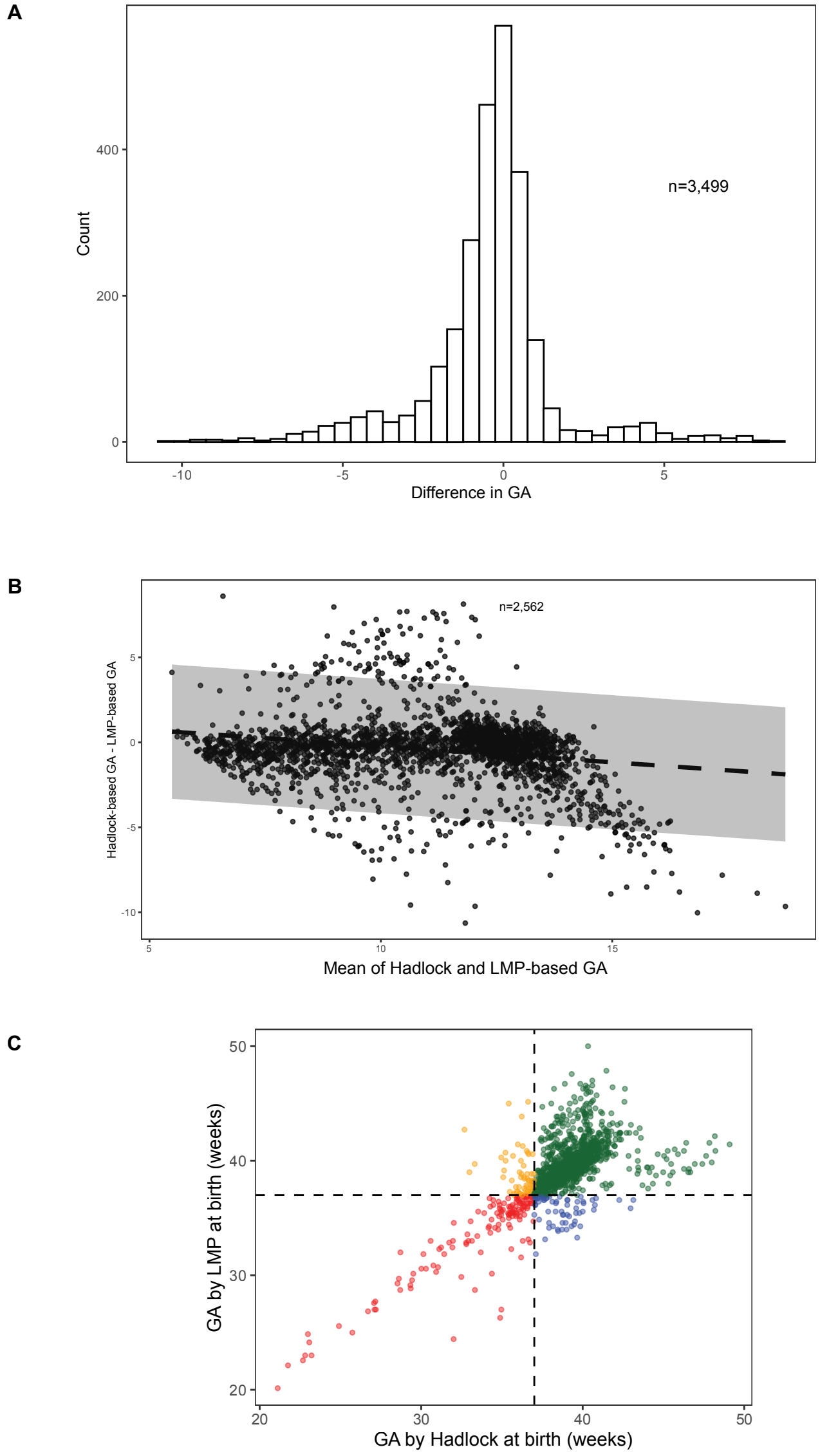
Table 2: Pairwise comparison of mean difference and LOA between different first trimester dating formulae (Difference: Column formula - Row formula). Values shown in white are for the *MAIN DATASET* and values shown in grey are for the *TEST DATASET* (see Methods for details).

