Data-driven identification of novel preterm birth phenotypes in Indian population

A THESIS

Submitted by

ADITI SADHU

for the award of the degree

of

Dual Degree (BTech + MTech)

DEPARTMENT OF BIOTECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY MADRAS
CHENNAI-600036

JUNE 2021

THESIS CERTIFICATE

This is to certify that the thesis entitled **"Data-driven identification of novel preterm birth phenotypes in Indian population"** submitted by **Aditi Sadhu (BE16B014)** to the Indian Institute of Technology, Madras for the award of the degree of **Dual Degree (BTech+MTech)** is a bona fide record of research work carried out by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Himanshu Sinha
Associate Professor
Department of Biotechnology
Indian Institute of Technology Madras
Chennai – 600 036.

Place: Chennai

Date: June 2021

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Himanshu Sinha, for giving me an opportunity to pursue this project under him. I am grateful for the invaluable guidance, feedback, time and support he has given me over the past one year. He has always been readily available whenever I had questions and has been a great mentor to me.

I would also like to express my gratitude to the members from Translation Health Science and Technology Institute, Dr. Ramchandran Thiruvengadam and Dr. Koundinya Desiraju, for the valuable inputs.

I would also like to thank my friend, Abhinav Bhatnagar, who helped me whenever I faced difficulty while doing the project.

ABSTRACT

India contributes 25% of the overall global preterm related deaths, the highest worldwide. Of the 3.6 million annual preterm births, 300,000 die soon after birth. Keeping this in mind, a risk stratification based on multidimensional factors assessed during pregnancy is needed for prevention of preterm birth. Preterm birth has been difficult to study and prevent because of its complex syndromic nature- having multiple etiological factors requiring different preventive strategies. Also, a significant percentage of these births are difficult to associate to specific phenotypes.

In a bid to promote healthcare of high-risk pregnant mothers and reduce neonate mortality specifically for India, we will be applying data-driven approaches to identify novel preterm birth phenotypes for the Indian population. Using machine learning algorithms and statistical methods on a dataset having anthropometric, clinical and obstetric data of 4700 pregnant women (enrolled within 20 weeks of gestation and followed until delivery and postpartum), preterm birth features will be extracted and associated with birth outcomes.

This study aims to identify phenotypic subgroups of preterm births in GARBH-Ini so that they are treated differentially and better intervention practices can be employed specific to these subgroups.

Keywords: preterm births; phenotypic classification; India; clustering

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

USG	Ultrasonography
GA	Gestational Age
BW	Birth Weight
PPROM	Preterm Prelabor Rupture of Membranes
NICU	Neonatal Intensive Care Unit
GARBH-Ini	Interdisciplinary Group for Advanced Research on Birth Outcomes—DBT India Initiative
BPM	Beats per minute

CHAPTER 1

INTRODUCTION

1.1. Overview

Preterm birth is defined as a birth taking place before 37 weeks of gestation are completed. India sees a preterm birth rate of approximately 13%, that is, 3.6 million of the 27 million babies born annually. It is major reason for infant mortality and adverse neurocognitive, visual, and respiratory outcomes. However, it has been hard to study due to multiple etiological factors and a time-based definition. It is defined as a syndrome rather than a disease because of its large set of symptoms without having a definite cause. It involves interactions of biological, psychosocial and environmental factors. A classification system would be useful both in population surveillance and research. It may help in narrowing down to the causes, which will lead to targeted interventions for treatment purposes. It may also be useful in developing prediction tools for preterm births also.

Studying preterm birth has been challenging for decades because of its time-based definition. If the example of premature death (before 65 years of age) is taken, the deaths can be due to a large number of factors ranging from cancer to accidents. An etiological study of such an entity would be too vast and can be inconclusive. Preterm birth syndrome is a similar entity.

A classification system may be made on the basis of either causes or phenotypes. Since the cause is often not known with confidence, an ideal classification system for preterm births should be made on the basis of clinical phenotypes rather than causes. Existing phenotypic classifications include a classification on the basis of:

- GA: early preterm (<32 weeks) or late preterm (>=32 weeks)
- Clinical presentations: spontaneous, preterm premature rupture of membranes (PPROM), indicated

- Pathology: infectious or stress-induced

However, such classifications do not look at preterm birth wholesomely and focus on only one aspect at a time. To cover multiple aspects revolving around preterm birth, a classification system can be defined using multiple features of the mother (eg, infections, short cervical length), the fetus (eg, abnormal amniotic fluid traits), placenta (eg, abruption, placenta previa), and the delivery presentation (eg, contractions, PPROM, advanced cervical dilation, bleeding or none).

For a reliable phenotypic classification system, the clinical data should constitute information on prior pregnancies, medical history, clinical information, obstetrics data, the reason for physician initiating a preterm delivery, a gross examination and microscopic evaluation of the placenta and, a pathology report for stillbirths and most importantly antepartum and intrapartum data.

1.2. Objective of the Work

To identify novel phenotypic classes of preterm birth for the Indian Population using data-driven techniques

2.4. Problem Definition and Approach

Currently, all preterm births are treated the same at the neonatal age, infancy and childhood, but that is not the best approach. The hypothesis here is that no two preterm babies at a certain gestational age are same in terms of their outcomes. They are different in the virtue of what happens in their delivery and how they are delivered. Hence, they shouldn't be treated at the same scale. The project aims to identify distinct groups of preterm births, particular to GARBH-Ini cohort, on the basis of groups of events that occur together and differential outcomes and understand how they differ

from other groups biologically. For example, one can say to clinician if a preterm baby falls in this phenotype of preterm, then the baby would have a defined set of outcomes as it grows in its neonatal and infancy, so perform interventions accordingly. So, the model will be preventive from baby's perspective.

Also, these groups behave differently in terms of outcomes, so if we may not be able to predict preterm births as a whole, accurate results may be obtained within the groups. Hence, from research's perspective, the study will assist in prediction models.

We will be solving the problem by defining 5 sets of characteristics for preterm births in GARBH-Ini cohort – maternal, fetal, placental, signs of initiation of parturition and pathway to delivery- as suggested in the conceptual framework. The data of preterm births that we will be using has been collected at the Gurgaon Civil Hospital. We will employ agglomerative hierarchical algorithm for clustering the preterm births to find phenotypic classes. These classes would be then analyzed, compared and validated.

1.4. Thesis Outline

The thesis is organized in 4 chapters. Chapter 1 introduces the reader to problem statement and briefs on the approach that will be applied. Chapter 2 is a literature review on related work, a reference study, data collection process in the GARBH-Ini cohort and explanation of the agglomerative hierarchical clustering. Chapter 3 has details on the methodology and experiments that were done in this study, along with a discussion on the results obtained. In Chapter 4 we summarize the work and discuss future work. Lastly an appendix is attached for supplementary information on the statistical methods used in the experiments.

CHAPTER 2

LITERATURE REVIEW

2.1. Conceptual Framework for Classification

A conceptual framework was put forward (Jose Villar, Aris T. Papageorghiou, Hannah E. Knight et al) that classifies preterm births based on these 5 phenotypic components:

1) maternal conditions before presentation for delivery
2) fetal conditions before presentation for delivery
3) placental pathologic conditions
4) signs of the initiation of parturition
5) the pathway to delivery

Maternal conditions include extrauterine infections, clinical chorioamnionitis, maternal trauma, worsening maternal diseases, uterine rupture, preeclampsia eclampsia. Intrauterine fetal death, intrauterine growth restriction, abnormal fetal heart rate, fetal infections, fetal anomalies, fetal anemia, polyhydramnios, oligohydramnios, multiple fetuses are some examples of fetal conditions. Placental conditions include histological chorioamnionitis, placental abruption, placenta previa and other placental abnormalities. Signs of initiation of parturition are cervical shortening, PPROM, regular contractions, cervical dilation, vaginal bleeding. The pathway to delivery is either spontaneously or caregiver initiated. Spontaneous labour can be either through regular contractions or augmented. Caregiver initiated deliveries can be clinically mandated, clinically discretionary, iatrogenic or no discernable reason, pregnancy termination, or no documented clinical indication.

The framework has been made such that relevant conditions can be a part of the classification without forcing any case into a predefined phenotype. Also, a class can have more than 1 characteristics from the set of 5 components.

2.2. Similar Research

A study on the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project performed classification aimed at identifying preterm birth phenotypes using the above conceptual framework as a priori (Fernando C. Barros et al, 2015). It was conducted at 8 different countries (Brazil, Italy, Oman, Kenya, UK, the USA, China and India,). Out of 60,058 total births, 53,871 were considered in this study because these had an ultrasound-based estimate of gestational age. 5828 of these were preterm births (10.8%). A birth (live or stillbirth) before 37 weeks of gestation was considered as preterm. They found a 12-cluster model to be congruous with the a priori framework. Their methodology will be taken as a reference in this study and a comparison of with their results will be done in the end.

