Data Preparation of the nuMoM2b Dataset

Anton Goretsky* Anastasia Dmitrienko† Irene Tang‡ Nicolae Lari†
Owen Kunhardt* Raiyan Rashid Khan† Cassandra Marcussen† Adam Catto*
Daniel Mallia* Alisa Leshchenko* Adam (Yun Chao) Lin† Anita Raja*
Ansaf Salleb-Aouissi†§ Itsik Pe'er† Ronald Wapner‡ Cynthia Gyamfi-Bannerman¶

Abstract

In 2010, the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) started the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be (nuMoM2b), a prospective cohort study of a racially/ethnically/geographically diverse population of nulliparous women with singleton gestation. The nuMoM2b is a very large dataset, consisting of data for 10,038 patients with over 4,600 features per patient, spread out over 80 files. In this report, we share our experience preparing and working with this dataset. We present our data preprocessing of the nuMoM2b dataset to get a deeper understanding of the data, extract the most relevant features, make the fewest assumptions when filling in unknown values, and reducing the dimensionality of the data. We hope this report is useful to researchers interested in building machine learning and statistical models from the nuMoM2b dataset.

1 Introduction to nuMoM2b and Data Processing

The primary goal of the nuMoM2b study [1] was to determine the maternal characteristics, both clinical and genetic factors, physiological response to pregnancy and environmental factors that could be used to derive models that accurately predict adverse pregnancy outcomes (APOs). Our team has extensive prior experience working on medical data for preterm birth (PTB) prediction [2, 4, 5].

As originally organized, the dataset is not immediately conducive to analysis to those unfamiliar with the medical background, nor is it conducive to quick placement into machine learning models. The ratio of instances to features would result in an inevitable model overfitting. Various medical categories exist within an individual file, data relevant to features exist throughout many different files, and dependencies and redundancies exist across the whole dataset. As such, nuMoM2b required extensive review and processing of features, their dependencies and relations in order to reduce the complexity of the dataset, and shape it in a form amenable to a variety of machine learning (ML) algorithms and exploratory data analysis (EDA).

The intended audience for this document is researchers interested in building machine learning and statistical models from the nuMoM2b dataset. Our aim is to share our experience preparing and working with this dataset. The intent is not to share the processed dataset nor the scripts that are specific to research aims.

Our specific research goal is to build machine learning models for the prediction and prevention of preterm birth in nulliparous women using the nuMoM2b dataset. This project is funded by the NIH/NLM (Project # 1R01LM013327-01). The data review and processing work described in this document was conducted through the direct collaboration of the Computer Science departments at Hunter College and Columbia University, along with maternal and fetal medicine experts at the Columbia University Medical Center. The following are the goals of this collaborative effort:

- Significant reduction of the feature space for a better management of the data and a a reduction of the risk for overfitting.

*Department of Computer Science, CUNY Hunter College
†Department of Computer Science, Columbia University
‡Department of Obstetrics and Gynecology, Columbia University
§Contact author ansaf@cs.columbia.edu
¶Department of Obstetrics, Gynecology, and Reproductive Sciences. UC San Diego Health Sciences

- A reformatting of the dataset, to allow for easy configuration of the data, conducive to exploration and machine learning modeling.
- Extensive literature review of existing risk factors related to data categories and causal pathways of PTB.
- Exploratory data analysis of both the reduced and the unmodified feature space.

As a result of the effort, the dataset was reduced to 364 features at the most general level of complexity, and 465 features at a higher level of detail. Extensive filtering and imputation rules were created to accomplish this goal, along with a system for both human readability and easy script interpretation. An extensive literature review was performed documenting odds ratios. EDA was performed on the dataset, comparing calculated odds ratios to the literature review, and discovering and correcting data inconsistencies. Finally, we summarize our thoughts on data preprocessing and nuMoM2b in Section 7.

We summarize and visualize the PTB statistics in the dataset in Figure 1 and 2. See [3] for more details about the dataset.

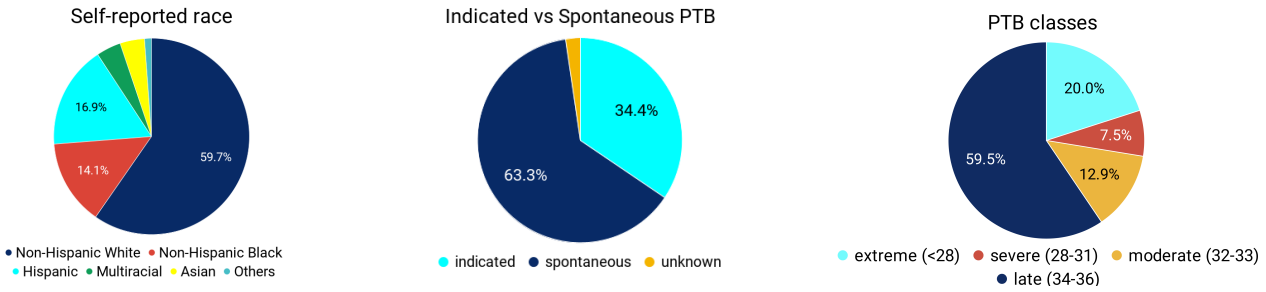

Figure 1: Preliminary statistics in the nuMoM2b data

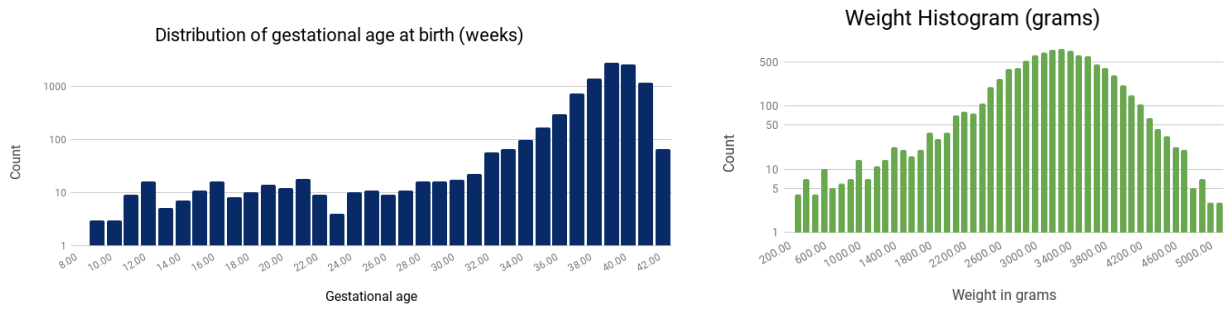

Figure 2: Distribution of gestational age and weight in the nuMoM2b data

2 Preprocessing

As part of the preprocessing stage, the information present in the provided codebooks on basic variable statistics, distribution, and type was transferred over via a script to a workable spreadsheet. Then to support our analysis, the following steps were taken:

- **"Data type" was corrected to a more accurate label.** Many variables were labeled "num" for numeric even though they actually represented categorical data. For example, Country of Birth (V1AF08) was labeled numeric. However, country cannot be treated as a numeric feature in modeling. Numeric, according to the codebook, simply meant the values used in labeling were numerals. This problem was repeated across the entire dataset. As such, each variable was looked at manually and categorized properly for typical modeling and analysis. Many variables labeled "num" were changed to "categorical".

- **A "Missingness" metric was calculated** using the variable statistics provided in the codebook to help with data analysis.
- **A new "Temporal" label was added** in order to label variables that appear at multiple unique time points throughout the study. nuMoM2b data was recorded at several distinct time points, and some variables are updated over time / questions are asked more than once. Temporal had values of True or False.
- **A new "Temporal Detail" label was added** in order to label at which time point a variable is relevant. Not all data collected are relevant only to the moment of collection in the study. For example, many demographic related questions were asked at Visit 2. However, those are static features and significant at all points in the study. This concept will be explained further under the section titled "Timeline".

Accuracy of action timing, such as treatment administration like medications, and the creation of these "temporal" labels are important for this team's goals of sequential treatment decision making modeling.

3 Organization into Groups

After the preprocessing stage, further organization was desired to better understand and manage the dataset. In order to better organize the data, **Filtering Groups** were created. The goal of these groups was to break down the dataset into more understandable portions for those unfamiliar with the detailed medical content, and for simpler processing of data in bulk. The following were the filtering groups created for the nuMoM2b dataset.

- **Treatment** - Variables relating to intervention in cases of potential or immediate PTB risk, such as progesterone administration, steroids, and last minute medical administrations at delivery.
- **Psychological** - Variables relating to the psychological state of the patient, through multiple scales.
- **Physiological** - Variables relating to instantaneous measurements of physiology, such as temperature, symptoms of flu-like illnesses, blood pressure, etc.
- **Medical History** - Variables relating to long term medical history and conditions, but also various tests performed on both the mother and the fetus or newborn found in the study.
- **Demographics** - Variables relating to the patient's demographic factors, such as race, income, education, etc.
- **Ultrasound** - Variables that were recorded from the research and clinical ultrasounds the patient went through and marked as such in the dataset.
- **Outcomes** - A metadata file containing various variables that are useful mainly as classifier labels, or are features that were collected post-delivery about the mother or newborn.
- **Activity** - Variables relating to the physical activity of the patient.
- **Toxicology** - Variables relating to the medications taken just before and during pregnancy, and / or their relations to particular reasons / conditions.
- **Family History** - Variables relating to family medical conditions and history.
- **Food Frequency Analysis** - Variables relating to food, diet and vitamins, in the three months prior to pregnancy.
- **Sleep Substudy** - Select variables from the two sleep substudies included in nuMoM2b.

4 Processing

Now that the dataset was more simply organized, the process of feature reduction began. Filtering groups were divided among teams, who worked in consultation with the Columbia University Medical Center OB/GYN collaborators, to determine which features to keep as is, which features to summarize into scales, scores, or other aggregate forms, and which features to remove for redundancy or other reasons. As a result of this effort, a system was created to organize the efforts in a form both readable by people, and interpretable by scripts, to allow for easy updates. This system consists of the creation of **Filtering** and **Imputation Rules**, and **Layering**. Filtering rules serve to keep, drop, or summarize features, while imputation rules serve to impute missing data as required by many modeling algorithms. Layering served to organize the data into different levels of abstraction from the most general to the most specific level of information.

4.1 Layers

Layers were decided upon given the high feature complexity of nuMoM2b, even after much feature reduction. As many questions are structured around a format of a general question followed by several sub-questions, it was reasoned that the general question should in most cases be representative of the data points that follow. For example, V1AD06 asks, *Have you had any 'flu-like illnesses', 'really bad colds', fever, a rash, or any muscle or joint aches since you became pregnant?* This question is then followed by questions in regard to which symptoms are actually present in this 'flu-like illness'. V1AD06 covers all, and is such a more general question, and thus would be selected into a more general layer, while symptom specific questions would be reserved for the detail-oriented layers, or dropped. If dropped, they may be brought back if desired, or if some significance is found in the most general feature. Internally, we decided upon 3 layers.

- Layer 0 would consist of known risk factors for PTB, along with variables shown to have high odds ratios in our EDA.
- Layer 1 would be the most general layer, consisting of L0 and all general questions that cover as much information as possible.
- Layer 2 would bring back detail that may have been lost, or not included given the generalization and simplicity of L0 and L1.

We will not go into variable-level detail in each layer, but we believe this concept can serve as an organizational method for large complex datasets.

4.2 Filtering and Imputation Rules

As part of the data cleaning process, data first passes through a general filtering script, with the rules shown in Table 1. It then passes through an imputation script, with the rules shown in Table 2. Throughout the processing of this dataset, we strived to hold to a set of generally applicable rules for imputation. Below is a sample of the imputation rules used for nuMoM2b.

- If a numeric-like feature is applicable to a vast majority of patients, and the missingness was relatively low, we attempted to impute with a value such as MEAN or MODE. If multiple measurements were made on the same information, the mean took into account all of them (excluding those marked as incomplete or inaccurate).
- If a feature serves as a general precursor to a list of follow-up questions, such as "*Have you had any flu-like illnesses, 'really bad cold', fever, a rash...*" followed by questions regarding symptoms, if data is present in the follow-up but the is missing for the general question, we impute the general question to whatever value represents True. Otherwise we often impute to an unknown or not applicable determiner such as 999, especially if assumption is misleading in the understanding of treatment. Often, these imputation rules looked at follow-up questions that were not included in the current layer or were excluded from modeling, due to the general question covering the topic. For the follow-up questions themselves, imputation may have been left at unknown or imputed to a value depending on the missingness and applicability.
- Negative one (-1) was often used to represent unknown or inapplicable for numeric features.

Table 1: List of Filter Rules and their Descriptions

Filter Type	**Description**
TEMP	Converts complementary Celsius and Fahrenheit temperature features into one Fahrenheit feature
WEIGHT	Converts complementary LB and KG weight features into one LB feature
ADD	Adds up given numeric features into one feature
ONEHOT	Aggregates given binary features into one-hot vector
MEAN	Take the mean value across given numeric features and converts to one feature
GROUPCONDITION	Categorizes 9 binary features corresponding to family history prevalence into 3 separate binary features: spontaneous, indicated, and fetal conditions
ALCSUM	Combined drinking days/week and drinks/day into a single numeric feature
ALCSCALE1	Combines drinking days/week and drinks/day for one binary feature representing > 7 drinks per week
CALC_POLYHYDRAMNIOS	Sum up the 4 quadrants
WEEKSUM	Converts three features (years, months, and weeks) into one weeks feature: $52 * years + 4 * months + weeks$
GROUPCOUNT	Creates 3 scores among spontaneous, indicated, and fetal conditions for family history risk factors by summing up the following prevalance: 1 for mother, 0.75 for sister, 0.5 for half-sisters, 0.25 for cousins
QUITSMOKE	Combines smoking cessation features into a single binary feature for attempted to quit smoking
SECONDHAND	Combines secondhand smoke exposure features into a single binary yes/no feature for exposure
SMOKESCALE	Creates 3 categories of cigarettes/day: 0 cpd, 1 to 19 cpd, or 20 cpd
DRUGSCALE	Nonprescribed stimulants (cocaine, amphetamines), nonprescribed depressants (narcotics, heroin all types), methadone (note this is not illicit), other nonprescribed (inhalants, hallucinogens)
GEST_AGE	Computes the average age of gestational loss for the first two pregnancies
SLEEP_AVG	Computes the average hours of sleep per night combining across weekdays and weekends
ADRENAL_MEASURE	Computes adrenal gland measurements based on ultrasound data, either averaging original and repeated measurements or using whichever is available

- When there is parent feature and there are several related child features that are too detailed, DROP the child features.

5 Timeline

This timeline is meant to serve as a high level representation of the way we processed and organized data for internal analysis and modeling. This timeline does not cover all variables and is thus not an exhaustive representation of nuMoM2b. It serves rather as a quick glance into the data currently in use, accurate at time of publication, but subject to change. Each internal substudy may use a different collection of variables for their goals.