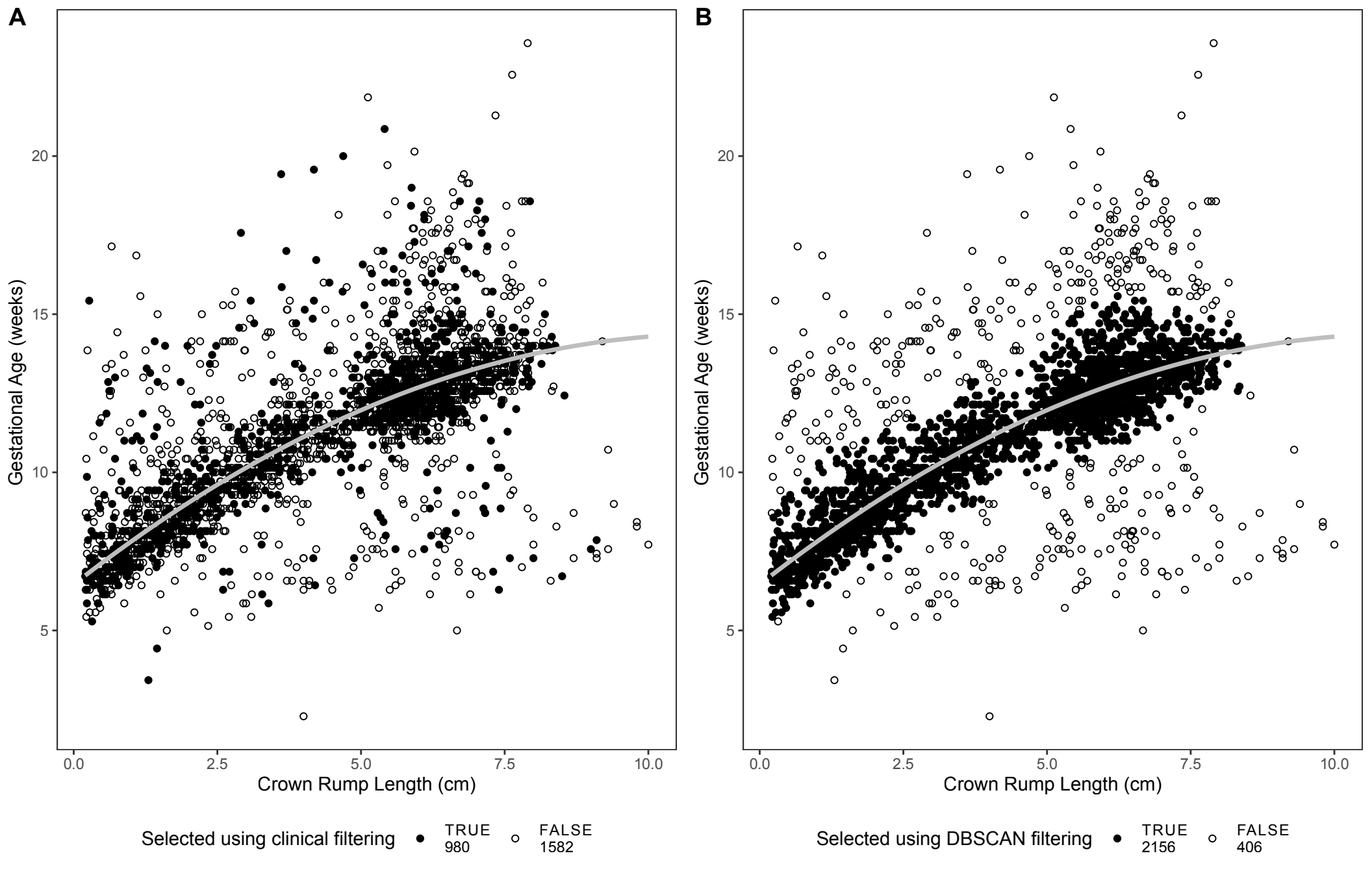
Formula	**Hadlock**	**McLennan-Schluter**	**Robinson-Fleming**	**Sahota**	**Verburg**	**INTERGROWTH-21**	**Garbhini-1**
Hadlock		-0.16 (-0.40,0.079)	-0.17 (-0.36,0.016)	0.034 (-0.22,0.29)	0.037 (-0.41,0.48)	0.079 (-0.54,0.70)	0.30 (-0.23,0.82)
McLennan-Schluter	0.14 (-0.032,0.31)		-0.015 (-0.16,0.13)	0.19 (0.05,0.34)	0.20 (-0.10,0.50)	0.24 (-0.36,0.83)	0.46 (0.042,0.87)
Robinson-Fleming	0.17 (-0.019,0.35)	0.024 (-0.095,0.14)		0.21 (0.082,0.33)	0.21 (-0.097,0.52)	0.25 (-0.35,0.85)	0.47 (-0.021,0.96)
Sahota	-0.052 (-0.30,0.19)	-0.19 (-0.33,-0.057)	-0.22 (-0.35,-0.088)		0.002 (-0.20,0.20)	0.044 (-0.46,0.55)	0.26 (-0.12,0.65)
Verburg	-0.065 (-0.51,0.39)	-0.21 (-0.52,0.11)	-0.23 (-0.54,0.08)	-0.013 (-0.22,0.19)		0.042 (-0.45,0.53)	0.26 (-0.14,0.66)
INTERGROWTH-21	-0.12 (-0.79,0.55)	-0.26 (-0.90,0.38)	-0.28 (-0.94,0.38)	-0.066 (-0.62,0.49)	-0.053 (-0.59,0.49)		0.22 (-0.098,0.53)
Garbhini-1	-0.34 (-0.84,0.16)	-0.48 (-0.93,-0.034)	-0.51 (-1.02,0.001)	-0.29 (-0.69,0.11)	-0.28 (-0.70,0.15)	-0.22 (-0.49,0.046)	

Table 3: The Jaccard similarity coefficient of PTB prediction between each pair of the method.