Table 1 The 12 clusters of preterm births obtained in INTERGROWTH study

Note: Reprinted from "The Distribution of Clinical Phenotypes of Preterm Birth Syndrome: Implications for Prevention" by Fernando C. Barros at al, 2015, *Journal of the American Medical Association*

Cluster	No. (%)	Main Condition (%)	Most Frequent Associated Conditions (%)
1	1747 (30.0)	None	None
2	689 (11.8)	Pre-eclampsia (100)	Third-trimester bleeding and pre-eclampsia (72.6), extrauterine infection (28.6), and suspected IUGR (24.4)
3	607 (10.4)	Multiple births (100)	Extrauterine infection (21.9) and suspected IUGR (21.3)
4	450 (7.7)	Extrauterine infection (100)	Mid-pregnancy bleeding (20.4), chorioamnionitis (12.7), and severe maternal conditions (12.7)
5	443 (7.6)	Chorioamnionitis (100)	Multiple births (25.1), perinatal sepsis (14.7), and suspected IUGR (9.7)
6	362 (6.2)	Mid-/late-pregnancy bleeding (100)	Chorioamnionitis (21.8), perinatal sepsis (16.0), and multiple births (14.9)
7	337 (5.8)	Suspected IUGR (100)	Fetal distress (18.4), severe maternal conditions (18.4), and mid-/late-pregnancy bleeding (7.7)
8	319 (5.5)	Perinatal sepsis (68.0)	Congenital anomalies (41.4), multiple births (30.1), and fetal anemia (23.8)
9	280 (4.8)	Early bleeding (100)	Multiple births (27.9), extrauterine infection (25.0), and mid-/late-pregnancy bleeding (22.5)
10	213 (3.7)	Antepartum stillbirth (100)	Severe maternal condition (23.9), extrauterine infection (13.6), and mid-/late-pregnancy bleeding (13.1)
11	200 (3.4)	Fetal distress (100)	Severe maternal conditions (7.5), congenital anomalies (6.5), and chorioamnionitis (4.5)
12	181 (3.1)	Severe maternal conditions (100)	Multiple births (28.7), chorioamnionitis (24.3), and congenital anomalies (8.3)
All	5828 (100)		

2.3. GARBH-Ini Cohort

GARBH-Ini cohort aims to study preterm births on a multidimensional scale including clinical, epidemiologic, genomic, epigenomic, proteomic, and microbial correlates, identify molecular-risk markers, and develop a PTB prediction model. It collects longitudinal data clinical, epidemiologic, obstetrics, ultrasound as well as social-demographic variables of pregnant women at the Gurgaon Civil Hospital. (Shinjini Bhatnagar, Partha P. Majumder, Dinakar M. Salunke, 2018)

The cohort framework is designed such that the enrolment of a woman is completed in less than 20 weeks of gestation (calculated through last menstrual period), followed by an ultrasound scan in the same week to check for viable pregnancy. Follow-up antenatal visits are scheduled once in each of 11–14, 18–20, 26–28, and 30–32 weeks, and one unscheduled visit takes place in the postpartum period.

The first follow-up visit at 11-14 weeks has ultrasound for checking gestational age, fetal growth, fetal wellbeing, placental morphologies, and signs of genetic abnormalities. The abdominal ultrasound scans at 18–20 weeks check for fetal dysmorphic developmental anomalies. Cervical length is measured in transvaginal scans and the successive changes in the length are noted. The 3rd follow-up visit at 26-28 weeks involves collection of clinical data and biospecimens. The follow-up visit between 30 and 32 weeks involve Doppler scans checking for fetal wellbeing through measurements of uterine and umbilical artery pulse indices and amniotic fluid index. Also, fetal growth restriction is checked through the measurements of fetal middle cerebral artery and ductus venosus blood flow. Figure 2, by Shinjini Bhatnagar et al., summarises the cohort framework.

It was calculated that the number of enrolments required to achieve statistical significance of 5% with a statistical power of 80% to detect a 10% difference in preterm and term births is 400 women in each group. Assuming a safe preterm birth rate of 5%,

a sample size of 8,000 pregnant women would be required to be enrolled in the cohort. The data obtained and used in this study has 8473 enrolled pregnant women.

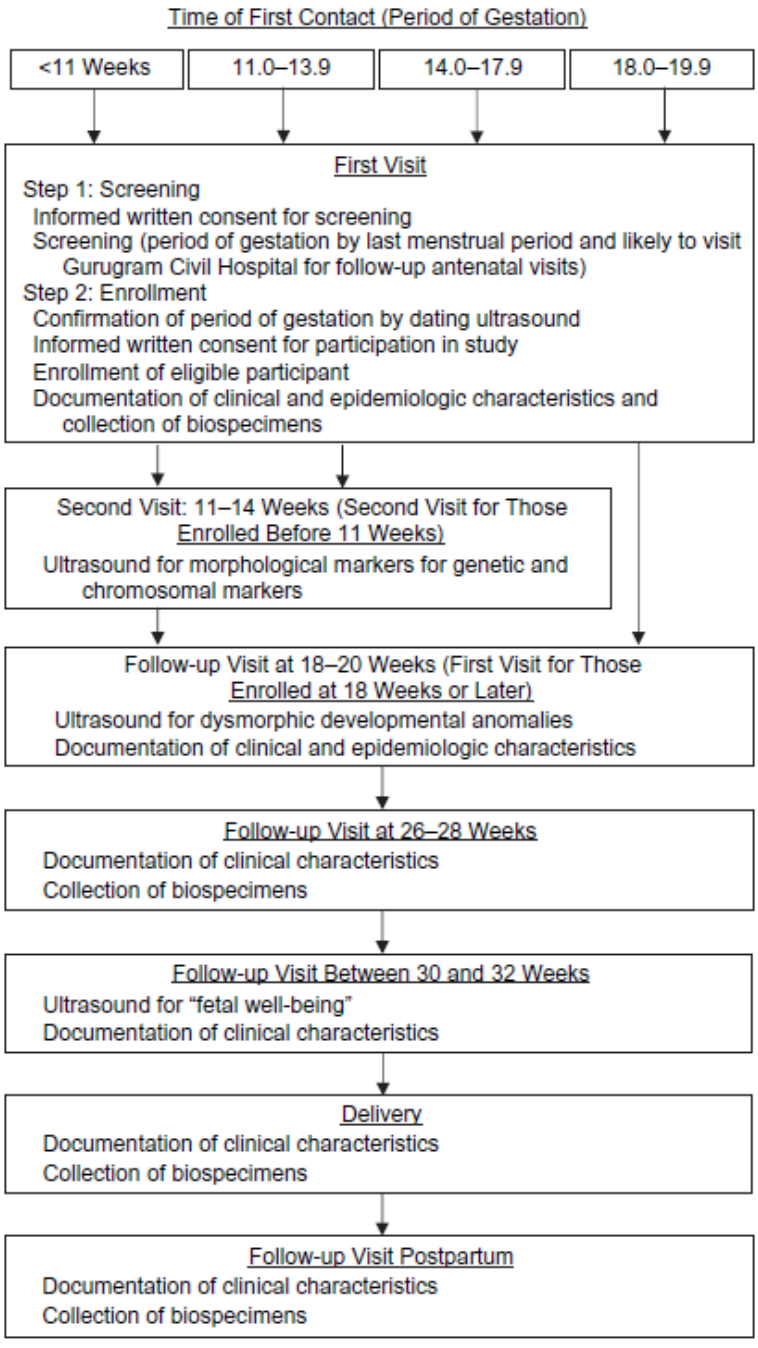


Figure 1 Participant enrollment and data collection in the GARBH-Ini cohort

Note: Reprinted from "A Pregnancy Cohort to Study Multidimensional Correlates of Preterm Birth in India: Study Design, Implementation, and Baseline Characteristics of the Participants" by Shinjini Bhatnagar et al, 2018, *American Journal of Epidemiology*

2.4. Hierarchical Clustering

Clustering is an unsupervised learning method which hypothesizes that data points in the same cluster behave similarly. (Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, 2008) Although flat clustering is efficient and conceptually simple, they return unstructured set of clusters and requires a predefined number of clusters. Hierarchical clustering returns a structured and informative set of clusters without the need of predefined number of clusters. However, it has lower efficiency but our sample size is relatively small so complexity is not an issue.

2.4.1. Agglomerative Hierarchical Clustering

Hierarchical clustering algorithms can be top-down or bottom-up. The top-down algorithms are called Divisive hierarchical clustering and the bottom-up are called agglomerative hierarchical clustering. Agglomerative algorithms are used more commonly. In this, each sample is considered a cluster initially, and then, pairs of clusters are merged to form a single cluster until a single cluster containing all the samples is obtained. It can be visualized using a dendrogram, where horizontal lines represent merging and vertical lines represent the dissimilarity.

CHAPTER 3

EXPERIMENTAL DETAILS

3.1. Methodology

3.1.1 Identifying variables from the data

The GARBH-Ini data obtained was thoroughly studied and all the conditions relevant for this study were found. Table 1 explains the variables used for this study and the conditions that have to be satisfied for each variable.