This timeline shows the existence of six unique time points that nuMoM2b represents. 5 of those time points are directly sequential, namely **Before Pregnancy**, **Visit 1**, **Visit 2**, **Visit 3**, **Visit 4 (Delivery Visit)**, and **Post Delivery**. There also exists a **Constant** time point, which groups data that may have been collected at different points in the study, but applies at all times to the patient, such as race and pregnancy history. In an ideal full-term pregnancy, the patient would pass through all these time points, and have data recorded under the Constant point. However, if a birth was preterm, still, or a patient missed a visit, their timeline could skip some of either Visit 1, 2 or 3, and move straight to Visit 4, which represents data collected at delivery. This is represented

Table 2: List of Imputation Rules and their Descriptions

Imputation Name	**Description**
MEAN	Imputes with the mean (average) of the existing values
BMICAT	Imputes based on exising BMI
ONEHOT	Encodes current values by assigning each unique value to an index and assigns a vector of zeros for missing values
ANY	Imputes with a value of True if any of given features is present: False otherwise
MULTI	Imputes with a value of True if any of given features is present: False otherwise (*different internal organization, outcome same as ANY*)
SINGLE	Imputes with a value of True if a given feature is present: False otherwise
NEG1	Imputes missing values with -1
NUMERIC	Imputes missing values with the given numeric value
MODE	Imputes with the mode of the existing values
SUM_LEQ	Imputes with a value of True for Oligohydramnios if sum of Amniotic Fluid Index across 4 quadrants < 5: False otherwise

by the dashed arrow in the timeline. Variables in the timeline are organized by the Filtering Groups described earlier, and are abstractly summarized and simplified in the tables below. On the bottom right we see two boxed sections. These represent time points that occurred in between time points in this timeline. Antepartum evaluations may have occurred between visits, and enrollment screening occurred before visit 1. ***To reiterate, this timeline is not representative of all data available or used, but should rather serve as a quick guide.***

Table 3: Amount of patients available at each visit. *Present* numbers are patients who filled out the main maternal interview form. Patients designated as other / missing may still have information available at that visit, especially under Chart Abstractions. **Delivery outcomes are available for almost every patient except those withdrawn.** *Withdrawn* patient counts were approximated using the interval at form A05 recording (official withdrawal). *Other / Missing* represents those not having taken the main maternal interview form, (V#A)

	Visit 1	**Visit 2**	**Visit 3**	**Visit 4 (Delivery)**	**Post Delivery**
Present ($V\#A$)	10,028	9,412	9,217	7,167	9,430
Withdrawn	7	167	213	276	62
Other or Missing	2	407	489	2,595	608

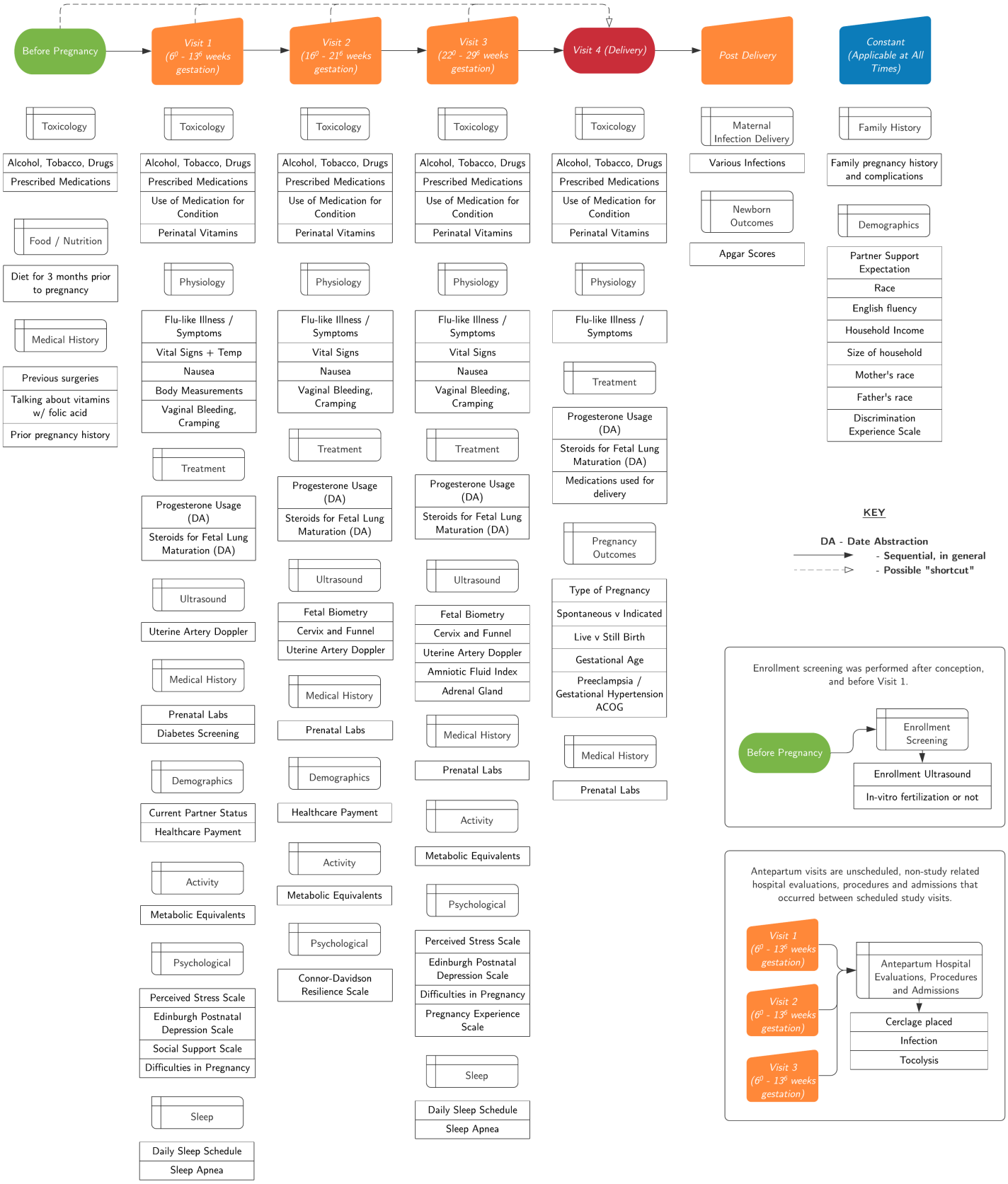

Figure 3: nuMoM2b Processed Data Timeline

6 Filtering Group Breakdown

The following is a breakdown of the filtering groups our team decided upon at the time of writing. These groups are subject to change in name and organization, and simply serve to understand and organize the data in a simpler and more manageable fashion. Each group below contains a description of itself, the total number of features marked belonging to said group, the number of features used in layer 1 and layer 2, the files which comprise the group, and a general description of the features that were dropped from out modeling. Following each description, shown is a table of the features both compiled and used as-is. Features used as-is are at the top of the table. Compiled features – those constructed from other variables in the data – follow, and are surrounded by horizontal lines and bolded. Those below a compiled variable are used to construct it, using the rule listed next to the compiled feature. The column **NAME** represents the original or compiled variable name. **FILE** represents the file in the data from which the original feature comes from. **RULE** represents which filtering rule was used to compile a group of features. **TEMPORAL** represents at which point in the timeline this feature is relevant. TEMPORAL ranges from -1 to 5, where 1 to 4 represent Visits 1 to 4 (delivery), -1 represents "applies at all time", 0 represents before pregnancy, and 5 represents post-pregnancy. **IMPUTE** represents which rule was used for imputation. **MISSING** represents the missingness value for each original feature. **DESCRIPTION** is a shortened description from the original data set of each variable.

6.1 Family History

The family history filtering compiled all of the questions regarding diagnoses of family members related to diabetes, blood clotting disorders, pregnancy complications, heart disease, and hypertension. The 9 pregnancy-related conditions were grouped into 3 categories:

Spontaneous

- Early or preterm rupture of the membranes
- Spontaneous preterm delivery (less than 37 weeks)

Indicated

- Delivery of a child more than 3 weeks before the expected due date
- Preeclampsia, eclampsia, toxemia or pregnancy-induced hypertension

Fetal Conditions

- Delivery of a child weighing less than 5 lb 8 oz (or 2500 grams)
- Stillbirth
- Delivery of an infant with a birth defect
- Other pregnancy complication

Layer 1 includes binary features indicating any presence of family history in the 3 categories. Layer 2 compiled a more detailed score in the 3 categories which aggregates a score based on the genetic proximity to the stated family member, based on the following scale: 1 for mother, 0.75 for sister, 0.5 for half-sisters or cousins.

Total # Features: 113
Layer 1 # Features: 3
Layer 2 # Features: 3
Relevant Files: V2A
Dropped Features: We chose to discard family history related to diabetes, blood clotting disorders, heart disease, and hypertension, in order to focus on pregnancy-related conditions. Family history does NOT include the patient's own medical history.

Table 4: Family History Filtering

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V2A_spontaneous_risk, V2A_indicated_risk, V2A_fetal_condition_risk	V2A	GROUPCONDITION	-1	GROUPMODE		
V2AE06a	V2A		2		0.0395	(V2A) Have any of your biological mother, sisters, half-sisters, or female first cousins ever had: Delivery of a child more than 3 weeks before the expected due date
V2AE06b	V2A		2		0.0414	(V2A) ' ' Delivery of a child weighing less than 5 lb 8 oz (or 2500 grams)
V2AE06c	V2A		2		0.044	(V2A) ' ' Spontaneous preterm delivery (<37 weeks)
V2AE06d	V2A		2		0.0639	(V2A) ' ' Early or preterm rupture of the membranes
V2AE06e	V2A		2		0.0542	(V2A) ' ' Preeclampsia, eclampsia, toxemia or pregnancy induced hypertension
V2AE06f	V2A		2		0.0507	(V2A) ' ' Induction of labor due to low amniotic fluid or poor fetal growth
V2AE06g	V2A		2		0.0268	(V2A) ' ' Stillbirth
V2AE06h	V2A		2		0.0227	(V2A) ' ' Delivery of an infant with a birth defect
V2AE06i	V2A		2		0.0266	(V2A) ' ' Other pregnancy complication, specify
V2A_spontaneous_score, V2A_indicated_score, V2A_fetal_condition_score	V2A	GROUPCONDITION	-1	GROUPMEAN		
V2AE07_1a	V2A		2		0.6137	(V2A) Family members with pregnancy complications - Family relation (1)
V2AE07_1b1	V2A		2		0.6142	(V2A) ' ' Pregnancy complication 1 (1)
V2AE07_1b2	V2A		2		0.8259	(V2A) ' ' Pregnancy complication 2 (1)
V2AE07_1b3	V2A		2		0.9062	(V2A) ' ' Pregnancy complication 3 (1)
V2AE07_1b4	V2A		2		0.9629	(V2A) ' ' Pregnancy complication 4 (1)
V2AE07_1b5	V2A		2		0.9902	(V2A) ' ' Pregnancy complication 5 (1)
V2AE07_1c1	V2A		2		0.93	(V2A) ' ' Pregnancy complication 6, specify (1)
V2AE07_1c2	V2A		2		0.9939	(V2A) ' ' Pregnancy complication 7, specify (1)
V2AE07_2a	V2A		2		0.9261	(V2A) ' ' Family relation (2)
V2AE07_2b1	V2A		2		0.9269	(V2A) ' ' Pregnancy complication 1 (2)
V2AE07_2b2	V2A		2		0.9833	(V2A) ' ' Pregnancy complication 2 (2)
V2AE07_2b3	V2A		2		0.9925	(V2A) ' ' Pregnancy complication 3 (2)
V2AE07_2b4	V2A		2		0.9973	(V2A) ' ' Pregnancy complication 4 (2)
V2AE07_2b5	V2A		2		0.9993	(V2A) ' ' Pregnancy complication 5 (2)
V2AE07_2c1	V2A		2		0.9873	(V2A) ' ' Pregnancy complication 6, specify (2)
V2AE07_2c2	V2A		2		0.9982	(V2A) ' ' Pregnancy complication 7, specify (2)
V2AE07_3a	V2A		2		0.9849	(V2A) ' ' Family relation (3)
V2AE07_3b1	V2A		2		0.985	(V2A) ' ' Pregnancy complication 1 (3)
V2AE07_3b2	V2A		2		0.9985	(V2A) ' ' Pregnancy complication 2 (3)
V2AE07_3b3	V2A		2		0.9993	(V2A) ' ' Pregnancy complication 3 (3)
V2AE07_3b4	V2A		2		0.9998	(V2A) ' ' Pregnancy complication 4 (3)
V2AE07_3b5	V2A		2		1	(V2A) ' ' Pregnancy complication 5 (3)
V2AE07_3c1	V2A		2		0.9967	(V2A) ' ' Pregnancy complication 6, specify (3)
V2AE07_3c2	V2A		2		0.9996	(V2A) ' ' Pregnancy complication 7, specify (3)
V2AE07_4a	V2A		2		0.9964	(V2A) ' ' Family relation (4)
V2AE07_4b1	V2A		2		0.9964	(V2A) ' ' Pregnancy complication 1 (4)
V2AE07_4b2	V2A		2		0.9999	(V2A) ' ' Pregnancy complication 2 (4)
V2AE07_4b3	V2A		2		1	(V2A) ' ' Pregnancy complication 3 (4)
V2AE07_4b4	V2A		2		1	(V2A) ' ' Pregnancy complication 4 (4)
V2AE07_4b5	V2A		2		1	(V2A) ' ' Pregnancy complication 5 (4)
V2AE07_4c1	V2A		2		0.9995	(V2A) ' ' Pregnancy complication 6, specify (4)
V2AE07_4c2	V2A		2		1	(V2A) ' ' Pregnancy complication 7, specify (4)
V2AE07_5a	V2A		2		0.9992	(V2A) ' ' Family relation (5)
V2AE07_5b1	V2A		2		0.9992	(V2A) ' ' Pregnancy complication 1 (5)
V2AE07_5b2	V2A		2		1	(V2A) ' ' Pregnancy complication 2 (5)
V2AE07_5b3	V2A		2		1	(V2A) ' ' Pregnancy complication 3 (5)
V2AE07_5b4	V2A		2		1	(V2A) ' ' Pregnancy complication 4 (5)
V2AE07_5b5	V2A		2		1	(V2A) ' ' Pregnancy complication 5 (5)
V2AE07_5c1	V2A		2		1	(V2A) ' ' Pregnancy complication 6, specify (5)
V2AE07_5c2	V2A		2		0.9999	(V2A) ' ' Pregnancy complication 7, specify (5)

6.2 Toxicology

The toxicology filtering relates to questions and chart abstractions regarding any kind of drug use, prescription or non-prescription, by the patient. This includes questions regarding alcohol, tobacco usage, second-hand smoke, smoking cessation, illegal drugs, all prescription medications noted, vitamins, vaccines, and whether medication was being taken for a particular condition. The information used comes from both patient interviews and ancillary files, which corrected the medication information through consensus and chart review.