Formula	**LMP**	**Hadlock**	**McLennan-Schluter**	**Robinson-Fleming**	**Sahota**	**Verburg**	**INTERGROWTH-21**	**Garbhini-1**
LMP	1.00	0.49	0.50	0.50	0.52	0.53	0.53	0.53
Hadlock		1.00	0.90	0.88	0.88	0.81	0.80	0.79
McLennan-Schluter			1.00	0.98	0.83	0.82	0.80	0.71
Robinson-Fleming				1.00	0.82	0.81	0.79	0.70
Sahota					1.00	0.92	0.89	0.85
Verburg						1.00	0.87	0.86
INTERGROWTH-21							1.00	0.89
Garbhini-1								1.00


A
Dataset for model building
Number of participants of GARBH-Ini cohort considered for model building for GA assessment in first trimester
N_p = 3,499
EXCLUDED (N_p = 1,753)
• Participants enrolled after 14 weeks (N_p = 1,590)
• Multifetal pregnancy (N_p = 125)
• Abortion (N_p = 38)
Live singleton births with multiple visits (<11 weeks and 11-14 weeks) considered as different observations
N_p = 1,746, N_o = 2,609
EXCLUDED (N_p = 25, N_o = 47)
• CRL ∉ (0,10] cm[a] (N_p = 23, N_0 = 44)
• GA values as -40 and 120 weeks[b] (N_p = 2, N_0 = 3)
Dataset with permissible CRL, GA values
N_p = 1,721, N_o = 2,562
MAIN DATASET
EXCLUDED (N_p = 1,071, N_o = 1,582)
• Presence of smokers
• Chewing tobacco
• Drink alcohol
• Smoking
• BMI>25, BMI<18.5
• Assisted conception
• Breastfeed
• Contraception
EXCLUDED
(N_p = 245, N_o = 406)
Dataset fulfilling clinical standards
N_p = 650, N_o = 980
CLINICALLY-FILTERED DATASET
Data points selected by DBSCAN
(minpoints = 20, eps = 0.5)
N_p = 1,476, N_o = 2,156
DBSCAN DATASET
B
Dataset for model validation
Number of participants of GARBH-Ini cohort considered for validation of models for GA assessment in first trimester
N_p = 999
EXCLUDED (N_p = 440)
• Participants enrolled after 14 weeks (N_p = 419)
• Multifetal pregnancy (N_p = 5)
• Abortion (N_p = 16)
Live singleton births with multiple visits (<11 weeks and 11-14 weeks) considered as different observations
N_p = 559, N_o = 809
EXCLUDED (N_p = 0, N_o = 1)
• CRL ∉ (0,10] cm[a] (N_p = 0, N_0 = 0)
• GA values as -40 and 120 weeks[b] (N_p = 0, N_0 = 1)
Dataset with permissible CRL, GA values
N_p = 559, N_o = 808
TEST DATASET
EXCLUDED (N_p = 342, N_o = 502)
• Presence of smokers
• Chewing tobacco
• Drink alcohol
• Smoking
• BMI>25, BMI<18.5
• Assisted conception
• Breastfeed
• Contraception
EXCLUDED
(N_p = 80, N_o = 112)
Dataset fulfilling clinical standards
N_p = 217, N_o = 306
TEST DATASET with clinical filtering
Data points selected by DBSCAN
(minpoints = 20, eps = 0.6)
N_p = 479, N_o = 696
TEST DATASET with DBSCAN filtering


A
Count
400
200
0
n=3,499
-10
-5
0
5
Difference in GA
B
Hadlock-based GA - LMP-based GA
5
0
-5
-10
n=2,562
5
10
15
Mean of Hadlock and LMP-based GA
C
GA by LMP at birth (weeks)
50
40
30
20
20
30
40
50
GA by Hadlock at birth (weeks)


A
B
Gestational Age (weeks)
Crown Rump Length (cm)
20
15
10
5
0.0
2.5
5.0
7.5
10.0
Selected using clinical filtering
TRUE
980
FALSE
1582
Selected using DBSCAN filtering
TRUE
2156
FALSE
406