Table 2 Characteristics considered for this study

Characteristics	**Definition**
Maternal Characteristics	
Clinical chorioamnionitis	A bacterial infection affecting the membranes (chorion and amnion) and amniotic fluid. Confirmed by the presence of maternal fever (>100.4°F), rupture of membranes (water bag breakage) and any two of maternal tachycardia (>100 bpm), uterine tenderness, purulent amniotic fluid (foul smelling amniotic fluid), fetal tachycardia (>160 bpm) & maternal leukocytosis
Extrauterine infection during index pregnancy	Presence of at least one of typhoid, jaundice, bacterial vaginosis, upper respiratory tract infection, lower respiratory tract infection, periodontal infection, tuberculosis, chickenpox, malaria, dengue, urinary tract infections, TORCH infections, rheumatic fever, HIV, hepatitis or acute gastroenteritis. This was captured by subject presenting a physical temperature greater than 100.4 F and bacteremia/ malaria/ pyelonephritis.
Maternal trauma	Subject suffered from serious bodily injury or shock during the index pregnancy
Medical diseases	Presence of at least one of maternal cardiac, respiratory, renal diseases, hypothyroid, hyperthyroid, diabetes, asthma, anemia, malignancy, epilepsy, hypertension or cardiac disorders

Uterine rupture	Tearing of the muscular wall of the uterus. Confirmed by the presence of vaginal bleeding with abdominal pain, pallor, anemia and previous caesarean sections
Preeclampsia	Presence of elevated blood pressure (≥140/90 mm Hg) and damaged kidney (proteins in urine)
Eclampsia	A severe complication of preeclamsia. Confirmed by the presence of preeclampsia and convulsions
Fetal Characteristics	
Perinatal sepsis	Defined by positive neonatal septic screen, confirmed by presence of systemic infections, pneumonia and antibiotic prescription
Multiple pregnancy	Presence of more than 2 fetuses in the same pregnancy
Fetal anomaly	Presence of visceral or dysmorphic developmental anomaly identified at 18-20 weeks or at neonatal examinations
Polyhydramnios	Excess accumulation of amniotic fluid. Defined by an amniotic fluid index >95th centile (19cm for the GARBH-Ini dataset)
Oligohydramnios	Low amount of amniotic fluid. Defined by an amniotic fluid index <5th centile (7.97cm for the GARBH-Ini dataset)
Fetal distress	Presence of any one of meconium aspiration, perinatal asphyxia of stage 2 or 3, or provision of bag and mask ventilation
Birth weight	Weight of baby at birth (in grams)
Placental Characteristics	
Placental abruption	Separation of placenta before delivery. Confirmed by the presence of vaginal bleeding with abdominal pain and retroplacental clot at delivery
Placenta previa	Implantation of placenta spanning the internal orifice of the cervix uteri
Evidence of initiation of parturition	
Cervical length	Length of cervix in mm

Cervical shortening	Length of cervix <25 mm at 28-32 weeks USG
Cervical dilatation	Dilatation of uterine cervix on examination
PPROM	Rupture of membranes before the onset of labor at GA < 37 weeks Confirmed by the breaking of the water bag before labor pain
Bleeding	Any bleeding from uterus or cervix in the peripartum period
Pathway to delivery	
Caregiver initiated	This includes labor induced by the caregiver because of any of these reasons: Clinically mandated: due to immediate life-threatening risk to mother or fetus. Clinically discretionary: no imminent risk but for better outcomes No clinical indication: due to errors in GA estimation, maternal request, precious fetus
Spontaneous	Labor initiated spontaneously followed by either regular contractions or augmentation of labor

3.1.2 Preparing the data frame

For this study, only live preterm births have been considered, i.e. the neonate was alive when born. A filter was put for only known GA at delivery. As a result, the study comprised of 577 subjects of 7082 participants with known GA (8.15%). It is observed that the gestational age range among these live preterm births happened to be from 25 weeks. The preterm cases before 2 weeks were of abortions, stillbirths and intrauterine deaths. The reason for high percentage of missing values in perinatal sepsis and fetal distress is because the data collection of these variables started later.

(Note: The gestational age from here on will mean gestational age at delivery.)

3.1.3 Imputation of missing values

Table 3 Summary of variables used for clustering

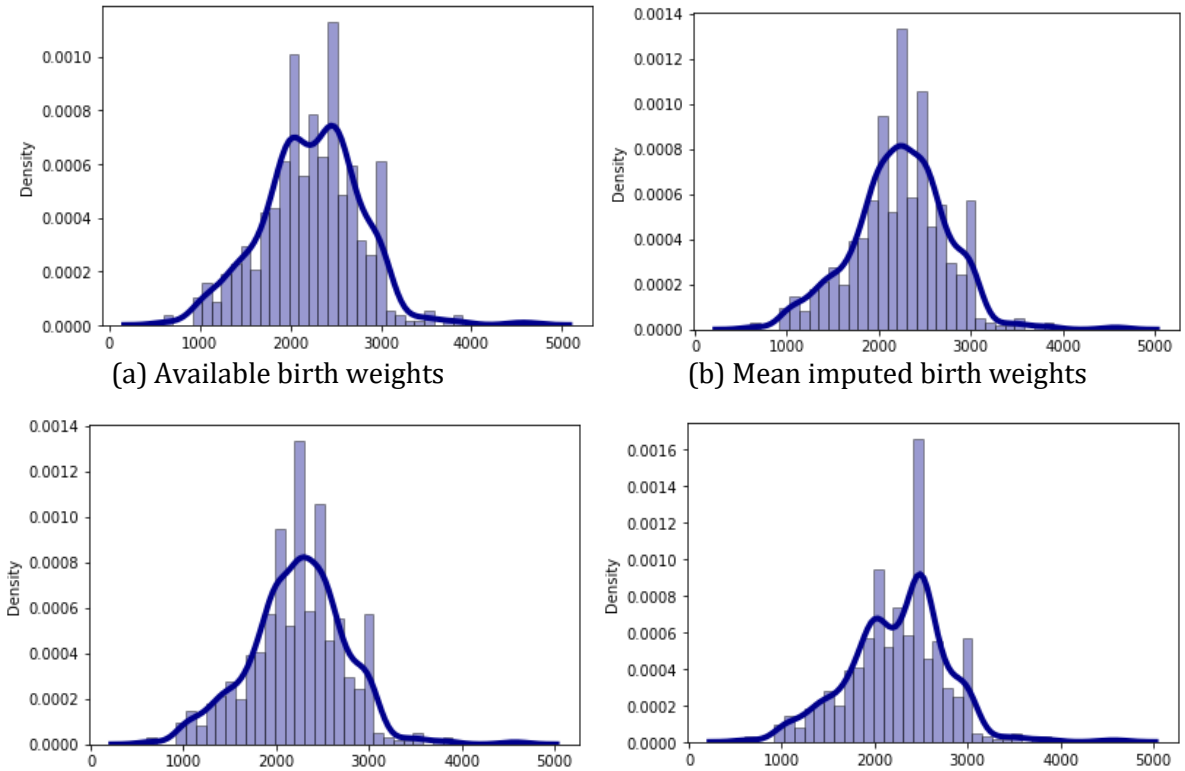
	Variable	Missing %	Missing	Data type	0	1	2
1	Clinical chorioamnionitis	64.70588	374	Categorical	203	0	
2	Extrauterine infections	50.69204	293	Categorical	278	6	
3	Maternal trauma	1.384083	8	Categorical	569	0	
4	Medical disorders	57.26644	331	Categorical	70	176	
5	Uterine rupture	55.19031	319	Categorical	257	1	
6	Preeclampsia	60.0346	347	Categorical	133	97	
7	Eclampsia	1.384083	8	Categorical	566	3	
8	Fetal anomalies	26.47059	153	Categorical	403	21	
9	Multiple birth	0	0	Categorical	558	19	
10	Amniotic fluid index	42.04152	243	Categorical	270	22	42
11	Perinatal sepsis	80.44983	465	Categorical	41	71	
12	Fetal distress	91.86851	531	Categorical	3	43	
13	Birth weight	6.401384	37	Continuous			
14	Abruption	43.59862	252	Categorical	324	1	
15	Placenta previa	43.94464	254	Categorical	317	6	

Table 4 Summary of variables used for analyzing clusters

	Variable	Missing %	Missing	Data type	0	1
1	Cervical length	42.56055	246	Categorical		
2	Cervical shortening	42.56055	246	Categorical	291	40
3	PPROM	13.84083	80	Categorical	421	76
4	Cervical dilation	46.02076	266	Categorical	63	248
5	Bleeding	1.384083	8	Categorical	529	40
6	SPL	33.04498	191	Categorical	147	239
7	Induction	1.730104	10	Categorical	502	65

All the variables except birth weight and cervical length were categorical, where 0 represents healthy state w.r.t to the condition and 1 represents that the condition is active. Polyhydramnios and oligohydramnios were depicted in a single categorical variable where *0* represented the healthy state, *1* represents oligohydramnios and *2* represents polyhydramnios. Also, those cases in which the variable was not applicable was taken as 0. As no positive cases for clinical chorioamnionitis and maternal trauma were found in the dataset of live preterm births, these variables were dropped.

Birth weight and cervical length were continuous data type. The distribution of birth weights was studied using a normed histogram, where the sum of products of width and height of each column is equal to the total count, along with a gaussian kernel density estimate. The mean was found to be 2229.952 g, median 2255.0 g and mode 2500.0g. For the purpose of imputation of missing values, mean imputed birth weights was chosen due to its symmetric bell curve.