Total # Features: 806
Layer 1 # Features: 63
Layer 2 # Features: 63
Relevant Files: drugs_in_pregnancy (ancillary), VXX, V1A, V2A, V3A, V4A
Dropped Features: General questions on if a class of drug (non-prescribed) was used were kept in L1. Some detail was saved for L2, such as illegal drug use breakdown. Prescription drugs were organized back into drug categories originally used by the nuMoM2b team in order to reduce dimensionality (see form VXX Section C), perinatal vitamins were separated from the vitamins group due to feature compiled feature *Vitamin_Multi_Perinatal_Folate*, covering that information as well. Metadata features were used to organize and preprocess data, and then dropped from modeling. "Medication taken for condition" features in VXX are dropped as prescribed drugs (as mentioned earlier) and "condition noted" features from VXX in medical history filtering cover said information.

Table 5: Toxicology Filtering

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V1AG01	V1A		1	999	0.0007	(V1A) Have you ever drunk alcohol?
V1AG04	V1A		1	999	0.0009	(V1A) Have you ever used any tobacco products including cigarettes and smokeless tobacco?
V1AG05	V1A		1	999	0.5825	(V1A) Did you smoke any tobacco products in the three months prior to this pregnancy?
V1AG07	V1A		1	999	0.584	(V1A) Did you smoke any tobacco products in the last month?
V1AG11	V1A		1	999	0.0013	(V1A) Have you ever used illegal drugs or drugs not prescribed for you?
V2AH02	V2A		2	999	0.0007	(V2A) Did you smoke any tobacco products in the last month?
V2AH06	V2A		2	999	0.0008	(V2A) Have you used illegal drugs or drugs not prescribed for you in the last month?
V3AF02	V3A		3	999	0.0001	(V3A) Did you smoke any tobacco products in the last month?
V3AF06	V3A		3	999	0.0005	(V3A) Have you used illegal drugs or drugs not prescribed for you in the last month?
V4AF02	V4A		4	999	0.0011	(V4A) Did you smoke any tobacco products in the month before your delivery?
V4AF06	V4A		4	999	0.001	(V4A) Have you used illegal drugs or drugs not prescribed for you in the month before delivery?
DrugName	drugs_in_pregnancy (ancillary)		VXXC01g, VXXC01h	N/A	0	Drug Name corrected from VXXC01b value provided
DrugCode	drugs_in_pregnancy (ancillary)		VXXC01g, VXXC01h	N/A	0	Drug Code corrected from VXXC01c value provided
ReasonCode	drugs_in_pregnancy (ancillary)		VXXC01g, VXXC01h	N/A	0	Reason Code corrected from VXXC01e value provided
VXXC01g	drugs_in_pregnancy (ancillary)		Itself	N/A	0.0251	(VXX) Medications and vaccinations - Start timing
VXXC01h	drugs_in_pregnancy (ancillary)		Itself	N/A	0.5536	(VXX) Medications and vaccinations - Stopped timing
V1AG05a	V1A		0	0	0.8234	(V1A) How many cigarettes did you smoke per day in the three months prior to this pregnancy? - # per day,
V1AG07a	V1A		1	0	0.9413	(V1A) How many cigarettes did you smoke per day in the last month? - # per day,
V2AH02a	V2A		2	0	0.9471	(V2A) How many cigarettes did you smoke per day? - # per day
V3AF02a	V3A		3	0	0.9532	(V3A) How many cigarettes did you smoke per day? - # per day
V4AF02a	V4A		4	0	0.9618	(V4A) In the month before delivery - How many cigarettes did you smoke per day? # per day
V1AG10	V1A		1	0	0.5834	(V1A) In the last month, did you use smokeless tobacco (chew or snuff)?,

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V2AH05	V2A		2	0	0.0045	(V2A) In the last month, did you use smokeless tobacco (chew or snuff)?
V3AF05	V3A		3	0	0.004	(V3A) In the last month, did you use smokeless tobacco (chew or snuff)?
V4AF05	V4A		4	0	0.0045	(V4A) In the month before your delivery - Did you use smokeless tobacco (chew or snuff)?
drug_category	drugs_in_pregnancy (ancillary)	drug_cat_creation	VXXC01g, VXXC01h	N/A	0	Compilation of DrugCode's into the respective medication categories as specified in VXX Section C.
DrugCode	drugs_in_pregnancy (ancillary)		VXXC01g, VXXC01h		0	Drug Code corrected from VXXC01c value provided
VXXC01k	drugs_in_pregnancy (ancillary)		N/A		0	(VXX) Medications and vaccinations - Final assessment: Took medication/vaccine
Alc_Usage_V0		AlcSum	0	999		
V1AG02a	V1A		0		0.3559	(V1A) How many days per week did you drink in the three months prior to this pregnancy? - days/week
V1AG02b	V1A		0		0.3558	(V1A) How many drinks did you drink per drinking day in the three months prior to this pregnancy? - drinks
Alc_Usage_V1		AlcSum	1	999		
V1AG03a	V1A		1		0.9621	(V1A) How many days per week did you drink in the last month? - days/week
V1AG03b	V1A		1		0.9621	(V1A) How many drinks did you drink per drinking day in the last month? - drinks
Quit_Smoke_V1		QuitSmoke	1			
V1AG08	V1A		1	2	0.5833	(V1A) In the last month, did you use nicotine gum, a nicotine patch, a nicotine spray or a nicotine inhaler?
V1AG09	V1A		1	2	0.5833	(V1A) In the last month, did you take a pill like Zyban (also known as Wellbutrin or Bupropion) to help you quit smoking?
SecH_Smoke_V1		SecondHand	1			
V1AG07b	V1A		1	2	0.4472	(V1A) In the last month, on average how many hours per week were you exposed to cigarette smoke because of smoking...
V1AG07c	V1A		1	2	0.4463	(V1A) In the last month, how many people (excluding yourself) smoked cigarettes inside the home...
Drug_Usage_V1		DrugScale	1	999		
V1AG12a	V1A		1		0.6612	(V1A) Every used any of these drugs - Marijuana (THC)
V1AG12b	V1A		1		0.6622	(V1A) Every used any of these drugs - Cocaine
V1AG12c	V1A		1		0.6636	(V1A) Every used any of these drugs - Prescription narcotics that were not prescribed for you
V1AG12d	V1A		1		0.6627	(V1A) Every used any of these drugs - Heroin
V1AG12d1	V1A		1		0.9899	(V1A) Every used any of these drugs - Heroin, Snorted
V1AG12d2	V1A		1		0.99	(V1A) Every used any of these drugs - Heroin, Injected
V1AG12d3	V1A		1		0.9907	(V1A) Every used any of these drugs - Heroin, Smoked
V1AG12d4	V1A		1		0.9906	(V1A) Every used any of these drugs - Heroin, Other
V1AG12d4_SP	V1A		1		0.9999	(V1A) Every used any of these drugs - Heroin, Other, specify
V1AG12e	V1A		1		0.6692	(V1A) Every used any of these drugs - Methadone
V1AG12f	V1A		1		0.666	(V1A) Every used any of these drugs - Amphetamines (speed) not prescribed for you
V1AG12g	V1A		1		0.6665	(V1A) Every used any of these drugs - Inhalants not prescribed for you
V1AG12h	V1A		1		0.6658	(V1A) Every used any of these drugs - Hallucinogens
V1AG12i	V1A		1		0.6702	(V1A) Every used any of these drugs - Other
V1AG12i_SP	V1A		1		0.9765	(V1A) Every used any of these drugs - Other , specify
Alc_Usage_V2		AlcSum	2	999		
V2AH01a	V2A		2		0.9439	(V2A) How many days per week did you drink? - days/week
V2AH01b	V2A		2		0.9443	(V2A) How many drinks did you drink per drinking day? - drinks
Quit_Smoke_V2		QuitSmoke	2			
V2AH03	V2A		2	2	0.0046	(V2A) In the last month, did you use nicotine gum, a nicotine patch, a nicotine spray or a nicotine inhaler?
V2AH04	V2A		2	2	0.0044	(V2A) In the last month, did you take a pill like Zyban (also known as Wellbutrin or Bupropion) to help you quit smoking?

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
SecH_Smoke_V2		SecondHand	2			
V2AH02b	V2A		2	2	0.0045	(V2A) In the last month, on average how many hours per week were you exposed to cigarette smoke because of smoking...
V2AH02c	V2A		2	2	0.0008	(V2A) In the last month, how many people (excluding yourself) smoked cigarettes inside the home...
Drug_Usage_V2		DrugScale	2	999		
V2AH07a	V2A		2		0.9888	(V2A) Drug use in the past month - Marijuana (THC)
V2AH07b	V2A		2		0.989	(V2A) Drug use in the past month - Cocaine
V2AH07c	V2A		2		0.989	(V2A) Drug use in the past month - Prescription narcotics that were not prescribed for you
V2AH07d	V2A		2		0.989	(V2A) Drug use in the past month - Heroin
V2AH07d1	V2A		2		0.9998	(V2A) Drug use in the past month - Heroin, Snorted
V2AH07d2	V2A		2		0.9998	(V2A) Drug use in the past month - Heroin, Injected
V2AH07d3	V2A		2		0.9998	(V2A) Drug use in the past month - Heroin, Smoked
V2AH07d4	V2A		2		0.9998	(V2A) Drug use in the past month - Heroin, Other
V2AH07d4_SP	V2A		2		1	(V2A) Drug use in the past month - Heroin, Other, specify
V2AH07e	V2A		2		0.9891	(V2A) Drug use in the past month - Methadone
V2AH07f	V2A		2		0.9891	(V2A) Drug use in the past month - Amphetamines (speed) not prescribed for you
V2AH07g	V2A		2		0.9891	(V2A) Drug use in the past month - Inhalants not prescribed for you
V2AH07h	V2A		2		0.9891	(V2A) Drug use in the past month - Hallucinogens
V2AH07i	V2A		2		0.9891	(V2A) Drug use in the past month - Other
V2AH07i_SP	V2A		2		0.9998	(V2A) Drug use in the past month - Other, specify
Alc_Usage_V3		AlcSum	3	999		
V3AF01a	V3A		3		0.9303	(V3A) How many days per week did you drink? - days/week
V3AF01b	V3A		3		0.9308	(V3A) How many drinks did you drink per drinking day? - drinks
Quit_Smoke_V3		QuitSmoke	3			
V3AF03	V3A		3	2	0.004	(V3A) In the last month, did you use nicotine gum, a nicotine patch, a nicotine spray or a nicotine inhaler?
V3AF04	V3A		3	2	0.0041	(V3A) In the last month, did you take a pill like Zyban (also known as Wellbutrin or Bupropion) to help you quit smoking?
SecH_Smoke_V3		SecondHand	3			
V3AF02b	V3A		3	2	0.3266	(V3A) In the last month, on average how many hours per week were you exposed to cigarette smoke because of smoking...
V3AF02c	V3A		3	2	0.3262	(V3A) In the last month, how many people (excluding yourself) smoked cigarettes inside the home...
Drug_Usage_V3		DrugScale	3	999		
V3AF07a	V3A		3		0.9932	(V3A) Drug use in the past month - Marijuana (THC)
V3AF07b	V3A		3		0.9932	(V3A) Drug use in the past month - Cocaine
V3AF07c	V3A		3		0.9932	(V3A) Drug use in the past month - Prescription narcotics that were not prescribed for you
V3AF07d	V3A		3		0.9932	(V3A) Drug use in the past month - Heroin
V3AF07d1	V3A		3		1	(V3A) Drug use in the past month - Heroin, Snorted
V3AF07d2	V3A		3		1	(V3A) Drug use in the past month - Heroin, Injected
V3AF07d3	V3A		3		1	(V3A) Drug use in the past month - Heroin, Smoked
V3AF07d4	V3A		3		1	(V3A) Drug use in the past month - Heroin, Other
V3AF07d4_SP	V3A		3		1	(V3A) Drug use in the past month - Heroin, Other, specify
V3AF07e	V3A		3		0.9932	(V3A) Drug use in the past month - Methadone
V3AF07f	V3A		3		0.9932	(V3A) Drug use in the past month - Amphetamines (speed) not prescribed for you
V3AF07g	V3A		3		0.9932	(V3A) Drug use in the past month - Inhalants not prescribed for you
V3AF07h	V3A		3		0.9932	(V3A) Drug use in the past month - Hallucinogens
V3AF07i	V3A		3		0.9932	(V3A) Drug use in the past month - Other
V3AF07i_SP	V3A		3		0.9998	(V3A) Drug use in the past month - Other, specify
Alc_Usage_V4		AlcSum	4	999		
V4AF01a	V4A		4		0.9399	(V4A) In the month before delivery - How many days per week did you drink? days/week
V4AF01b	V4A		4		0.94	(V4A) In the month before delivery - How many drinks did you drink per drinking day?

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
Quit_Smoke_V4		QuitSmoke	4			
V4AF03	V4A		4	2	0.0045	(V4A) In the month before your delivery - Did you use nicotine gum, a nicotine patch, a nicotine spray or a nicotine inhaler?
V4AF04	V4A		4	2	0.0045	(V4A) In the month before your delivery - Did you take a pill like Zyban (also known as Wellbutrin or Bupropion) to help you quit smoking?
SecH_Smoke_V4		SecondHand	4			
V4AF02b	V4A		4	2	0.2594	(V4A) In the last month, on average how many hours per week were you exposed to cigarette smoke because of smoking...
V4AF02c	V4A		4	2	0.2591	(V4A) In the last month, how many people (excluding yourself) smoked cigarettes inside the home...
Drug_Usage_V4		DrugScale	4	999		
V4AF07a	V4A		4		0.9964	(V4A) Drug use in the last month - Marijuana (THC)
V4AF07b	V4A		4		0.9964	(V4A) Drug use in the last month - Cocaine
V4AF07c	V4A		4		0.9964	(V4A) Drug use in the last month - Prescription narcotics that were not prescribed for you
V4AF07d	V4A		4		0.9964	(V4A) Drug use in the last month - Heroin
V4AF07d1	V4A		4		0.9996	(V4A) Drug use in the last month - Heroin, Snorted
V4AF07d2	V4A		4		0.9996	(V4A) Drug use in the last month - Heroin, Injected
V4AF07d3	V4A		4		0.9996	(V4A) Drug use in the last month - Heroin, Smoked
V4AF07d4	V4A		4		0.9996	(V4A) Drug use in the last month - Heroin, Other
V4AF07d4_SP	V4A		4		1	(V4A) Drug use in the last month - Heroin, Other, specify
V4AF07e	V4A		4		0.9964	(V4A) Drug use in the last month - Methadone
V4AF07f	V4A		4		0.9964	(V4A) Drug use in the last month - Amphetamines (speed) not prescribed for you
V4AF07g	V4A		4		0.9964	(V4A) Drug use in the last month - Inhalants not prescribed for you
V4AF07h	V4A		4		0.9964	(V4A) Drug use in the last month - Hallucinogens
V4AF07i	V4A		4		0.9964	(V4A) Drug use in the last month - Other
V4AF07i_SP	V4A		4		0.9997	(V4A) Drug use in the last month - Other , specify
Vitamin_Multi_Perinatal_Folate		WeekSum	N/A	Mean		
VXXC02a	VXX		N/A		0.7405	(VXX) How long have you been taking [perinatal vitamins/other multivitamins/ folate supplements]? - wks
VXXC02b	VXX		N/A		0.499	(VXX) How long have you been taking [perinatal vitamins/other multivitamins/ folate supplements]? - months
VXXC02c	VXX		N/A		0.8739	(VXX) How long have you been taking [perinatal vitamins/other multivitamins/ folate supplements]? - years

Toxicology Filtering

6.3 Psychological

The psychological filtering compiled all of the features relating to psychological health and wellbeing, including previous treatment for mental health conditions in addition to surveys about factors related to depression, stress, resiliency and social support at the 3 visits.

Total # Features: 172
Layer 1 # Features: 8
Layer 2 # Features: 8
Relevant Files: CMA, V1A, V2A, V1C, V1E, V1G, V1H, V2I, V3A, V3C, V3E, V3J
Dropped Features: We chose to discard features related to prenatal and postpartum treatment for mental health conditions, in the case that some conditions were not diagnosed, to focus on the current state of mind throughout the pregnancy as measured by the surveys.

Table 6: Psychological Filtering

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V1EA01	V1E		1	ONEHOT	0.0887	(V1E) Are you feeling bothered, upset or worried at this point in your pregnancy?
V3EA01	V3E		3	ONEHOT	0.0826	(V3E) Are you feeling bothered, upset or worried at this point in your pregnancy?
V1A_STRESS_SUMMARY	V1A	ADD	1	MEAN		
V1AH01	V1A		1		0.0039	(V1A) In the last month, how often have you been upset because of something that happened unexpectedly?
V1AH02	V1A		1		0.004	(V1A) In the last month, how often have you felt that you were unable to control the important things in your life?
V1AH03	V1A		1		0.0043	(V1A) In the last month, how often have you felt nervous and 'stressed'?
V1AH04	V1A		1		0.0041	(V1A) In the last month, how often have you felt confident about your ability to handle your personal problems?
V1AH05	V1A		1		0.0045	(V1A) In the last month, how often have you felt that things were going your way?
V1AH06	V1A		1		0.0041	(V1A) In the last month, how often have you found that you could not cope with all the things that you had to do?
V1AH07	V1A		1		0.0048	(V1A) In the last month, how often have you been able to control irritations in your life?
V1AH08	V1A		1		0.0048	(V1A) In the last month, how often have you felt that you were on top of things?
V1AH09	V1A		1		0.0045	(V1A) In the last month, how often have you been angered because of things that were outside of your control?
V1AH10	V1A		1		0.0043	(V1A) In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V1C_DEPRESSION_SUMMARY	V1C	ADD	1	MEAN		
V1CA01	V1C		1		0.0003	(V1C) Edinburgh Postnatal Depression Scale - I have been able to laugh and see the funny side of things
V1CA02	V1C		1		0.0004	(V1C) Edinburgh Postnatal Depression Scale - I have looked forward with enjoyment to things.
V1CA03	V1C		1		0.0005	(V1C) Edinburgh Postnatal Depression Scale - I have blamed myself unnecessarily when things went wrong.
V1CA04	V1C		1		0.0006	(V1C) Edinburgh Postnatal Depression Scale - I have been anxious or worried for no good reason.
V1CA05	V1C		1		0.0003	(V1C) Edinburgh Postnatal Depression Scale - I have felt scared or panicky for no very good reason.
V1CA06	V1C		1		0.0041	(V1C) Edinburgh Postnatal Depression Scale - Things have been getting on top of me.
V1CA07	V1C		1		0.004	(V1C) Edinburgh Postnatal Depression Scale - I have been so unhappy that I have had difficulty sleeping.
V1CA08	V1C		1		0.004	(V1C) Edinburgh Postnatal Depression Scale - I have felt sad or miserable.
V1CA09	V1C		1		0.004	(V1C) Edinburgh Postnatal Depression Scale - I have been so unhappy that I have been crying.
V1CA10	V1C		1		0.0038	(V1C) Edinburgh Postnatal Depression Scale - The thought of harming myself has occurred to me.
V1G_SOCIAL_SCALE	V1G	ADD	1	MEAN		
V1GA01	V1G		1		0.0008	(V1G) Social Support - There is a special person who is around when I am in need
V1GA02	V1G		1		0.0005	(V1G) Social Support - There is a special person with whom I can share my joys and sorrows
V1GA03	V1G		1		0.0019	(V1G) Social Support - My family really tries to help me
V1GA04	V1G		1		0.0008	(V1G) Social Support - I get the emotional help and support I need from my family
V1GA05	V1G		1		0.0009	(V1G) Social Support - I have a special person who is a real source of comfort to me
V1GA06	V1G		1		0.0006	(V1G) Social Support - My friends really try to help me
V1GA07	V1G		1		0.001	(V1G) Social Support - I can count on my friends when things go wrong
V1GA08	V1G		1		0.0006	(V1G) Social Support - I can talk about my problems with my family
V1GA09	V1G		1		0.0013	(V1G) Social Support - I have friends with whom I can share my joys and sorrows
V1GA10	V1G		1		0.0007	(V1G) Social Support - There is a special person in my life who cares about my feelings
V1GA11	V1G		1		0.0013	(V1G) Social Support - My family is willing to help me make decisions
V1GA12	V1G		1		0.0006	(V1G) Social Support - I can talk about my problems with my friends

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V2I_RESELIENCE_SUMMARY	V2I	ADD	2	MEAN		
V2IA01	V2I		2		0.0006	(V2I) Able to adapt to change
V2IA02	V2I		2		0.0004	(V2I) Close and secure relationships
V2IA03	V2I		2		0.0021	(V2I) Sometimes fate or God can help
V2IA04	V2I		2		0.0004	(V2I) Can deal with whatever comes
V2IA05	V2I		2		0.0015	(V2I) Past success gives confidence for new challenge
V2IA06	V2I		2		0.0009	(V2I) See the humorous side of things
V2IA07	V2I		2		0.001	(V2I) Coping with stress strengthens
V2IA08	V2I		2		0.0008	(V2I) Tend to bounce back after illness or hardship
V2IA09	V2I		2		0.0013	(V2I) Things happen for a reason
V2IA10	V2I		2		0.0004	(V2I) Best effort no matter what
V2IA11	V2I		2		0.0006	(V2I) You can achieve your goals
V2IA12	V2I		2		0.0004	(V2I) When things look hopeless, I don't give up
V2IA13	V2I		2		0.0006	(V2I) Know where to turn for help
V2IA14	V2I		2		0.0006	(V2I) Under pressure, focus and think clearly
V2IA15	V2I		2		0.0059	(V2I) Prefer to take the lead in problem solving
V2IA16	V2I		2		0.0063	(V2I) Not easily discouraged by failure
V2IA17	V2I		2		0.0083	(V2I) Think of self as strong person
V2IA18	V2I		2		0.0064	(V2I) Make unpopular or difficult decisions
V2IA19	V2I		2		0.0062	(V2I) Can handle unpleasant feelings
V2IA20	V2I		2		0.0063	(V2I) Have to act on a hunch
V2IA21	V2I		2		0.0062	(V2I) Strong sense of purpose
V2IA22	V2I		2		0.0061	(V2I) In control of your life
V2IA23	V2I		2		0.0063	(V2I) I like challenges
V2IA24	V2I		2		0.006	(V2I) You work to attain your goals
V2IA25	V2I		2		0.0061	(V2I) Pride in your achievements
V3A_STRESS_SUMMARY	V3A	ADD	3	MEAN		
V3AG01	V3A		3		0.0016	(V3A) Perceived Stress Scale - In the last month, how often have you been upset because of something that happened unexpectedly?
V3AG02	V3A		3		0.0014	(V3A) Perceived Stress Scale - In the last month, how often have you felt that you were unable to control the important things in your life?
V3AG03	V3A		3		0.0014	(V3A) Perceived Stress Scale - In the last month, how often have you felt nervous and 'stressed'?
V3AG04	V3A		3		0.0016	(V3A) Perceived Stress Scale - In the last month, how often have you felt confident about your ability to handle your personal problems?
V3AG05	V3A		3		0.0016	(V3A) Perceived Stress Scale - In the last month, how often have you felt that things were going your way?
V3AG06	V3A		3		0.0017	(V3A) Perceived Stress Scale - In the last month, how often have you found that you could not cope with all the things that you had to do?
V3AG07	V3A		3		0.0028	(V3A) Perceived Stress Scale - In the last month, how often have you been able to control irritations in your life?
V3AG08	V3A		3		0.002	(V3A) Perceived Stress Scale - In the last month, how often have you felt that you were on top of things?
V3AG09	V3A		3		0.0015	(V3A) Perceived Stress Scale - In the last month, how often have you been angered because of things that were outside of your control?
V3AG10	V3A		3		0.0015	(V3A) Perceived Stress Scale - In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V3C_DEPRESSION_SUMMARY	V3C	ADD	3	MEAN		
V3CA01	V3C		3		0.0007	(V3C) Edinburgh Postnatal Depression Scale - I have been able to laugh and see the funny side of things
V3CA02	V3C		3		0.0007	(V3C) Edinburgh Postnatal Depression Scale - I have looked forward with enjoyment to things.
V3CA03	V3C		3		0.0007	(V3C) Edinburgh Postnatal Depression Scale - I have blamed myself unnecessarily when things went wrong.
V3CA04	V3C		3		0.0008	(V3C) Edinburgh Postnatal Depression Scale - I have been anxious or worried for no good reason.
V3CA05	V3C		3		0.0009	(V3C) Edinburgh Postnatal Depression Scale - I have felt scared or panicky for no very good reason.
V3CA06	V3C		3		0.0064	(V3C) Edinburgh Postnatal Depression Scale - Things have been getting on top of me.
V3CA07	V3C		3		0.006	(V3C) Edinburgh Postnatal Depression Scale - I have been so unhappy that I have had difficulty sleeping.
V3CA08	V3C		3		0.0059	(V3C) Edinburgh Postnatal Depression Scale - I have felt sad or miserable.
V3CA09	V3C		3		0.0059	(V3C) Edinburgh Postnatal Depression Scale - I have been so unhappy that I have been crying.
V3CA10	V3C		3		0.0059	(V3C) Edinburgh Postnatal Depression Scale - The thought of harming myself has occurred to me.

Psychological Filtering

6.4 Activity

The activity filtering related to details about participant physical activity prior to and during pregnancy. The activities were measured in METs, or metabolic equivalents, for Layer 1 and expanded upon in further detail, namely activity type, minutes, miles and duration in Layer 2. Activity restriction information available at V3 and V4 was also included, to account for influence on the MET score.

Total # Features: 131
Layer 1 # Features: 3
Layer 2 # Features: 22
Relevant Files: physical_activity (ancillary), V1A, V2A, V3A, V4A
Dropped Features: We chose to discard features relating to any personal or care provider weight change goals, as these were suggestions that would merely be reflected in the actual activity data.

Table 7: Activity Filtering

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
Visit1_MET	physical_activity		1	mean	0	Visit 1 MET Equivalents
Visit2_MET	physical_activity		2	mean	0	Visit 2 MET Equivalents
Visit3_MET	physical_activity		3	mean	0	Visit 3 MET Equivalents
V1A_ACTIVITY_1	V1A	V1AI01	1	onehot	0.001	During the past 4 weeks, did you participate in any physical activities?
V1A_ACTIVITY_1_code	V1A	V1AI01a	1	onehot	0.2987	(Activity #1) - Code
V1A_ACTIVITY_1_weekly	V1A	V1AI01a1	1	mean	0.2991	How many times per week did you take part in this activity during the past 4 weeks?
V1A_ACTIVITY_1_minutes	V1A	V1AI01a2	1	mean	0.3063	And when you took part in this activity, for how many minutes did you usually keep at it?
V1A_ACTIVITY_1_miles	V1A	V1AI01a3	1	mean	0.6079	How far do you usually walk/run/jog/cycle/swim? (Activity #1) - miles
V1A_ACTIVITY_2	V1A	V1AI02	1	onehot	0.3016	During the past 4 weeks, did you participate in any physical activities?
V1A_ACTIVITY_2_code	V1A	V1AI02a	1	onehot	0.7045	(Activity #2) - Code
V1A_ACTIVITY_2_weekly	V1A	V1AI02a1	1	mean	0.7051	How many times per week did you take part in this activity during the past 4 weeks?
V1A_ACTIVITY_2_minutes	V1A	V1AI02a2	1	mean	0.7074	And when you took part in this activity, for how many minutes did you usually keep at it?
V1A_ACTIVITY_2_miles	V1A	V1AI02a3	1	mean	0.9092	How far do you usually walk/run/jog/cycle/swim? (Activity #2) - miles
V1A_ACTIVITY_3	V1A	V1AI03	1	onehot	0.7083	During the past 4 weeks, did you participate in any physical activities?
V1A_ACTIVITY_3_code	V1A	V1AI03a	1	onehot	0.8957	(Activity #3) - Code
V1A_ACTIVITY_3_weekly	V1A	V1AI03a1	1	mean	0.8961	How many times per week did you take part in this activity during the past 4 weeks?
V1A_ACTIVITY_3_minutes	V1A	V1AI03a2	1	mean	0.8965	And when you took part in this activity, for how many minutes did you usually keep at it?
V1A_ACTIVITY_3_miles	V1A	V1AI03a3	1	mean	0.9733	How far do you usually walk/run/jog/cycle/swim? (Activity #3) - miles
V2A_ACTIVITY_1	V2A	V2AJ01	2	onehot	0.0003	During the past 4 weeks, did you participate in any physical activities?
V2A_ACTIVITY_1_code	V2A	V2AJ01a	2	onehot	0.2581	(Activity #1) - Code
V2A_ACTIVITY_1_weekly	V2A	V2AJ01a1	2	mean	0.2592	How many times per week did you take part in this activity during the past 4 weeks?