(a) Available birth weights (b) Mean imputed birth weights

(c) Median imputed birth weights (d) Mode imputed birth weights

Figure 2 Density plots for birth weight

Another method for imputation chosen was regression. The polynomial of degree 2 and 3 were overfitting, as can be seen from the plot, and the linear curve is underfitting. Power and exponential curves showed similar R^2 values, so exponential curve was chosen for imputation due to its relatively higher R^2. The equation of this curve is $y = 115.87e^{0.0846x}$.

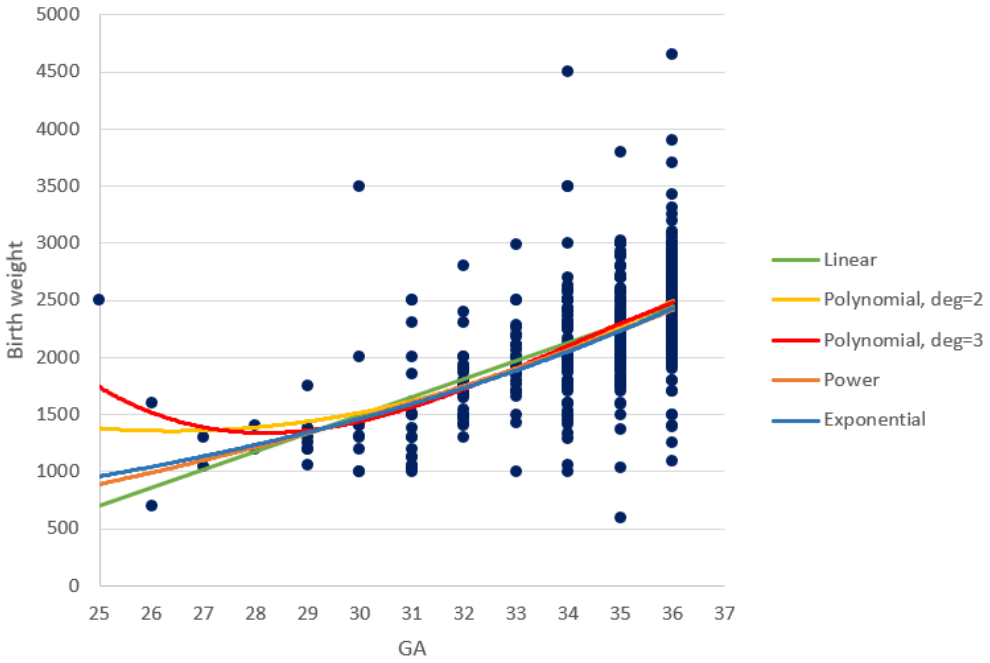


Figure 3 Various regression curves for birth weight

Table 5 R^2 value for various regressions

Regression	**R^2**
Linear	0.3046
Polynomial, deg=2	0.3205
Polynomial, deg=3	0.3241
Power	0.3133
Exponential	0.3159

3.1.4 Normalization

The variables birth weight and amniotic fluid index were normalized using the Min-Max Scaler, which is defined as below. This normalization has been used so that the range of the variable is between 0 and 1.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

3.2. Clustering

Agglomerative hierarchical clustering was performed using the sklearn library of Python. The linkage method used is Ward's method, as it gave the most separated clusters as seen from dendrogram.

3.2.1 Experiment I: Mean imputed birth weight and N=4

The maternal, fetal and placental characteristics were inputted for plotting the dendrogram, for which the Scipy library was used. The optimal number of clusters was found to be 4. Also, the sum of squared distances within cluster was plotted against increasing number of clusters (N). This method is called the elbow method. The "elbow" of the plot was found to be at N=4. As both the methods gave the same N, the number of clusters chosen was 4.

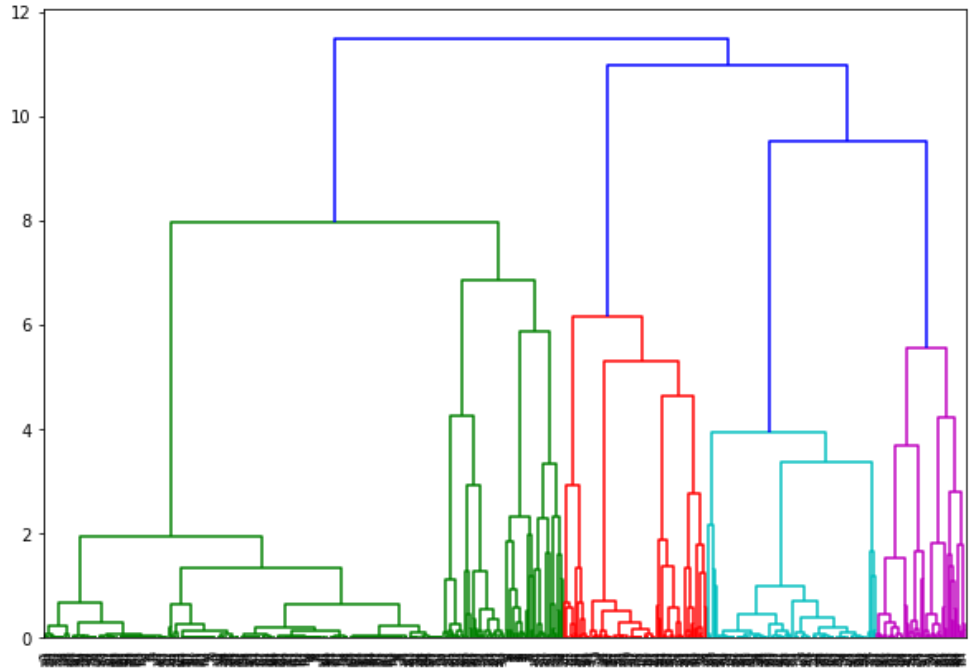


Figure 4 Dendrogram for dataset with mean imputed birth weights

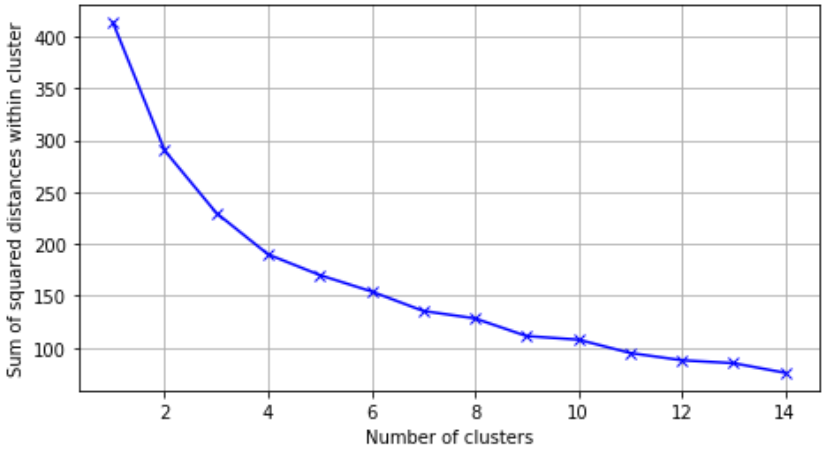


Figure 5 Elbow method to find optimal number of clusters

Clustering results obtained using N=4 are shown below.

Table 6 Results for Mean imputed birth weight and N=4

	Cluster Size	Mean GA (weeks)	Mean BW (g)	Cervical length (mm)	Major Conditions (%)	Parturition and delivery characteristics (%)
0	325	34.538	2279.460	33.498	Polyhydramnios 10.154 Medical disorders 7.692 Fetal anomalies 6.154 Multiple birth 5.538 Oligohydramnios 4.000 Preeclampsia 3.077 Perinatal sepsis 2.154 Eclampsia 0.615 Fetal distress 0.308	Cervical dilation 81.982 SPL 51.445 Cervical shortening 13.725 Induction 10.410 PPROM 10.078 Bleeding 4.101
1	106	35.236	2356.183	35.035	Medical disorders 95.283 Extrauterine infections 5.660 Placenta previa 4.717 Preeclampsia 1.887 Fetal distress 0.943 Eclampsia 0.943	Cervical dilation 79.612 SPL 71.111 PPROM 23.232 Induction 13.333 Cervical shortening 7.229 Bleeding 6.604
2	90	34.511	2188.512	34.400	Preeclampsia 90.000 Medical disorders 26.667 Fetal distress 21.111 Perinatal sepsis 15.556 Oligohydramnios 4.444 Polyhydramnios 3.333 Placenta previa 1.111 Abruption 1.111 Fetal anomalies 1.111	Cervical dilation 77.273 SPL 72.973 PPROM 14.943 Induction 13.483 Bleeding 12.222 Cervical shortening 10.526
3	56	33.036	1774.722	34.021	Perinatal sepsis 89.286 Medical disorders 46.429 Fetal distress 39.286 Polyhydramnios 10.714 Oligohydramnios 8.929 Preeclampsia 7.143 Multiple birth 1.786 Uterine rupture 1.786	Cervical dilation 77.358 SPL 65.306 PPROM 26.415 Cervical shortening 18.421 Bleeding 16.071 Induction 10.714

3.2.2 Experiment II Mean imputed birth weight and N=10

In experiment I, though most had a dominating condition, all the clusters showed mixed conditions. Also, the parturition and delivery characteristics, mean birth weight and mean gestational age were not differentiable. For practical purposes, we need clusters which can be easily distinguished from one another.

Since the clusters obtained with N=4 did not look useful, the number of clusters was decided to be increased. For this, we plotted the silhouette score, a measure of inter-cluster distance along with cohesion, against increasing number of clusters. Generally, a score greater than 0.6 is considered satisfactory, hence a cutoff of 0.6 was set. Using this, the new optimal number of clusters obtained was N=10.

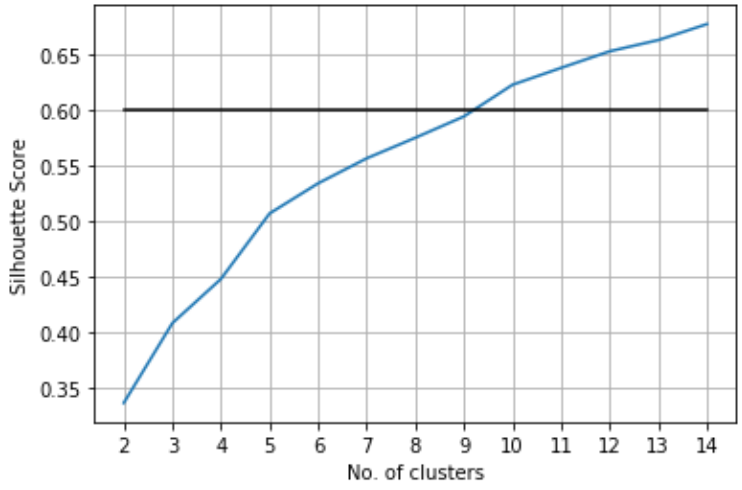


Figure 6 Silhouette score for dataset with mean imputed birth weights

The clusters obtained are shown below.

Table 7 Results for Mean imputed birth weight and N=10

	Cluster Size	Mean GA (weeks)	Mean BW (g)	Cervical length (mm)	Major Conditions (%)	Parturition and delivery characteristics (%)
0	250	34.576	2339.739	33.007	None	Cervical dilation 87.692 SPL 44.915 Cervical shortening 13.830 Induction 8.264 PPROM 7.407 Bleeding 3.719
1	106	35.236	2356.183	35.035	Medical disorders 95.283 Extrauterine infections 5.660 Placenta previa 4.717 Preeclampsia 1.887 Fetal distress 0.943 Eclampsia 0.943	Cervical dilation 79.612 SPL 71.111 PPROM 23.232 Induction 13.333 Cervical shortening 7.229 Bleeding 6.604
2	41	34.927	2233.500	33.032	Preeclampsia 100.0	Cervical dilation 83.333 SPL 75.676 Bleeding 14.634 PPROM 12.821 Cervical shortening 9.091 Induction 7.500
3	39	35.256	2372.333	34.336	Polyhydramnios 82.051 Medical disorders 33.333 Oligohydramnios 17.949 Preeclampsia 12.821 Eclampsia 2.564	Cervical dilation 66.667 SPL 64.000 Induction 25.641 PPROM 18.919 Cervical shortening 13.889 Bleeding 2.564
4	32	33.906	2061.188	35.136	Preeclampsia 100.000 Medical disorders 75.000 Perinatal sepsis 43.750 Polyhydramnios 9.375 Oligohydramnios 6.250 Fetal distress 6.250 Placenta previa 3.125 Abruption 3.125	Cervical dilation 82.609 SPL 70.370 PPROM 21.875 Induction 15.625 Cervical shortening 13.636 Bleeding 9.375

5	30	32.633	1748.179	33.211	Perinatal sepsis 100.000 Fetal distress 33.333 Oligohydramnios 10.000 Polyhydramnios 6.667 Multiple birth 3.333	Cervical dilation 79.310 SPL 73.077 PPROM 24.138 Cervical shortening 21.053 Bleeding 13.333 Induction 10.000
6	26	33.500	1803.308	34.832	Medical disorders 100.000 Perinatal sepsis 76.923 Fetal distress 46.154 Polyhydramnios 15.385 Preeclampsia 15.385 Oligohydramnios 7.692 Uterine rupture 3.846	Cervical dilation 75.000 SPL 56.522 PPROM 29.167 Bleeding 19.231 Cervical shortening 15.789 Induction 11.538
7	18	33.667	1910.118	35.736	Fetal anomalies 100.000 Oligohydramnios 16.667 Perinatal sepsis 16.667 Medical disorders 16.667 Preeclampsia 11.111	Cervical dilation 81.818 SPL 70.588 PPROM 17.647 Cervical shortening 9.091 Induction 5.556 Bleeding 5.556
8	18	33.333	1670.222	32.775	Multiple birth 100.000 Medical disorders 50.000 Perinatal sepsis 22.222 Oligohydramnios 16.667 Preeclampsia 16.667 Fetal anomalies 11.111 Polyhydramnios 5.556 Fetal distress 5.556 Eclampsia 5.556	Cervical dilation 81.818 SPL 61.538 Cervical shortening 16.667 PPROM 13.333 Induction 11.111 Bleeding 11.111
9	17	34.647	2347.562	35.469	Fetal distress 100.000 Preeclampsia 47.059 Oligohydramnios 11.765 Fetal anomalies 5.882	SPL 70.000 Cervical dilation 55.556 Induction 23.529 Bleeding 11.765 Cervical shortening 7.692 PPROM 6.250

3.2.3 Experiment III: Regression imputed birth weight and N=10

Although the clusters obtained in experiment II are well differentiable, we used birth weight imputed using regression based on gestational age in experiment III as it was believed to be more reliable, since we know that the birth weight varies with gestational age. The dendrogram obtained in this experiment was different from before.

Again, from the dendrogram, the optimal number of clusters was seen as 4. However, since we know that this would not give desirable results, we followed the silhouette score plot with a cutoff of 0.6. Once again, we obtained N=10.

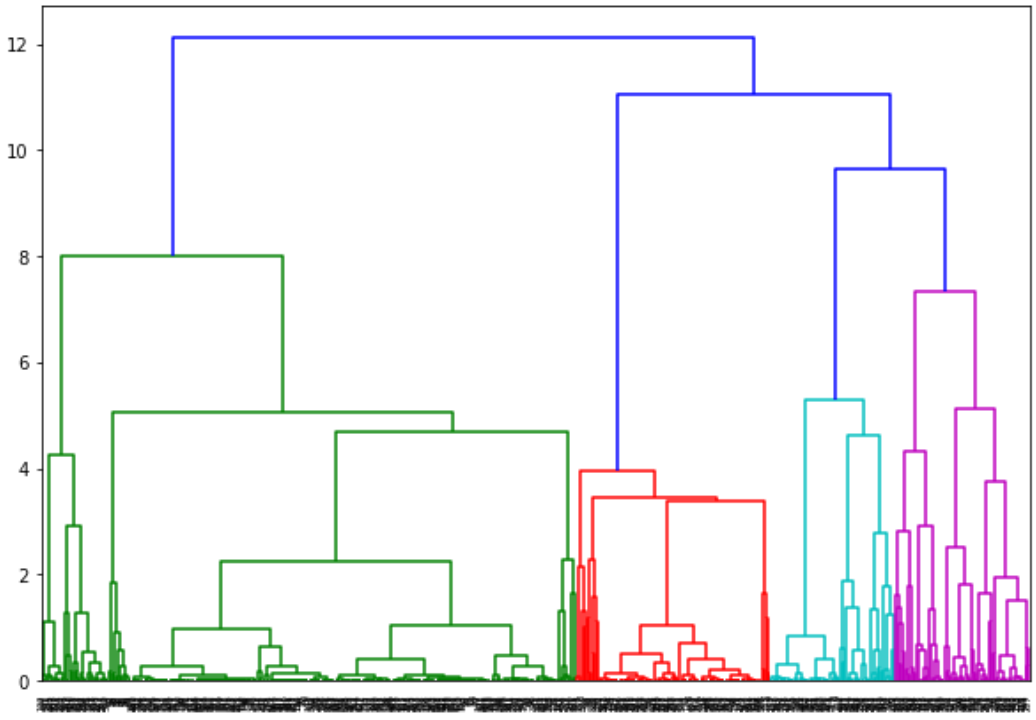


Figure 7 Dendrogram for dataset with regression imputed birth weight

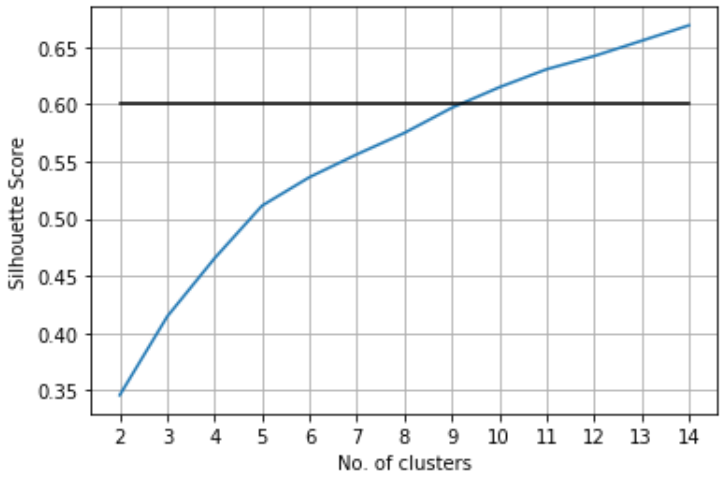


Figure 8 Silhouette score for dataset with regression imputed birth weights

The results obtained are shown below.