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V2A_ACTIVITY_1_minutes	V2A	V2AJ01a2	2	mean	0.2695	And when you took part in this activity, for how many minutes did you usually keep at it?
V2A_ACTIVITY_1_miles	V2A	V2AJ01a3	2	mean	0.5907	How far do you usually walk/run/jog/cycle/swim? (Activity #1) - miles
V2A_ACTIVITY_2	V2A	V2AJ02	2	onehot	0.258	During the past 4 weeks, did you participate in any physical activities?
V2A_ACTIVITY_2_code	V2A	V2AJ02a	2	onehot	0.696	(Activity #2) - Code
V2A_ACTIVITY_2_weekly	V2A	V2AJ02a1	2	mean	0.6972	How many times per week did you take part in this activity during the past 4 weeks?
V2A_ACTIVITY_2_minutes	V2A	V2AJ02a2	2	mean	0.7003	And when you took part in this activity, for how many minutes did you usually keep at it?
V2A_ACTIVITY_2_miles	V2A	V2AJ02a3	2	mean	0.911	How far do you usually walk/run/jog/cycle/swim? (Activity #2) - miles
V2A_ACTIVITY_3	V2A	V2AJ03	2	onehot	0.6924	During the past 4 weeks, did you participate in any physical activities?
V2A_ACTIVITY_3_code	V2A	V2AJ03a	2	onehot	0.9019	(Activity #3) - Code
V2A_ACTIVITY_3_weekly	V2A	V2AJ03a1	2	mean	0.902	How many times per week did you take part in this activity during the past 4 weeks?
V2A_ACTIVITY_3_minutes	V2A	V2AJ03a2	2	mean	0.9031	And when you took part in this activity, for how many minutes did you usually keep at it?
V2A_ACTIVITY_3_miles	V2A	V2AJ03a3	2	mean	0.978	How far do you usually walk/run/jog/cycle/swim? (Activity #3) - miles
V3A_ACTIVITY_1	V3A	V3AH01	3	onehot	0.0002	During the past 4 weeks, did you participate in any physical activities?
V3A_ACTIVITY_1_code	V3A	V3AH01a	3	onehot	0.2896	(Activity #1) - Code
V3A_ACTIVITY_1_weekly	V3A	V3AH01a1	3	mean	0.2906	How many times per week did you take part in this activity during the past 4 weeks?
V3A_ACTIVITY_1_minutes	V3A	V3AH01a2	3	mean	0.3001	And when you took part in this activity, for how many minutes did you usually keep at it?
V3A_ACTIVITY_1_miles	V3A	V3AH01a3	3	mean	0.6231	How far do you usually walk/run/jog/cycle/swim? (Activity #1) - miles
V3A_ACTIVITY_2	V3A	V3AH02	3	onehot	0.289	During the past 4 weeks, did you participate in any physical activities?
V3A_ACTIVITY_2_code	V3A	V3AH02a	3	onehot	0.739	(Activity #2) - Code
V3A_ACTIVITY_2_weekly	V3A	V3AH02a1	3	mean	0.7392	How many times per week did you take part in this activity during the past 4 weeks?
V3A_ACTIVITY_2_minutes	V3A	V3AH02a2	3	mean	0.7411	And when you took part in this activity, for how many minutes did you usually keep at it?
V3A_ACTIVITY_2_miles	V3A	V3AH02a3	3	mean	0.9277	How far do you usually walk/run/jog/cycle/swim? (Activity #2) - miles
V3A_ACTIVITY_3	V3A	V3AH03	3	onehot	0.7376	During the past 4 weeks, did you participate in any physical activities?
V3A_ACTIVITY_3_code	V3A	V3AH03a	3	onehot	0.923	(Activity #3) - Code
V3A_ACTIVITY_3_weekly	V3A	V3AH03a1	3	mean	0.923	How many times per week did you take part in this activity during the past 4 weeks?

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V3A_ACTIVITY_3_minutes	V3A	V3AH03a2	3	mean	0.9234	And when you took part in this activity, for how many minutes did you usually keep at it?
V3A_ACTIVITY_3_miles	V3A	V3AH03a3	3	mean	0.9817	How far do you usually walk/run/jog/cycle/swim? (Activity #3) - miles
V3AD07	V3A		3	mode	0.0435	(V3A) Has a care provider told you to restrict your activity in any way?
V4AE02	V4A		4	mode	0.006	(V4A) Before delivery, did a care provider tell you to restrict your activity in any way?

Activity Filtering

6.5 Demographics

The demographics filtering includes details about social factors including education and marital status, in addition to the more medically relevant factors of race, healthcare access and stress levels related to partner support.

Total # Features: 170
Layer 1 # Features: 21
Layer 2 # Features: 22
Relevant Files: V1A, V2A
Dropped Features: We chose to discard redundant features such as the racial background of the mother's parents since that information is inherently stored in the mother's race herself. Social constructs like ethnicity and perceived race (by others) were not deemed medically relevant, so those were also removed. Experiences in discrimination would further account for some of their social effects. We also believed that expectations for partner support was a more relevant proxy of social support than questions about the nature of the relationship such as length of relationship.

Table 8: Demographics Filtering

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V1AF02	V1A		1	Mean	0.0008	(V1A) How many years of education have you completed?
V1AF03	V1A		1	Onehot	0.0015	(V1A) Do you currently have a partner or 'significant other'?
V1AF04	V1A		1	Onehot	0.0015	(V1A) What is your current marital status?
V1AF05	V1A		-1	Onehot	0.0001	(V1A) Are you of Hispanic or Latino origin or descent?
V1AF09	V1A		-1	Mean	0.0057	(V1A) How many years have you lived in the United States? - years,
V1AF10	V1A		-1	Onehot	0.0005	(V1A) How would you rate your ability to speak and understand English?
V1AF13	V1A		-1	Mean	0.0018	(V1A) Counting yourself, how many people live in your household? - people
V1AF14	V1A		-1	Mean	0.1893	(V1A) Total family income for the past 12 months
V2AF02	V2A		2	Onehot	0.0007	(V2A) Do you currently have a partner or 'significant other'?
V2AF02a	V2A		2	Onehot	0.0611	(V2A) Are you currently living with your partner?
V2AF03	V2A		2	Onehot	0.0007	(V2A) What is your current marital status?
V2AF04	V2A		-1	Onehot	0.0048	(V2A) Is your mother of Hispanic or Latino origin or descent?
V2AF13	V2A		2	Mean	0.0283	(V2A) How old is the father of the baby? - years
V2AG01	V2A		-1	Onehot	0.0017	(V2A) If you feel you have been treated unfairly, do you usually accept it as a fact of life and keep it to yourself?
V2AG02	V2A		-1	Onehot	0.0019	(V2A) If you have been treated unfairly, do you usually talk to people or keep it to yourself?

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V1A PATIENT RACE	V1A	ONEHOT	-1	Onehot		
V1AF07a	V1A		-1		0.0004	(V1A) Racial background - White
V1AF07b	V1A		-1		0.0004	(V1A) Racial background - Black, African American or African Descent
V1AF07c	V1A		-1		0.0004	(V1A) Racial background - American Indian or Alaska Native
V1AF07d	V1A		-1		0.0004	(V1A) Racial background - Asian Indian
V1AF07e	V1A		-1		0.0004	(V1A) Racial background - Other Asian
V1AF07f	V1A		-1		0.0004	(V1A) Racial background - Native Hawaiian or Other Pacific Islander
V1AF07g	V1A		-1		0.0004	(V1A) Racial background - Other
V1A HEALTHCARE	V1A	ONEHOT	1	Onehot		
V1AF15a	V1A		1		0	(V1A) How is your health care currently paid for? - Government insurance (federal, state, or local)
V1AF15b	V1A		1		0	(V1A) How is your health care currently paid for? - Military insurance (Veteran's Administration or active duty)
V1AF15c	V1A		1		0	(V1A) How is your health care currently paid for? - Commercial health insurance/Commercial HMO
V1AF15d	V1A		1		0	(V1A) How is your health care currently paid for? - Personal household income
V1AF15e	V1A		1		0	(V1A) How is your health care currently paid for? - Other
V1AF15f	V1A		1		0	(V1A) How is your health care currently paid for? - Don't know
V1AF15g	V1A		1		0	(V1A) How is your health care currently paid for? - Refused
V1A PARTNER SUPPORT	V1A	ADD	1	Onehot		
V1AF03a1	V1A		1		0.0584	(V1A) What support do you expect your partner to give you during this pregnancy? - Emotional support
V1AF03a2	V1A		1		0.0584	(V1A) What support do you expect your partner to give you during this pregnancy? - Financial support
V1AF03a3	V1A		1		0.0584	(V1A) What support do you expect your partner to give you during this pregnancy? - To be present for my prenatal visits
V1AF03a4	V1A		1		0.0584	(V1A) What support do you expect your partner to give you during this pregnancy? - To be present for the delivery

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V2A HEALTHCARE	V2A	ONEHOT	2	Onehot		
V2AF01a	V2A		2		0.0005	(V2A) How is your health care currently paid for? - Government insurance (federal, state, or local)
V2AF01b	V2A		2		0.0005	(V2A) How is your health care currently paid for? - Military insurance (Veteran's Administration or active duty)
V2AF01c	V2A		2		0.0005	(V2A) How is your health care currently paid for? - Commercial health insurance/Commercial HMO
V2AF01d	V2A		2		0.0005	(V2A) How is your health care currently paid for? - Personal household income
V2AF01e	V2A		2		0.0005	(V2A) How is your health care currently paid for? - Other
V2AF01f	V2A		2		0.0005	(V2A) How is your health care currently paid for? - Don't know
V2AF01g	V2A		2		0.0005	(V2A) How is your health care currently paid for? - Refused
V2A MOTHER RACE	V2A	ONEHOT	-1	Onehot		
V2AF05a	V2A		-1		0.0006	(V2A) Mother's racial background - White
V2AF05b	V2A		-1		0.0006	(V2A) Mother's racial background - Black, African American or African Descent
V2AF05c	V2A		-1		0.0006	(V2A) Mother's racial background - American Indian or Alaska Native
V2AF05d	V2A		-1		0.0006	(V2A) Mother's racial background - Asian or Asian Indian
V2AF05e	V2A		-1		0.0006	(V2A) Mother's racial background - Native Hawaiian or Other Pacific Islander
V2AF05f	V2A		-1		0.0006	(V2A) Mother's racial background - Other
V2AF05g	V2A		-1		0.0006	(V2A) Mother's racial background - Don't know
V2AF05h	V2A		-1		0.0006	(V2A) Mother's racial background - Refused
V2A FATHER RACE	V2A	ONEHOT	-1	Onehot		
V2AF08a	V2A		-1		0.0007	(V2A) Father's racial background - White
V2AF08b	V2A		-1		0.0007	(V2A) Father's racial background - Black, African American or African Descent
V2AF08c	V2A		-1		0.0007	(V2A) Father's racial background - American Indian or Alaska Native
V2AF08d	V2A		-1		0.0007	(V2A) Father's racial background - Asian or Asian Indian
V2AF08e	V2A		-1		0.0007	(V2A) Father's racial background - Native Hawaiian or Other Pacific Islander
V2AF08f	V2A		-1		0.0007	(V2A) Father's racial background - Other
V2AF08f_SP	V2A		-1		0.8986	(V2A) Father's racial background - Other, specify
V2AF08g	V2A		-1		0.0007	(V2A) Father's racial background - Don't know
V2AF08h	V2A		-1		0.0007	(V2A) Father's racial background - Refused

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V2A DISCRIMINATION	V2A	ADD	-1	Mean		
V2AG02	V2A		-1		0.0019	(V2A) If you have been treated unfairly, do you usually talk to people or keep it for yourself
V2AG03a	V2A		-1		0.0017	(V2A) Have you ever experienced discrimination, been prevented from doing something, or been hassled or made to feel inferior in any of the following situations because of your race, ethnicity, or color: at school?
V2AG03b	V2A		-1		0.0017	(V2A) ' ' Getting hired or getting a job?
V2AG03c	V2A		-1		0.0017	(V2A) ' ' At work?
V2AG03d	V2A		-1		0.0017	(V2A) ' ' Getting housing?
V2AG03e	V2A		-1		0.0017	(V2A) ' ' Getting medical care?
V2AG03f	V2A		-1		0.0017	(V2A) ' ' Getting service in a store or restaurant?
V2AG03g	V2A		-1		0.0017	(V2A) ' ' Getting credit, bank loans, or a mortgage?
V2AG03h	V2A		-1		0.0018	(V2A) ' ' On the street or in a public setting?
V2AG03i	V2A		-1		0.0018	(V2A) ' ' From the police or in the courts?

Demographics Filtering

6.6 Physiology

The physiology filtering includes questions regarding the current and recent physiological health and questioning of the patient. This includes factors such as temperature, weight, body measurements, flu-like symptoms, nausea, blood pressure, and recent evidence of contractions or vaginal bleeding.

Total # Features: 115
Layer 1 # Features: 38
Layer 2 # Features: 38
Relevant Files: CMA, CMB, Demographics (ancillary), V1A, V1B, V2A, V2B, V3A, V3B, V4A
Dropped Features: For this group, we chose to exclude features that were detail oriented to questions that were asked earlier in the questionnaire. For example, details on flu-like illness symptoms were dropped, while the general question on if any such illness was present was kept, as the general question implicitly covers the presence of a response in the details that follow. Certain details were added back within L2, such as contraction and vaginal bleeding times. Multiple measurements of the same region were averaged into a single feature. Maximum, rather than average, blood pressure and temperature on admission were kept as they were deemed a better representation of potential risk or medical issues. Metadata features were used to organize data and later dropped.

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSINGNESS	DESCRIPTION
CMBC02b1	CMB		Date Abstraction	Mean	0.0266	(CMB) Blood pressure - Highest systolic blood pressure during stay (mmHg),
CMBC02b2	CMB		Date Abstraction	Mean	0.0266	(CMB) Blood pressure - Highest diastolic blood pressure during stay (mmHg),
BMI_Cat	demographics (ancillary)		1	Bmicat	0.0225	BMI category (kg/mˆ2) at visit 1 calculated from height and weight (form V1B)
V1AD06	V1A		1	Any	0.0004	(V1A) Have you had any 'flu-like illnesses,' 'really bad colds,' fever, a rash , or any muscle or joint aches since you
V1AD07	V1A		1	Mode	0.0008	(V1A) Have you had any cramping since you became pregnant?
V1AD08	V1A		1	Multi	0.0005	(V1A) Since you became pregnant, have you had vaginal bleeding more than spotting?
V1AD09	V1A		1	Mode	0.0004	(V1A) In the last 12 hours, for how many hours have you felt nauseated?
V1AD10	V1A		1	Mode	0.0003	(V1A) In the last 12 hours, how many times have you vomited?
V1AD11	V1A		1	Mode	0.0003	(V1A) In the last 12 hours, how many times have you had retching or dry heaves without emesis (vomiting)?
V2AD01	V2A		2	Any	0.0008	(V2A) Have you had any 'flu-like illnesses,' 'really bad colds,' fever, a rash, or any muscle or joint aches since last study
V2AD02	V2A		2	Single	0.001	(V2A) Have you had any cramping/contractions since last study visit?
V2AD03	V2A		2	Multi	0.001	(V2A) Since last study visit, have you had vaginal bleeding more than spotting?
V2AD04	V2A		2	Mode	0.0012	(V2A) In the last 12 hours, for how many hours have you felt nauseated?
V2AD05	V2A		2	Mode	0.0012	(V2A) In the last 12 hours, how many times have you vomited?
V2AD06	V2A		2	Mode	0.0013	(V2A) In the last 12 hours, how many times have you had retching or dry heaves without emesis (vomiting)?
V3AD01	V3A		3	Any	0.0001	(V3A) Have you had any 'flu-like illnesses,' 'really bad colds,' fever, a rash, or any muscle or joint aches since last study
V3AD02	V3A		3	Single	0.0002	(V3A) Have you had any cramping/contractions since last study visit?
V3AD03	V3A		3	Multi	0.0009	(V3A) Since last study visit, have you had vaginal bleeding more than spotting?
V3AD04	V3A		3	Mode	0.0007	(V3A) In the last 12 hours, for how many hours have you felt nauseated?
V3AD05	V3A		3	Mode	0.0007	(V3A) In the last 12 hours, how many times have you vomited?
V3AD06	V3A		3	Mode	0.0007	(V3A) In the last 12 hours, how many times have you had retching or dry heaves without emesis (vomiting)?
V4AE01	V4A		4	Any	0.0001	(V4A) Did you have any 'flu-like illnesses,' 'really bad colds,' fever, a rash, or any muscle or joint aches between date of

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSINGNESS	DESCRIPTION
V1B_Resting_BP_Sys	V1B	Mean	1	Mean		
V1BA06a1	V1B				0.0092	(V1B) Resting blood pressure - Systolic (mm Hg) measure 1
V1BA06a2	V1B				0.9749	(V1B) Resting blood pressure - Systolic (mm Hg) measure 2
V1B_Resting_BP_Dia	V1B	Mean	1	Mean		
V1BA06b1	V1B				0.0092	(V1B) Resting blood pressure - Diastolic (mm Hg) measure 1
V1BA06b2	V1B				0.9749	(V1B) Resting blood pressure - Diastolic (mm Hg) measure 2
V1B_Neck_Circ	V1B	Mean	1	Mean		
V1BA07a	V1B				0.1392	(V1B) Neck circumference - Measure 1 (cm)
V1BA07b	V1B				0.1442	(V1B) Neck circumference - Measure 2 (cm)
V1BA07c	V1B				0.9924	(V1B) Neck circumference - Measure 3 (cm)
V2B_Weight_lbs	V2B	Mean	2	Mean		
V2BA01_KG	V2B				0.8669	(V2B) Weight - kg
V2BA01_LB	V2B				0.1361	(V2B) Weight - lbs
V2B_Resting_BP_Sys	V2B	Mean	2	Mean		
V2BA02a1	V2B				0.006	(V2B) Resting blood pressure - Systolic measure 1 (mmHg)
V2BA02a2	V2B				0.9806	(V2B) Resting blood pressure - Systolic measure 2 (mmHg)
V2B_Resting_BP_Dia	V2B	Mean	2	Mean		
V2BA02b1	V2B				0.006	(V2B) Resting blood pressure - Diastolic measure 1 (mmHg)
V2BA02b2	V2B				0.9808	(V2B) Resting blood pressure - Diastolic measure 2 (mmHg)
V3B_Weight_lbs	V3B	Weight	3	Mean		
V3BA01_KG	V3B				0.8675	(V3B) Weight - kg
V3BA01_LB	V3B				0.1352	(V3B) Weight - lbs
V3B_Resting_BP_Sys	V3B	Mean	3	Mean		
V3BA02a1	V3B				0.0052	(V3B) Resting blood pressure - Systolic measure 1 (mmHg)
V3BA02a2	V3B				0.9763	(V3B) Resting blood pressure - Systolic measure 2 (mmHg)
V3B_Resting_BP_Dia	V3B	Mean	3	Mean		
V3BA02b1	V3B				0.0052	(V3B) Resting blood pressure - Diastolic measure 1 (mmHg)
V3BA02b2	V3B				0.9763	(V3B) Resting blood pressure - Diastolic measure 2 (mmHg)
CMA_Weight_Pre_Delivery_lbs	CMA	Weight	4	Mean		
CMAB01a1	CMA				0.5736	(CMA) Participant's last weight prior to delivery - kilograms
CMAB01a2	CMA				0.4468	(CMA) Participant's last weight prior to delivery - pounds
CMB_Max_Admission_Temp	CMB	Temp	Date Abstraction	Mean		
CMBC01b1	CMB				0.481	(CMB) Maximum temperature during admission or encounter - degrees Celsius,
CMBC01b2	CMB				0.5494	(CMB) Maximum temperature during admission or encounter - degrees Fahrenheit,
V1B_Weight_lbs	V1B	Weight	1	Mean		
V1BA01_KG	V1B				0.8577	(V1B) Weight - kg
V1BA01_LB	V1B				0.1439	(V1B) Weight - lbs
V1B_Height	V1B	Mean	1	Mean		
V1BA02a	V1B				0.0093	(V1B) Height - Measure 1 (cm)
V1BA02b	V1B				0.0229	(V1B) Height - Measure 2 (cm)
V1BA02c	V1B				0.992	(V1B) Height - Measure 3 (cm)
V1B_Natural_Waist_Circumference	V1B	Mean	1	Mean		
V1BA03a	V1B				0.0183	(V1B) Natural waist circumference - Measure 1 (cm)
V1BA03b	V1B				0.0233	(V1B) Natural waist circumference - Measure 2 (cm)
V1BA03c	V1B				0.9745	(V1B) Natural waist circumference - Measure 3 (cm)
V1B_Waist_Lilac_Crest	V1B	Mean	1	Mean		
V1BA04a	V1B				0.0184	(V1B) Waist circumference over iliac crest - Measure 1 (cm)
V1BA04b	V1B				0.0234	(V1B) Waist circumference over iliac crest - Measure 2 (cm)
V1BA04c	V1B				0.9721	(V1B) Waist circumference over iliac crest - Measure 3 (cm)
V1B_Hip_Circ	V1B	Mean	1	Mean		
V1BA04a	V1B				0.0189	(V1B) Hip circumference - Measure 1 (cm)
V1BA04b	V1B				0.0238	(V1B) Hip circumference - Measure 2 (cm)
V1BA04c	V1B				0.9768	(V1B) Hip circumference - Measure 3 (cm)

6.7 Ultrasound

The ultrasound filtering group contains all features located within the ultrasound related files and chart abstractions. This includes fetal measurements, uterine artery analysis, adrenal gland study, noted conditions, amniotic fluid information, and more.

Total # Features: 542
Layer 1 # Features: 59
Layer 2 # Features: 131
Relevant Files: CUA, CUB, S02, U02, U1C, U2A, U2B, U2C, U3A, U3B, U3C, U3D
Dropped Features: For this group, there were many excess features that were dropped (100% missingness), for noted abnormalities. Abnormality code was the only feature used to represent fetal abnormalities. There was abundant metadata relevant only for either organizing data or populating other features and therefore redundant, which was dropped. Details for conditions were left for Layer 2, as were detailed measurements in the uterine artery.

NAME	FILE	RULE	TEMP	IMPUTE	MISSING	DESCRIPTION
CUAB02a2	CUA		1	999	0.9855	(CUA) Brain Abnormality - Code (a)
CUAC02a2	CUA		1	999	0.9981	(CUA) Head/neck abnormality - Code (a)
CUAD02a2	CUA		1	999	0.9868	(CUA) Chest abnormality - Code (a)
CUAD02b2	CUA		1	999	0.9987	(CUA) Chest abnormality - Code (b)
CUAD02c2	CUA		1	999	0.9994	(CUA) Chest abnormality - Code (c)
CUAE02a2	CUA		1	999	0.9779	(CUA) Abdomen abnormality - Code (a)
CUAE02b2	CUA		1	999	0.9981	(CUA) Abdomen abnormality - Code (b)
CUAF02a2	CUA		1	999	0.9975	(CUA) Extremities abnormality - Code (a)
CUAH02a2	CUA		1	999	0.9924	(CUA) Malformation(s) not mentioned under brain, head/neck, chest, abdomen, extremities or spine - Code (a)
CUAH02b2	CUA		1	999	0.9987	(CUA) Malformation(s) not mentioned under brain, head/neck, chest, abdomen, extremities or spine - Code (b)
CUBB01	CUB		1	999	0.0008	(CUB) Were any structural abnormalities detected?
CUBB02a2	CUB		1	999	0.9417	(CUB) Abnormality - Code (a)
CUBB02b2	CUB		1	999	0.9935	(CUB) Abnormality - Code (b)
CUBB02c2	CUB		1	999	0.9974	(CUB) Abnormality - Code (c)
CUBB02d2	CUB		1	999	0.9987	(CUB) Abnormality - Code (d)
CUBB02e2	CUB		1	999	0.9992	(CUB) Abnormality - Code (e)
CUBB02f2	CUB		1	999	0.9997	(CUB) Abnormality - Code (f)
U1CB02	U1C		1	ANY	0.0014	(U1C) Disclosable condition
U1CB04c	U1C		1	999	0.9971	(U1C) Disclosable condition reported - Incidental detection of complete or partial placenta previa or vasa previa
U1CC04	U1C		1	MEAN	0.0287	(U1C) Left Uterine Artery - Peak Systolic Velocity (PSV) (cm/sec)
U1CC05	U1C		1	MEAN	0.0337	(U1C) Left Uterine Artery - Systolic/Diastolic Ratio (S/D)
U1CC06	U1C		1	MEAN	0.0294	(U1C) Left Uterine Artery - Resistance Index (RI)
U1CC07	U1C		1	MEAN	0.0309	(U1C) Left Uterine Artery - Pulsatility Index (PI)
U1CC09	U1C		1	999	0.0266	(U1C) Left Uterine Artery - Early diastolic notch present
U1CC12	U1C		1	999	0.023	(U1C) Left Uterine Artery - Systolic notch present
U1CD04	U1C		1	MEAN	0.0345	(U1C) Right Uterine Artery - Peak Systolic Velocity (PSV) (cm/sec)
U1CD05	U1C		1	MEAN	0.0402	(U1C) Right Uterine Artery - Systolic/Diastolic Ratio (S/D)
U1CD06	U1C		1	MEAN	0.0345	(U1C) Right Uterine Artery - Resistance Index (RI)
U1CD07	U1C		1	MEAN	0.0359	(U1C) Right Uterine Artery - Pulsatility Index (PI)
U1CD09	U1C		1	999	0.0302	(U1C) Right Uterine Artery - Early diastolic notch present
U1CD12	U1C		1	999	0.0294	(U1C) Right Uterine Artery - Systolic notch present
U2AB02	U2A		2	MEAN	0.0113	(U2A) Biparietal Diameter - cm
U2AB03	U2A		2	MEAN	0.011	(U2A) Head Circumference - cm
U2AB04	U2A		2	MEAN	0.0115	(U2A) Abdominal Circumference - cm
U2AB05	U2A		2	MEAN	0.0106	(U2A) Femur Diaphysis Length - cm
U2AB06	U2A		2	MEAN	0.1167	(U2A) Estimated fetal weight as reported - grams
U2AB07	U2A		2	N/A	0.595	(U2A) Estimated fetal weight percentile as reported - percentile
U2AC01	U2A		2	ANY	0.4733	(U2A) Disclosable condition
U2AC03c	U2A		2	999	0.9871	(U2A) Disclosable condition reported - Incidental finding of oligohydramnios
U2AC03e	U2A		2	999	0.9871	(U2A) Disclosable condition reported - Incidental detection of complete or partial placenta previa or vasa previa
U2BB02	U2B		2	NEG1	0.0208	(U2B) Cervical length - mm
U2BB03	U2B		2	999	0.0051	(U2B) Funnel
U2BB04	U2B		2	NEG1	0.9944	(U2B) Funnel length - mm
U2BB05	U2B		2	999	0.0059	(U2B) Debris