Table 8 Results for regression imputed birth weight and N=10

	Cluster Size	**Mean GA (weeks)**	**Mean BW (g)**	**Cervical length (mm)**	**Major Conditions (%)**	**Parturition and delivery characteristics (%)**
0	250	34.576	2339.739	33.007	None	Cervical dilation 87.692 SPL 44.915 Cervical shortening 13.830 Induction 8.264 PPROM 7.407 Bleeding 3.719
1	113	35.168	2327.676	35.121	Medical disorders 95.575 Extrauterine infections 5.310 Placenta previa 4.425 Fetal anomalies 4.425 Multiple birth 3.540 Oligohydramnios 1.770 Fetal distress 1.770 Preeclampsia 1.770	Cervical dilation 80.909 SPL 71.134 PPROM 22.857 Induction 12.500 Bleeding 7.080 Cervical shortening 6.742

2	41	34.927	2233.500	33.032	Preeclampsia 100.0	Cervical dilation 83.333 SPL 75.676 Bleeding 14.634 PPROM 12.821 Cervical shortening 9.091 Induction 7.500
3	39	35.256	2372.333	34.336	Polyhydramnios 82.051 Medical disorders 33.333 Oligohydramnios 17.949 Preeclampsia 12.821 Eclampsia 2.564	Cervical dilation 66.667 SPL 64.000 Induction 25.641 PPROM 18.919 Cervical shortening 13.889 Bleeding 2.564
4	32	33.906	2061.188	35.136	Preeclampsia 100.000 Medical disorders 75.000 Perinatal sepsis 43.750 Polyhydramnios 9.375 Oligohydramnios 6.250 Fetal distress 6.250 Placenta previa 3.125 Abruption 3.125	Cervical dilation 82.609 SPL 70.370 PPROM 21.875 Induction 15.625 Cervical shortening 13.636 Bleeding 9.375
5	32	32.375	1698.300	33.420	Perinatal sepsis 100.000 Fetal distress 31.250 Oligohydramnios 12.500 Polyhydramnios 6.250 Fetal anomalies 6.250 Multiple birth 3.125	Cervical dilation 77.419 SPL 70.370 PPROM 25.806 Cervical shortening 20.000 Bleeding 15.625 Induction 12.500
6	29	33.897	2059.000	34.724	Fetal distress 100.000 Preeclampsia 41.379 Medical disorders 41.379 Perinatal sepsis 20.690 Oligohydramnios 10.345 Polyhydramnios 6.897 Fetal anomalies 3.448 Uterine rupture 3.448	Cervical dilation 61.905 SPL 55.000 Bleeding 24.138 Induction 20.690 PPROM 14.815 Cervical shortening 14.286
7	18	34.222	1883.722	35.543	Perinatal sepsis 100.000 Medical disorders 100.000 Multiple birth 22.222 Oligohydramnios 11.111 Polyhydramnios 11.111	Cervical dilation 87.500 SPL 76.471 PPROM 23.529 Cervical shortening 7.143 Induction 5.556

8	13	34.385	1980.000	36.000	Fetal anomalies 100.000 Oligohydramnios 15.385 Preeclampsia 15.385	Cervical dilation 83.333 SPL 69.231 PPROM 16.667 Cervical shortening 12.500
9	10	32.400	1633.500	27.840	Multiple birth 100.0 Preeclampsia 30.0 Polyhydramnios 10.0 Eclampsia 10.0 Medical disorders 10.0	SPL 40.000 Cervical shortening 40.000 Cervical dilation 33.333 Induction 20.000 PPROM 12.500 Bleeding 10.000

3.2.4 Experiment IV: Regression imputed birth weight and N=11

In this experiment, the number of clusters were increased to 11 to check if any important clusters are present which were missed out with N=10. It was observed that N=11 showed two new clusters, which branched out from Cluster-4 of experiment II, with although similar parturition and delivery characteristics, but very different mortality and morbidity outcomes, as will be discussed later.

Table 9 Results for regression imputed birth weight and N=11

	Cluster Size	**Mean GA (weeks)**	**Mean BW (g)**	**Cervical length (mm)**	**Major Conditions (%)**	**Parturition and delivery characteristics (%)**
0	250	34.576	2339.739	33.007	None	Cervical dilation 87.692 SPL 44.915 Cervical shortening 13.830 Induction 8.264 PPROM 7.407 Bleeding 3.719

1	113	35.168	2327.676	35.121	Medical disorders 95.575 Extrauterine infections5.31 Placenta previa 4.425 Fetal anomalies 4.425 Multiple birth 3.540 Oligohydramnios 1.770 Fetal distress 1.770 Preeclampsia 1.770	Cervical dilation 80.909 SPL 71.134 PPROM 22.857 Induction 12.500 Bleeding 7.080 Cervical shortening 6.742
2	41	34.927	2233.500	33.032	Preeclampsia 100.0	Cervical dilation 83.333 SPL 75.676 Bleeding 14.634 PPROM 12.821 Cervical shortening 9.091 Induction 7.500
3	39	35.256	2372.333	34.336	Polyhydramnios 82.051 Medical disorders 33.333 Oligohydramnios 17.949 Preeclampsia 12.821 Eclampsia 2.564	Cervical dilation 66.667 SPL 64.000 Induction 25.641 PPROM 18.919 Cervical shortening 13.889 Bleeding 2.564
4	32	32.375	1698.300	33.420	Perinatal sepsis 100.000 Fetal distress 31.250 Oligohydramnios 12.500 Polyhydramnios 6.250 Fetal anomalies 6.250 Multiple birth 3.125	Cervical dilation 77.419 SPL 70.370 PPROM 25.806 Cervical shortening 20.000 Bleeding 15.625 Induction 12.500
5	29	33.897	2059.000	34.724	Fetal distress 100.000 Preeclampsia 41.379 Medical disorders 41.379 Perinatal sepsis 20.690 Oligohydramnios 10.345 Polyhydramnios 6.897 Fetal anomalies 3.448 Uterine rupture 3.448	Cervical dilation 61.905 SPL 55.000 Bleeding 24.138 Induction 20.690 PPROM 14.815 Cervical shortening 14.286
6	18	34.222	1883.722	35.543	Perinatal sepsis 100.000 Medical disorders 100.000 Multiple birth 22.222 Oligohydramnios 11.111 Polyhydramnios 11.111	Cervical dilation 87.500 SPL 76.471 PPROM 23.529 Cervical shortening 7.143 Induction 5.556

7	18	34.500	2220.833	35.569	Preeclampsia 100.000 Medical disorder 100.000 Polyhydramnios 11.111 Oligohydramnios 5.556 Abruption 5.556	SPL 73.333 Cervical dilation 71.429 Induction 16.667 PPROM 16.667 Cervical shortening 15.385 Bleeding 5.556
8	14	33.143	1855.929	34.511	Perinatal sepsis 100.000 Preeclampsia 100.000 Medical disorder 42.857 Fetal distress 14.286 Oligohydramnios 7.143 Polyhydramnios 7.143 Placenta previa 7.143	Cervical dilation 100.000 SPL 66.667 PPROM 28.571 Induction 14.286 Bleeding 14.286 Cervical Shortening 11.111
9	13	34.385	1980.000	36.000	Fetal anomalies 100.000 Oligohydramnios 15.385 Preeclampsia 15.385	Cervical dilation 83.333 SPL 69.231 PPROM 16.667 Cervical shortening 12.500
10	10	32.400	1633.500	27.840	Multiple birth 100.0 Preeclampsia 30.0 Polyhydramnios 10.0 Eclampsia 10.0 Medical disorders 10.0	SPL 40.000 Cervical shortening 40.000 Cervical dilation 33.333 Induction 20.000 PPROM 12.500 Bleeding 10.000

3.2.5 Experiment V: Regression imputed birth weight and N>11

Clusters numbers were increased to 12 and 13 to check for new distinct clusters. However, clusters formed with numbers above N=11 showed branching out of the current clusters with similar parturition and delivery characteristics as well as mortality and morbidity outcomes, hence the number of clusters were not increased further. Also, the cluster sizes got reduced to single digits. Hence, the results are not shown here.

3.3. Results and Discussion

It is observed that the results obtained in experiments II and III are similar in terms of cluster sizes and maternal, fetal, placental, initiation of parturition and delivery characteristics. The results of experiment IV will be discussed finally due to the more reliable imputation, as birth weight varies with gestational age.

3.3.1 Cluster-wise observations

The behavior of the clusters was studied in terms of neonatal mortality, requirement of ICU by the newborn and infant mortality (for validation of the classification), and parturition and delivery characteristics. For the purpose of easy visualization of parturition and delivery characteristics, a heatmap has been plotted after standardizing the variables.