NAME	FILE	RULE	TEMP	IMPUTE	MISSING	DESCRIPTION
U2BC01	U2B		2	ANY	0.1082	(U2B) Disclosable condition
U2BC03d	U2B		2	999	0.9815	(U2B) Disclosable condition reported - Incidental detection of complete or partial placenta previa or vasa previa
U2CB03	U2C		2	ANY	0	(U2C) Disclosable condition reported
U2CB04c	U2C		2	999	0.9825	(U2C) Disclosable condition reported - Incidental detection of complete or partial placenta previa or vasa previa
U2CC04	U2C		2	MEAN	0.0167	(U2C) Left Uterine Artery - Peak Systolic Velocity (PSV) (cm/sec)
U2CC05	U2C		2	MEAN	0.0191	(U2C) Left Uterine Artery - Systolic/Diastolic Ratio (S/D)
U2CC06	U2C		2	MEAN	0.0171	(U2C) Left Uterine Artery - Resistance Index (RI)
U2CC07	U2C		2	MEAN	0.0188	(U2C) Left Uterine Artery - Pulsatility Index (PI)
U2CC09	U2C		2	999	0.0146	(U2C) Left Uterine Artery - Early diastolic notch present
U2CC12	U2C		2	999	0.016	(U2C) Left Uterine Artery - Systolic notch present
U2CD04	U2C		2	MEAN	0.0151	(U2C) Right Uterine Artery - Peak Systolic Velocity (PSV) (cm/sec)
U2CD05	U2C		2	MEAN	0.0182	(U2C) Right Uterine Artery - Systolic/Diastolic Ratio (S/D)
U2CD06	U2C		2	MEAN	0.0151	(U2C) Right Uterine Artery - Resistance Index (RI)
U2CD07	U2C		2	MEAN	0.0164	(U2C) Right Uterine Artery - Pulsatility Index (PI)
U2CD09	U2C		2	999	0.0131	(U2C) Right Uterine Artery - Early diastolic notch present
U2CD12	U2C		2	999	0.0148	(U2C) Right Uterine Artery - Systolic notch present
U3AB02	U3A		3	MEAN	0.0077	(U3A) Biparietal Diameter (cm)
U3AB03	U3A		3	MEAN	0.0076	(U3A) Head Circumference (cm)
U3AB04	U3A		3	MEAN	0.0064	(U3A) Abdominal Circumference (cm)
U3AB05	U3A		3	MEAN	0.0068	(U3A) Femur Diaphysis Length (cm)
U3AB07	U3A		3	MEAN	0.0127	(U3A) Estimated fetal weight as reported - grams
U3AC01	U3A		3	ANY	0.0437	(U3A) Disclosable condition
U3AC03c	U3A		3	SUM_LEQ	0.9915	(U3A) Disclosable condition reported - Oligohydramnios (maximal vertical pocket of <2 cm or AFI <5 cm)
U3AC03e	U3A		3	999	0.9915	(U3A) Disclosable condition reported - Incidental detection of complete or partial placenta previa or vasa previa
U3BB02	U3B		3	NEG1	0.0152	(U3B) Cervical length - mm
U3BB03	U3B		3	999	0.0052	(U3B) Funnel
U3BB04	U3B		3	NEG1	0.9799	(U3B) Funnel length - mm
U3BB05	U3B		3	999	0.0064	(U3B) Debris
U3BC01	U3B		3	ANY	0.0122	(U3B) Disclosable condition
U3BC03d	U3B		3	999	0.984	(U3B) Disclosable condition reported - Incidental detection of complete or partial placenta previa or vasa previa
U3CB02	U3C		3	ANY	0.0009	(U3C) Disclosable condition
U3CB04c	U3C		3	999	0.9954	(U3C) Disclosable condition reported - Incidental detection of complete or partial placenta previa or vasa previa
U3CC04	U3C		3	MEAN	0.0271	(U3C) Left Uterine Artery - Peak Systolic Velocity (PSV) (cm/sec)
U3CC05	U3C		3	MEAN	0.0271	(U3C) Left Uterine Artery - Systolic/Diastolic Ratio (S/D)
U3CC06	U3C		3	MEAN	0.0264	(U3C) Left Uterine Artery - Resistance Index (RI)
U3CC07	U3C		3	MEAN	0.0266	(U3C) Left Uterine Artery - Pulsatility Index (PI)
U3CC09	U3C		3	999	0.0243	(U3C) Left Uterine Artery - Early diastolic notch present
U3CC12	U3C		3	999	0.0254	(U3C) Left Uterine Artery - Systolic notch present
U3CD04	U3C		3	MEAN	0.0259	(U3C) Right Uterine Artery - Peak Systolic Velocity (PSV) (cm/sec)
U3CD05	U3C		3	MEAN	0.0275	(U3C) Right Uterine Artery - Systolic/Diastolic Ratio (S/D)
U3CD06	U3C		3	MEAN	0.0257	(U3C) Right Uterine Artery - Resistance Index (RI)
U3CD07	U3C		3	MEAN	0.0264	(U3C) Right Uterine Artery - Pulsatility Index (PI)
U3CD09	U3C		3	999	0.0234	(U3C) Right Uterine Artery - Early diastolic notch present
U3CD12	U3C		3	999	0.0268	(U3C) Right Uterine Artery - Systolic notch present
U3DC01	U3D		3	999	0.0677	(U3D) Fetal adrenal gland mass
U3DC02	U3D		3	999	0.0004	(U3D) Disclosable condition reported
U3DD01	U3D		3	999	0.0895	(U3D) Fetal Adrenal Gland Measures Visit 3 - Which gland was used for measurement
Calc_Polyhydramnios		Calc_ Polyhydramnios	3			
U3AB06a	U3A		3	MEAN	0.0761	(U3A) Amniotic Fluid Index (AFI) - Quadrant 1 (cm)
U3AB06b	U3A		3	MEAN	0.074	(U3A) Amniotic Fluid Index (AFI) - Quadrant 2 (cm)
U3AB06c	U3A		3	MEAN	0.0736	(U3A) Amniotic Fluid Index (AFI) - Quadrant 3 (cm)
U3AB06d	U3A		3	MEAN	0.0833	(U3A) Amniotic Fluid Index (AFI) - Quadrant 4 (cm)
adrenal_gland_length		adrenal_measure	3			
U3DD02a1	U3D		3	MEAN	0.1898	(U3D) Adrenal Gland Measurements, Adrenal gland - Length in mm (1)
U3DD02a2	U3D		3	MEAN	0.2111	(U3D) Adrenal Gland Measurements, Adrenal gland - Length in mm (2)
U3DD02a3	U3D		3	MEAN	0.257	(U3D) Adrenal Gland Measurements, Adrenal gland - Length in mm (3)
U3DD02a1	U3D		3	MEAN	0.9951	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland length in mm, Measure 1
U3DF02a2	U3D		3	MEAN	0.9955	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland length in mm, Measure 2
U3DF02a3	U3D		3	MEAN	0.9955	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland length in mm, Measure 3
U3DD05	U3D		3	N/A	0.0579	(U3D) Fetal Adrenal Gland Measures Visit 3 - Adrenal Glad measurements exam complete
U3DF05	U3D		3	N/A	0.9933	(U3D) Repeat Adrenal Gland Measurements - Exam complete

NAME	FILE	RULE	TEMP	IMPUTE	MISSING	DESCRIPTION
fetal_zone_length		adrenal_measure	3			
U3DD02b1	U3D		3	MEAN	0.1906	(U3D) Adrenal Gland Measurements, Fetal zone - Length in mm (1)
U3DD02b2	U3D		3	MEAN	0.212	(U3D) Adrenal Gland Measurements, Fetal zone - Length in mm (2)
U3DD02b3	U3D		3	MEAN	0.2584	(U3D) Adrenal Gland Measurements, Fetal zone - Length in mm (3)
U3DF02b1	U3D		3	MEAN	0.9951	(U3D) Repeat Adrenal Gland Measurements, Fetal zone length in mm, Measure 1
U3DF02b2	U3D		3	MEAN	0.9955	(U3D) Repeat Adrenal Gland Measurements, Fetal zone length in mm, Measure 2
U3DF02b3	U3D		3	MEAN	0.9955	(U3D) Repeat Adrenal Gland Measurements, Fetal zone length in mm, Measure 3
U3DD05	U3D		3	N/A	0.0579	(U3D) Fetal Adrenal Gland Measures Visit 3 - Adrenal Glad measurements exam complete
U3DF05	U3D		3	N/A	0.9933	(U3D) Repeat Adrenal Gland Measurements - Exam complete
adrenal_gland_width		adrenal_measure	3			
U3DD03a1	U3D		3	MEAN	0.1929	(U3D) Adrenal Gland Measurements, Adrenal gland - Width in mm (1)
U3DD03a2	U3D		3	MEAN	0.2134	(U3D) Adrenal Gland Measurements, Adrenal gland - Width in mm (2)
U3DD03a3	U3D		3	MEAN	0.2615	(U3D) Adrenal Gland Measurements, Adrenal gland - Width in mm (3)
U3DF03a1	U3D		3	MEAN	0.9951	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland length in mm, Measure 1
U3DF03a2	U3D		3	MEAN	0.9955	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland length in mm, Measure 2
U3DF03a3	U3D		3	MEAN	0.9955	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland length in mm, Measure 3
U3DD05	U3D		3	N/A	0.0579	(U3D) Fetal Adrenal Gland Measures Visit 3 - Adrenal Glad measurements exam complete
U3DF05	U3D		3	N/A	0.9933	(U3D) Repeat Adrenal Gland Measurements - Exam complete
fetal_zone_width		adrenal_measure	3			
U3DD03b1	U3D		3	MEAN	0.1929	(U3D) Adrenal Gland Measurements, Fetal zone - Width in mm (1)
U3DD03b2	U3D		3	MEAN	0.2143	(U3D) Adrenal Gland Measurements, Fetal zone - Width in mm (2)
U3DD03b3	U3D		3	MEAN	0.2633	(U3D) Adrenal Gland Measurements, Fetal zone - Width in mm (3)
U3DF03b1	U3D		3	MEAN	0.9951	(U3D) Repeat Adrenal Gland Measurements, Fetal zone width in mm, Measure 1
U3DF03b2	U3D		3	MEAN	0.9955	(U3D) Repeat Adrenal Gland Measurements, Fetal zone width in mm, Measure 2
U3DF03b3	U3D		3	MEAN	0.9955	(U3D) Repeat Adrenal Gland Measurements, Fetal zone width in mm, Measure 3
U3DD05	U3D		3	N/A	0.0579	(U3D) Fetal Adrenal Gland Measures Visit 3 - Adrenal Glad measurements exam complete
U3DF05	U3D		3	N/A	0.9933	(U3D) Repeat Adrenal Gland Measurements - Exam complete

NAME	FILE	RULE	TEMP	IMPUTE	MISSING	DESCRIPTION
adrenal_gland_depth		adrenal_measure	3			
U3DD04a1	U3D		3	MEAN	0.7657	(U3D) Adrenal Gland Measurements, Adrenal gland - Depth in mm (1)
U3DD04a2	U3D		3	MEAN	0.7679	(U3D) Adrenal Gland Measurements, Adrenal gland - Depth in mm (2)
U3DD04a3	U3D		3	MEAN	0.7719	(U3D) Adrenal Gland Measurements, Adrenal gland - Depth in mm (3)
U3DF04a1	U3D		3	MEAN	0.9982	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland depth in mm, Measure 1
U3DF04a2	U3D		3	MEAN	0.9982	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland depth in mm, Measure 2
U3DF04a3	U3D		3	MEAN	0.9982	(U3D) Repeat Adrenal Gland Measurements, Adrenal gland depth in mm, Measure 3
U3DD05	U3D		3	N/A	0.0579	(U3D) Fetal Adrenal Gland Measures Visit 3 - Adrenal Glad measurements exam complete
U3DF05	U3D		3	N/A	0.9933	(U3D) Repeat Adrenal Gland Measurements - Exam complete
fetal_zone_depth		adrenal_measure	3			
U3DD04b1	U3D		3	MEAN	0.7657	(U3D) Adrenal Gland Measurements, Fetal zone - Depth in mm (1)
U3DD04b2	U3D		3	MEAN	0.7679	(U3D) Adrenal Gland Measurements, Fetal zone - Depth in mm (2)
U3DD04b3	U3D		3	MEAN	0.7719	(U3D) Adrenal Gland Measurements, Fetal zone - Depth in mm (3)
U3DF04b1	U3D		3	MEAN	0.9982	(U3D) Repeat Adrenal Gland Measurements, Fetal zone, depth in mm, Measure 1
U3DF04b2	U3D		3	MEAN	0.9982	(U3D) Repeat Adrenal Gland Measurements, Fetal zone, depth in mm, Measure 2
U3DF04b3	U3D		3	MEAN	0.9982	(U3D) Repeat Adrenal Gland Measurements, Fetal zone, depth in mm, Measure 3
U3DD05	U3D		3	N/A	0.0579	(U3D) Fetal Adrenal Gland Measures Visit 3 - Adrenal Glad measurements exam complete
U3DF05	U3D		3	N/A	0.9933	(U3D) Repeat Adrenal Gland Measurements - Exam complete

6.8 Medical History

The medical history filtering includes questions regarding the current or recent (prenatal labs) medical conditions of the patient, which is a quite large category. As a result, using literature review and clinician input, we narrowed down our relevant features to PTB-related factors like STI screens, gynecological infections, and diabetes/hypertension data. Aside from diagnostics, we also included prior information about previous pregnancy loss history, given the trial's focus on nulliparous patients, as well as relevant CBC results like hematocrit levels, and patient previous procedures and pre-pregnancy weight.

Total Features: 2294
Layer 1 Features: 100
Layer 2 Features: 100
Relevant Files: CLA, CLB, CMA, CMB, CMD, CME, S02, V1A, VXX, demographics (ancillary), pregnancy_outcomes (ancillary)
Dropped Features: For this group, we dropped most features due to a large amount of information. As a result, we focused on health conditions that pertained most to PTB risk factors.

Table 9: Medical History Filtering

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
CLAA01a	CLA		1	Mean	0.035	(CLA) Complete blood count - Hemoglobin g/dL
CLAA01b	CLA		1	Mean	0.0376	(CLA) Complete blood count - Hematocrit %
CLAA01d	CLA		1	Mean	0.0945	(CLA) Complete blood count - Platelet count x10 3 /mm 3 (x10 3L)
CLAA06a2	CLA		1	Onehot	0.8675	(CLA) Urine culture - Organism code (a)
CLAA06b2	CLA		1	Onehot	0.9912	(CLA) Urine culture - Organism code (b)
CLAA07	CLA		1	Onehot	0.1827	(CLA) Gonorrheal screen
CLAA08	CLA		1	Onehot	0.1751	(CLA) Chlamydial screen
CLAA09	CLA		1	Onehot	0.0535	(CLA) VDRL/RPR (syphilis) test
CLAB02d2	CLA		1	Mean	0.5419	(CLA) First trimester serum screen PAPP-A Results - Multiples of the median (MoM)
CLAB02e2	CLA		1	Mean	0.5616	(CLA) First trimester serum screen beta HCG - Multiples of the median (MoM)
CLAC01c2	CLA		2	Mean	0.6079	(CLA) Second trimester screen AFP Results - Multiples of the median (MoM)
CLAC01d2	CLA		2	Mean	0.6848	(CLA) Second trimester screen Total beta HCG Results - Multiples of the median (MoM)
CLAC01e2	CLA		2	Mean	0.6831	(CLA) Second trimester screen uEstriol Results - Multiples of the median (MoM)
CLAC01f2	CLA		2	Mean	0.6867	(CLA) Inhibin A (DIA) Results - Multiples of the median (MoM)
CLAD01a2	CLA		3	Mean	0.1797	(CLA) Third Trimester Lab Studies (>24 0 weeks) Blood count, Lowest hemoglobin during third trimester before labor
CLAD01b2	CLA		3	Mean	0.2049	(CLA) Lowest hematocrit during third trimester before labor and delivery - Result %
CLAD02a	CLA		3	Onehot	0.7854	(CLA) Last third trimester urine culture - result
CLAD02a1b	CLA		3	Onehot	0.9561	(CLA) Last third trimester urine culture - Organism code (1)
CLAD02a2b	CLA		3	Onehot	0.9962	(CLA) Last third trimester urine culture - Organism code (2)
CLAD03a	CLA		3	Onehot	0.7255	(CLA) Last third trimester Gonorrheal screen - Result
CLAD04a	CLA		3	Onehot	0.7268	(CLA) Last third trimester Chlamydial screen - result
CLAD05a	CLA		3	Onehot	0.6993	(CLA) Last third trimester VDRL/RPR (syphilis) test - Result
CLAD06a	CLA		3	Onehot	0.1277	(CLA) Last third trimester GBS culture/test - Result