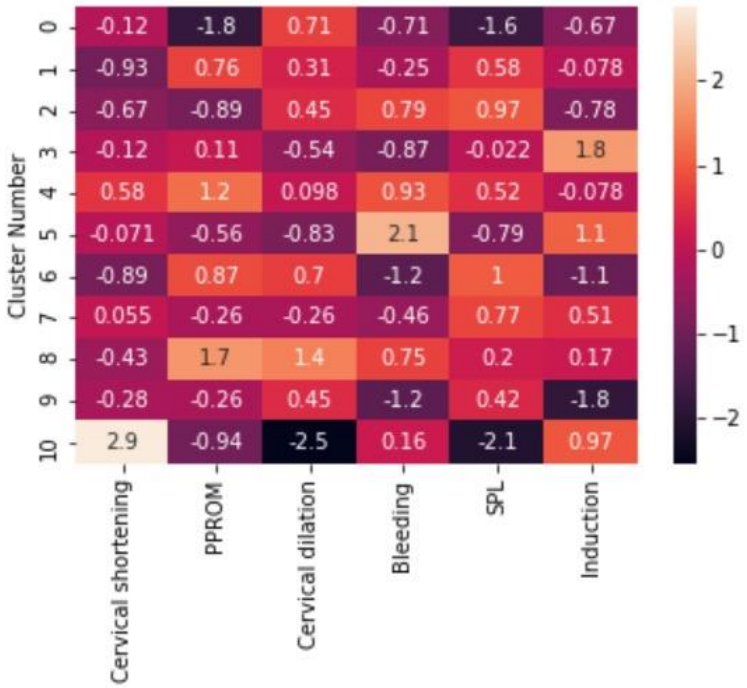


Figure 9 Heatmap of delivery and parturition characteristics

Table 10 Summary and outcomes of the obtained clusters

(Conditions constituting less than 5% are omitted)

	Mean GA (weeks)	**Mean BW (g)**	**Main condition (%)**	**Frequently Associated Conditions (%)**	**Neonatal death (%)**	**NICU (%)**	**Infant death (%)**
0	34.576	2339.739	None		0.00	7.94	17.95
1	35.168	2327.676	Medical disorders (95.6)	Extrauterine infections (5.31)	0.00	11.76	14.81
2	34.927	2233.500	Preeclampsia (100)		2.78	15.15	18.75
3	35.256	2372.333	Polyhydramnios (82.05)	Medical disorders (33.33) Oligohydramnios (17.95) Preeclampsia (12.82)	0.00	16.67	21.05
4	32.375	1698.300	Perinatal sepsis (100)	Fetal distress (31.25) Oligohydramnios (12.5) Polyhydramnios (6.25) Fetal anomalies (6.25)	18.75	100.00	26.32
5	33.897	2059.000	Fetal distress (100)	Preeclampsia (41.38) Medical disorders (41.38) Perinatal sepsis (20.69) Oligohydramnios (10.35) Polyhydramnios (6.90)	23.08	92.00	11.76
6	34.222	1883.722	Perinatal sepsis (100) Medical disorders (100)	Multiple birth (22.22) Oligohydramnios (11.11) Polyhydramnios (11.11)	5.56	100.00	10.00
7	34.500	2220.833	Preeclampsia (100) Medical disorders (100)	Polyhydramnios (11.11) Oligohydramnios (5.56) Abruption (5.56)	5.56	22.22	33.33
8	33.143	1855.929	Perinatal sepsis (100) Preeclampsia (100)	Medical disorders (42.86) Fetal distress (14.29) Oligohydramnios (7.14) Polyhydramnios (7.14) Placenta previa (7.143)	21.43	100.00	22.22

9	34.385	1980.000	Fetal anomalies (100)	Oligohydramnios (15.39) Preeclampsia (15.39)	0.00	16.67	0.00
10	32.400	1633.500	Multiple birth (100)	Preeclampsia (30) Polyhydramnios (10) Eclampsia (10) Medical disorders (10)	0.00	60.00	0.00

Cluster-0, the largest cluster, showed none of the conditions. It has one of the highest average birth weight and lowest neonatal mortality and neonatal morbidity, as seen from neonatal deaths and NICU requirement respectively. This group has the low percentages for PPROM and bleeding, and high percentage for cervical dilation.

Cluster-1, the second largest cluster, has medical disorders as its main condition and has one of the highest mean gestational age. It has the lowest percentage for cervical shortening and has a high percentage of spontaneous preterm labor. It has a low neonatal mortality and morbidity rate.

The sizes of clusters drop drastically after Cluster-1. Cluster-2 constituted of purely preeclampsia with no associated conditions. It has one of the highest percentages for spontaneous preterm labor.

Cluster-3 constitutes of majorly polyhydramnios. It has the highest mean gestational age and mean birth weight. It also shows lowest percentage for bleeding and highest percentage for induction (25%).

Custer-4, having perinatal sepsis associated most with fetal distress, shows one of the lowest mean gestational age and mean birth weight. It also has one of the highest PPROM percentage. The group shows one of the highest neonatal mortality and morbidity rates.

Cluster-5 has multiple conditions associated with fetal distress. It shows the highest percentage of bleeding. It has the highest neonatal mortality rate and extremely high neonatal morbidity.

Cluster-6 has perinatal sepsis as well as medical disorders, some having multiple birth as well. It shows the lowest percentage for cervical shortening and bleeding, and highest percentage for spontaneous preterm labor and cervical dilation. Although the neonatal mortality rate is on the lower side, the neonatal morbidity is high.

Cluster-7 has cases of preeclampsia and medical disorders. This group has the highest infant death rate.

Cluster-8 consists of cases of perinatal sepsis and preeclampsia associated with medical disorders. It has the highest percentage for PPROM and cervical dilation and is subject high mortality and morbidity rates.

Cluster-9 constitutes of mostly cases of fetal anomalies.

Cluster-10, consisting of multiple births, has the lowest mean gestational age and mean birth weight. It has the highest percentage of cervical shortening and the lowest percentages for cervical dilation and spontaneous preterm labor.

It is observed that the larger sized clusters have lower neonatal mortality, lower neonatal morbidity and higher mean birth weights and mean gestational ages. Also, overlaps of conditions in the clusters are seen. This may be due to common causal factors of the conditions.

3.3.2 A comparison with Newborn Cross-Sectional Study of the INTERGROWTH-21st study

A comparison was done with the results of the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project (Fernando C. Barros, Aris T. Papageorghiou, Cesar G. Victora, et al), which was taken as the reference study, to find clusters common to ours and theirs study. A common cluster means that it is dominated by the same conditions and shows similar parturition and delivery characteristics as well as similar neonatal mortality and morbidity. It was found that Cluster-0 (no serious conditions, and the largest one in both the studies), Cluster-1 (mostly medical disorders) and Cluster-10 (mostly multiples) behave similarly in both the studies in terms of the factors mentioned before. Also, it is noted that clusters with perinatal sepsis as major condition have high mortality and morbidity rates. Clusters with mostly preeclampsia and mostly fetal distress are present in both studies but the associated conditions are different.

The other clusters in the studies do not show similarity. This is due to the various reasons:

- Stillbirths were not considered in our study because we aim to find phenotypic classes of only live births. This difference in consideration is because the aims of the two studies differ slightly. The INTEGROWTH growth study tries to understand the underlying factors leading to preterm births. Our study is more interested in using the phenotype classes for risk-prediction purposes.

- Some of the variables used in the studies are different.
  The placental characteristics included in the INTEGROWTH growth study are early bleeding, mid-/late- vaginal bleeding and third trimester vaginal bleeding due to unavailability of precise diagnosis of placenta related conditions. Vaginal bleeding could be caused due to more than one reason, such as placental abruption and placenta previa. Hence, we have used these as placental characteristics since

we have the required data. The INTEGROWTH growth study also lacked information on cervical dilation and other signs of parturition.

The GARBH-Ini data had extremely high percentage of missing data (97%) on the types of caregiver-initiated delivery-clinically mandated, clinical discretion and no clinical indication, hence these could not be included in the study.

- There were no cases of chorioamnionitis in our sample of live preterm births, thus the variable was dropped.

- The complete data on fetal anemia and fetal growth restriction from USG scans are awaited, hence these could not be incorporated in the models yet. Suspected fetal growth restriction understood through birth weight was taken into in the birth weight variable itself. However, fetal growth restriction is also diagnosed through USG scans.

- Our study was specific to the Indian population. A rearrangement of the clusters may be possible for different populations.

CHAPTER 4

SUMMARY AND CONCLUSIONS

In this study, we proposed that for better medical interventions for preterm births and for an accurate risk-prediction model for preterm births for the Indian population, phenotypic classes should be identified.

A preterm population of 577 was selected from 8473 enrollments of pregnant women. We combined a conceptual framework suggested by Jose Villar, Aris T. Papageorghiou et al with an unsupervised learning technique to divide this population into distinct groups of phenotypes, based on maternal, fetal, placental characteristics, signs of initiation of parturition and pathway to delivery. 57% of the preterm births could be associated to at least one of maternal, fetal, placental conditions. These groups were found to show different outcomes in terms of birth weights, gestational ages and correspondingly neonatal deaths and requirement of neonatal ICU, proving that this classification could be useful.

However, 43%of the preterm births in the study were not associated with any of fetal, maternal, or placental conditions that were defined beforehand. It would require further investigation to study this group as it is possible that some features may have been left out of consideration.

RECOMMENDATIONS FOR THE FUTURE WORK

The model in this study has a few missing elements. Fetal growth restriction is recorded in two ways: (1) birth weight and (2) fetal weight estimated by ultrasonography. Fetal anemia is also diagnosed by Doppler ultrasound. The data on these two variables are awaited, and will be incorporated in the study once received. The groups that will be obtained then are expected to be more reliable.

According to this study, if a patient has a given set of conditions, we should be able to identify which group she falls in to perform effective medical intervention accordingly. This may not be easy as overlaps exist in the groups obtained. An experimental validation of the proposal needs to be done to understand this.

The clusters obtained in this study can be used as an input for preterm birth risk-prediction models, instead of considering all preterm births as a whole. It may give more accurate results as the model will be designed particularly for a group of conditions.

APPENDIX

A. Ward's method

The Ward's method is a criterion used in hierarchical agglomerative clustering to provide minimum within-cluster variance. It is the most common criterion used in agglomerative hierarchical clustering.

If q is a cluster and i is an observation in it, let $d(i,j)$ be the distance (or dissimilarity in this case) between two observations *i* and *j*, and let the mass associated with the observation *i* be

$p(i) = \frac{1}{|q|}$ when $i \in q$,

i.e. 1 over cluster cardinality of the respective cluster.

Let q^* be the cluster's center. Then,

$$q^* = \frac{1}{|q|} \sum_{i \in q} i$$

The error sum of squares can be defined as:

$$\sum_{i \in q} d^2(i, q^*)$$

The variance, or the centered sum of squares can be defined as:

$$\frac{1}{|q|} \sum_{i \in q} d^2(i, q^*)$$

Lance and Williams (1967) had developed a family of updation algorithms for agglomerative hierarchical clustering. According to this, the dissimilarity is updated recursively at each step of combining clusters (or possibly a singleton observations) *i* and *j* to form $i \cup j$ relative the an external cluster *k* as follows:

$$d(i \cup j, k) = a(i) * d(i,k) + a(j) * d(j,k) + b * d(i,j) + c * |d(i,k) - d(j,k)|$$

where the coefficients *a(i), a(j), b and c* are parameters which are dependent on the clustering criterion being used. At each step, those two clusters merge that bring the smallest increase in the combined error sum of squares.

B. Regression

Regression of birth weight based on gestational age was done sing Ordinary Least Squares (OLS) method. If a matrix **X** contains the values of gestational ages, and if a vector **y** contains the values of birth weights, let a matrix **c** contain the values of regression coefficients. The linear model would be given as:

$$\boldsymbol{y = c\,X + \varepsilon}$$

Where **ε** is the error matrix that has to be minimized. The estimated regression coefficients matrix which minimizes the error, the sum of squared deviations between actual ***y*** and predicted ***y***, is given by:

$$\boldsymbol{X\,\hat{c} = \hat{y}}$$

$$\boldsymbol{\hat{c} = (X^T X)^{-1} X^T \hat{y}}$$

Linear and Polynomial Regression

In polynomial regression of degree *p*,

$$y = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + ... + c_p x^p$$

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{1} & \boldsymbol{x_1^1} & \cdots & \boldsymbol{x_1^p} \\ \mathbf{1} & \boldsymbol{x_2^1} & \ldots & \boldsymbol{x_2^p} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \boldsymbol{x_n^1} & \cdots & \boldsymbol{x_n^p} \end{pmatrix} \qquad \hat{\boldsymbol{c}} = \begin{pmatrix} \boldsymbol{c_0} \\ \boldsymbol{c_1} \\ \vdots \\ \boldsymbol{c_p} \end{pmatrix} \qquad \hat{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_n} \end{pmatrix}$$

Linear regression would be solved with degree p =1.

Exponential Regression

The exponential equation is transformed to a linear model by taking natural logarithm on both sides, the solved as standard linear regression.

$$y = Ae^{Bx}$$

$$ln\,(y) = ln(A) + Bx$$

$$y' = c_0 + c_1 x$$

In this case, the sum of squared deviations between actual Ln(y) and predicted Ln(y) is minimized.

Power Regression

Similarly, the power equation is transformed to a linear model by taking natural logarithm on both sides, the solved as standard linear regression similar to the case of exponential regression.

$$y = Ax^B$$

$$ln\,(y) = ln(A) + B\,\ln\,(x)$$

$$y' = c_0 + c_1 x'$$

R^2 value

R^2 value, also known as the coefficient of determination, measures the closeness of the data to the fitted regression line. It is defined as the percentage of the variance in the dependent variable that the independent variables are explaining.

$$R^2 = \frac{Variance\ explained\ by\ the\ model}{Total\ variance}$$

$$R^2 = 1 - \frac{Sum\ of\ squares\ of\ residuals}{Total\ sum\ of\ squares}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - y_{mean})^2}$$

Its values range from 0 to 1, where 1 means that the variance of the dependent variable is completely explained by the model, and 0 means that the model could not explain any of the variation of the dependent variable.

C. METHODS TO CHOOSE NUMBERS OF CLUSTERS

Although there is no right way to select the number of clusters, and often domain knowledge is required and the number of clusters may be dependent on the problem statement that is being solved, there are some standard methods that are frequently followed for this purpose.

Dendrogram

If there is a requirement of minimum dissimilarity between clusters, a horizontal line is drawn, say *y=t,* on a dendrogram, the number of clusters obtained will be the number of vertical lines cut by the line *y=t*, where t is the threshold for dissimilarity that we have set.

Another way to decide the number of clusters is by cutting the dendrogram where the dissimilarity between 2 merges is maximum. This can naturally be seen as the tallest vertical line on the dendrogram.

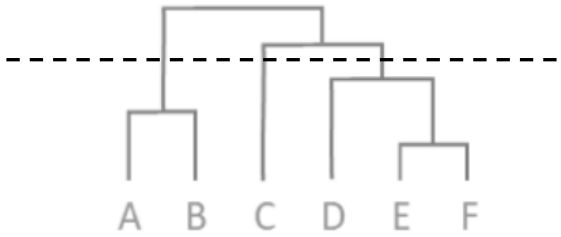


Figure 10 An example where 3 clusters are obtained

Note: Adapted from "What is a Dendrogram" by Tim Bock, *DisplayR*

Elbow method

In the elbow method in cluster analysis, the explained variation is plotted against the number of clusters following a k-means clustering. Then, the "elbow" of the curve is chosen as the number of clusters. The intuition is that the elbow represent the number of clusters beyond which the model is overfitting, because naturally, more the number of clusters better will be the model fit. The explained variation on the y-axis can be either of the two:

1. Inertia- the sum of squares of distances of observations from the centroid of the respective cluster
2. Distortion- the mean of Euclidean squared distance of observations from the centroid of the respective cluster

Silhouette score

Silhouette score validates the consistency within clusters and is defined as follow:

$$S_i = \frac{b_i - a_i}{Max(a_i, b_i)}$$

where ai = mean intra-cluster distance and bi = mean nearest-cluster distance for each sample. Intuitively, it tells how far away the observations of a cluster are from other clusters. It has a range from -1 to 1, and the value should be closer to 1.

REFERENCES

1. Michael S. Kramer, Aris Papageorghiou, Jennifer Culhane et al. *Challenges in defining and classifying the preterm birth syndrome*. American Journal of Obstetrics and Gynecology, 2012.

2. Robert L. Goldenberg, Michael G. Gravett, Jay Iams, Aris T. Papageorghiou et al. *The preterm birth syndrome: issues to consider in creating a classification system*. American Journal of Obstetrics and Gynecology, 2012.

3. Jose Villar, Aris T. Papageorghiou, Hannah E. Knight. *The preterm birth syndrome: a prototype phenotypic classification*. American Journal of Obstetrics and Gynecology, 2012.

4. Fernando C. Barros, Aris T. Papageorghiou, Cesar G. Victora, Julia A. Noble, Ruyan Pang. *The Distribution of Clinical Phenotypes of Preterm Birth Syndrome: Implications for Prevention*. Journal of the American Medical Association, 2015.

5. *Prediction and Prevention of Preterm Birth.* The American College of Obstetricians and Gynaecologists. Practice Bulletin Number 130, 2012.

6. Shinjini Bhatnagar, Partha P. Majumder, Dinakar M. Salunke, and Interdisciplinary Group for Advanced Research on Birth Outcomes—DBT India Initiative (GARBH-Ini). *A Pregnancy Cohort to Study Multidimensional Correlates of Preterm Birth in India: Study Design, Implementation, and Baseline Characteristics of the Participants*. American Journal of Epidemiology, December 2018.

7. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

8. Fionn Murtagh, Pierre Legendre. *Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?* Journal of Classification, 2014

9. Middleton, *M.R. Data Analysis Using Microsoft Excel 5.0*. Duxbury Press, 1995