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
CMAC16	CMA		4	Onehot	0.0021	(CMA) Did the participant have a diagnosis of chorioamnionitis documented in the chart by a care giver prior to delivery
CMAE05	CMA		4	Onehot	0.0018	(CMA) Abruptio placenta
CMAE09	CMA		4	Onehot	0.0019	(CMA) Cerebral vascular accident (CVA)
CMAE10	CMA		4	Onehot	0.0018	(CMA) SEver anemia (Hematocrit < 20 vol% treated with blood transfusion)
CMAH01b3	CMA		4	Mean	0.2241	(CMA) Results of complete blood count: - Hemoglobin (g/dL)
CMAH01b4	CMA		4	Mean	0.2244	(CMA) Results of complete blood count: - Hematocrit (%)
CMAH01b6	CMA		4	Mean	0.2276	(CMA) Results of complete blood count: - Platelet count x10^3 / mm^3 (x10^3 /_L)
CMAH02b	CMA		4	Onehot	0.9791	(CMA) Gonorrhea screen test result
CMAH03b	CMA		4	Onehot	0.9786	(CMA) Chlamydia screen test result
CMAH04b	CMA		4	Onehot	0.9583	(CMA) Urine drug screen result
CMBF01b	CMB		4	Onehot	0.8973	(CMB) Participant had a cervico-vaginal fetal fibronectin sampled collected, result
CMBG01	CMB		4	Onehot	0	(CMB) Participant had a cerclage placed
CMBI01	CMB		4	Onehot	0	(CMB) Did the participant receive tocolysis at any time during this hospital encounter?
CMBJ01	CMB		4	Onehot	0	(CMB) Did the participant have an infection or receive antibiotics during this hospital encounter?
CMBJ02	CMB		4	Onehot	0.8441	(CMB) Urinary tract infection or pyelonephritis
CMBJ03	CMB		4	Onehot	0.8441	(CMB) Sepsis
CMBJ04	CMB		4	Onehot	0.8422	(CMB) Pneumonia
CMBJ05	CMB		4	Onehot	0.8384	(CMB) Other maternal infections (e.g., viral, fungal, parasitic, spirochetal, other bacterial), excluding bacterial vaginosis
CMEA01	CME		4	Onehot	0	(CME) Urinary tract infection
CMEA02	CME		4	Onehot	0	(CME) Sepsis
CMEA03	CME		4	Onehot	0	(CME) Pneumonia
CMEA04	CME		4	Onehot	0	(CME) Other maternal infections (e.g., viral, fungal, parasitic, spirochetal, other bacterial), excluding bacterial vaginosis
GravCat	Demographics		1	Mode	0.001	Gravidity category (V1AE01)
oDM	Pregnancy outcomes		0	Mode	0.047	Diabetes based on CMAE03 & glucose tolerance testing results in CLA section E
ChronHTN	Pregnancy outcomes		1	Mode	0.0567	Chronic hypertension based on CMDA01 & CMAE01
PEgHTN	Pregnancy outcomes		3	Mode	0.0568	Preeclampsia/Gestational HTN (worst) using nuMoM2b criteria (CMDA08a)
S02A01	S02		0	Mean	0	(S02) Total number of pregnancies including current pregnancy - number of pregnancies
S02C01	S02		0	Onehot	0.0003	(S02) Assisted reproduction for this pregnancy
V1AD01b	V1A		1	Mean	0.0207	(V1A) How much did you weigh before you got pregnant? - lbs
V1AD12a	V1A		1	Onehot	0.0003	(V1A) Previous surgeries - Cervical surgery - cone
V1AD12b	V1A		1	Onehot	0.0003	(V1A) Previous surgeries - Cervical surgery - LEEP
V1AD12c	V1A		1	Onehot	0.0003	(V1A) Previous surgeries - Cervical surgery - Cryotherapy
VXXB01aa_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, High blood pressure (hypertension) - Final Assessment: Condition present
VXXB01ab_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Asthma - Final Assessment: Condition present
VXXB01ac_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Seizure disorder - Final Assessment: Condition present
VXXB01ad_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Migraine headaches - Final Assessment: Condition present
VXXB01ae_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Diabetes (excluding gestational diabetes in a prior pregnancy) - Final Assessment: Condition present

NAME	FILE	RULE	TEMP	IMPUTE	MISSING	DESCRIPTION
VXXB01ae1_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Diabetes, onset during pregnancy (gestational diabetes) - Final Assessment: Condition present
VXXB01af_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Hyperthyroidism - Final Assessment: Condition present
VXXB01ag_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Hypothyroidism - Final Assessment: Condition present
VXXB01ah_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Valvular heart disease - Final Assessment: Condition present
VXXB01ai_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Other structural heart disease - Final Assessment: Condition present
VXXB01aj_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Coronary artery disease / congestive heart failure (angina, heart attack) - Final Assessment: Condition present
VXXB01ak_FA	VXX		4	Onehot	0	(VXX) Medical conditions or diagnoses, Cardiac arrhythmias - Final Assessment: Condition present
VXXB01al_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, Kidney disease - Final Assessment: Condition present
VXXB01am_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, Urinary tract infection (provider diagnosed) - Final Assessment: Condition
VXXB01an_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, Sickle cell disease - Final Assessment: Condition present
VXXB01ao_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, Sickle cell trait - Final Assessment: Condition present
VXXB01ap_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, Thrombocytopenia - Final Assessment: Condition present
VXXB01aq_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, Anemia - Final Assessment: Condition present
VXXB01ar_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, History of blood clots (thrombosis or thromboembolism) or stroke - Final Assessment: Condition present
VXXB01as_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, Congenital or inherited bleeding disorder - Final Assessment: Condition present
VXXB01at_FA	VXX		4	Onehot	0.0004	(VXX) Medical conditions or diagnoses, Antiphospholipid syndrome (APA) or other acquired thrombophilia - Final Assessment: Condition present
VXXB01au_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Systemic lupus erythematosus (SLE) - Final Assessment: Condition present
VXXB01av_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Rheumatoid arthritis - Final Assessment: Condition present
VXXB01aw_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Other collagen vascular or autoimmune disease - Final Assessment: Condition
VXXB01ax_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Ulcerative colitis / Crohn's disease - Final Assessment: Condition present

NAME	FILE	RULE	TEMP	IMPUTE	MISSING	DESCRIPTION
VXXB01ay_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Cholestasis of pregnancy - Final Assessment: Condition present
VXXB01az_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Liver/gall bladder disease - Final Assessment: Condition present
VXXB01ba_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Cancer (malignancy) - Final Assessment: Condition present
VXXB01bb_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Mental health condition - Final Assessment: Condition present
VXXB01bc1_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Gynecological Conditions, Cervical dysplasia - Final Assessment: Condition Present
VXXB01bc2_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Gynecological Conditions, Fibroids - Final Assessment: Condition present
VXXB01bc3_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Gynecological Conditions, Polycystic ovary disease (PCOS) - Final Assessment: Condition present
VXXB01bd1_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, Gonorrhea - Final Assessment: Condition present
VXXB01bd2_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, Chlamydia - Final Assessment: Condition present
VXXB01bd3_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, Herpes - Final Assessment: Condition present
VXXB01bd4_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, Syphilis - Final Assessment: Condition present
VXXB01bd5_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, HIV/AIDS - Final Assessment: Condition
VXXB01bd6_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, Hepatitis B - Final Assessment: Condition
VXXB01bd7_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, Hepatitis C - Final Assessment: Condition
VXXB01bd8_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, Trichomonas - Final Assessment: Condition
VXXB01bd9_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Sexually transmitted diseases, Other - Final Assessment: Condition present
VXXB01be_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Yeast infection (provider diagnosed) - Final Assessment: Condition present
VXXB01bf_FA	VXX		4	Onehot	0.0005	(VXX) Medical conditions or diagnoses, Bacterial vaginosis - Final Assessment: Condition present
GEST_AGE_LOSS	S02	GEST_AGE	0	0		
S02A01a_1	S02		0		0.8565	(S02) Pregnancy #1 - Gestational age at time of loss, weeks
S02A01b_1	S02		0		0.8892	(S02) Pregnancy #1 - Gestation age at time of loss
S02A02_1	S02		0		0.7489	(S02) Pregnancy #1 - Spontaneous delivery or miscarriage
S02A01a_2	S02		0		0.9641	(S02) Pregnancy #2 - Gestational age at time of loss, weeks

Medical History Filtering

6.9 Outcomes

The outcomes filtering concerns post-pregnancy analysis of the newborn (all kinds of delivery), and variables that are highly predictive of preterm birth, such as major fetal conditions. Some of these variables may be used as class labels, such as pOUTCOME or GAwksEND, and others may provide privileged information for certain models. **The vast majority of this data is highly prone to class leakage and should not be used in modeling, but is used for data analysis.** Provided in the table are a **sample** of features that are useful in data analysis or as class labels.

Total # Features: 1165
Relevant Files: CBA, CBB, CBC, CMA, CMC, CPA, pregnancy_outcomes, S02, U02, U2A, U3A, V4A

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
CBAA04a	CBA		5		0.0051	(CBA) Apgar scores - One minute
CBAA04a_Check	CBA		5		0	(CBA) Apgar scores - One minute, not available
CBAA04b	CBA		5		0.0054	(CBA) Apgar scores - Five minute
CBAA04b_Check	CBA		5		0	(CBA) Apgar scores - Five minute, not available
CBAA04c	CBA		5		0.9609	(CBA) Apgar scores - Ten minute
CBAA04c_Check	CBA		5		0	(CBA) Apgar scores - Ten minute, not available
bw	pregnancy_outcomes		5		0.0602	Birth weight in grams based on CBAA05 (for livebirths) or CBCA04 (for stillbirths)
CBAA02	CBA		5 4		0.0013	(CBA) Gender of baby or fetus
pOUTCOME	pregnancy_outcomes		4		0.0352	Pregnancy outcome based on chart abstraction & A09
PROM	pregnancy_outcomes		4		0.0592	Premature rupture of membranes based on CMAC06c & CMAC06d
SPONTANEOUS	pregnancy_outcomes		4		0.0546	Spontaneous or indicated based on PROM and CMAC06
GAwksEND	pregnancy_outcomes		4		0.0374	Gestational age (weeks) at pregnancy end based on S02F01 and both chart abstraction & A09
TYPE_CA_A09	pregnancy_outcomes		4		0	Type of livebirth/stillbirth based on chart abstraction & A09, derived from multiple outcome variables

6.10 Treatment

The treatment filtering contains features that pertain to drug administration directly related to delivery or the prevention of preterm birth. This includes administration of steroids for fetal lung maturation, tocolytics, progesterone, and various classes of drugs used specifically for delivery. **These features are highly predictive of PTB, and are thus only used for analysis or sequential treatment decision making. Shown is a sample of these features.**

Total # Features: 149
Layer 1 # Features: N/A (27 Available)
Layer 2 # Features: N/A (34 Available)
Relevant Files: CMA
Dropped Features: For this group, there was much metadata that was used to place the administration within a particular time point, and then later dropped. Medication codes were used instead of names. Details on which particular drug within a class was used were dropped in L1, as the intended and possible effect were known and similar. Anticonvulsant details were returned in L2 as they can have dramatic differences.

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSINGNESS	DESCRIPTION
CMAF01	CMA		Date Abstraction	ANY	0.002	(CMA) Did the Participant receive steroids for fetal lung maturation at any time during this pregnancy?
CMAF02	CMA		Date Abstraction	ANY	0.0023	(CMA) Did the Participant receive progesterone at any time during this pregnancy?
CMAF01a	CMA		Date Abstraction	ANY	0.0011	(CMA) Steroids for fetal lung maturation: Betamethasone
CMAF01a3	CMA		Date Abstraction	ANY	0.9544	(CMA) Steroids for fetal lung maturation: Betamethasone - Total number of injections given over pregnancy
CMAF01b	CMA		Date Abstraction	ANY	0.0011	(CMA) Steroids for fetal lung maturation: Dexamethasone - Dexamethasone
CMAF01b3	CMA		Date Abstraction	ANY	0.9992	(CMA) Steroids for fetal lung maturation: Dexamethasone - Total number of doses given over pregnancy
CMAF02a	CMA		Date Abstraction	999	0.981	(CMA) Progesterone vehicle of administration:
CMAF02a_sp	CMA		Date Abstraction	999	0.9979	(CMA) Progesterone vehicle of administration: - Other, specify
CMAF02b1	CMA		Date Abstraction	-1	0.9979	(CMA) Progesterone dose - %
CMAF02b2	CMA		Date Abstraction	-1	0.989	(CMA) Progesterone dose - mg
CMAF02c	CMA		Date Abstraction	0	0.9743	(CMA) Progesterone frequency of administration
CMAF02c_sp	CMA		Date Abstraction	999	0.9875	(CMA) Progesterone frequency of administration - Other, specify
CMAG01	CMA		4	ANY	0.0021	(CMA) Did the Participant receive tocolysis at any time during this hospitalization for delivery?
CMAG02	CMA		4	ANY	0.0021	(CMA) Were antibiotics used anytime during this hospitalization for delivery?
CMAG03	CMA		4	ANY	0.002	(CMA) Were antihypertensives used anytime during this hospitalization for delivery?
CMAG04	CMA		4	ANY	0.002	(CMA) Were anticonvulsants used anytime during this hospitalization for delivery?
CMAG05	CMA		4	ANY	0.0018	(CMA) Was magnesium sulfate given specifically for neuroprophylaxis anytime during this hospitalization for delivery?
CMAG06	CMA		4	ANY	0.0021	(CMA) Were any other prescribed medications used during this hospitalization for delivery?
CMAG06a	CMA		4	2	0.0011	(CMA) Other prescribed medications used during this hospitalization for delivery - Insulin
CMAG06b	CMA		4	2	0.0011	(CMA) Other prescribed medications used during this hospitalization for delivery - Heparin / LMW heparin injection
CMAG06c1	CMA		4	777	0.7016	(CMA) Other prescribed medications used during this hospitalization for delivery - Medication code (1)
CMAG06d1	CMA		4	777	0.871	(CMA) Other prescribed medications used during this hospitalization for delivery - Medication code (2)
CMAG06e1	CMA		4	777	0.9373	(CMA) Other prescribed medications used during this hospitalization for delivery - Medication code (3)
CMAG06f1	CMA		4	777	0.9667	(CMA) Other prescribed medications used during this hospitalization for delivery - Medication code (4)
CMAG06g1	CMA		4	777	0.9821	(CMA) Other prescribed medications used during this hospitalization for delivery - Medication code (5)
CMAG06h1	CMA		4	777	0.9888	(CMA) Other prescribed medications used during this hospitalization for delivery - Medication code (6)
CMAG06i1	CMA		4	777	0.9935	(CMA) Other prescribed medications used during this hospitalization for delivery - Medication code (7)

6.11 Food Frequency Analysis

The food frequency analysis file takes into account the food and nutrients consumed by patients in the 3 months prior to conception. We were most interested in the calculated nutrient intake as opposed to less interpretable data like grams of food consumed or food pyramid group quantities. In addition to nutrients, we also attempted to capture the overall energy intake by using glycemic load and calorie intake as proxies.

Total Features: 737
Layer 1 Features: 38
Layer 2 Features: 38
Relevant Files: food_frequency_analysis (ancillary)
Dropped Features: For this group, we dropped the features that contained redundant information about food intake that was already captured by the vitamin amounts. For instance, quantities of food items such as "glasses of milk" would be reflected in the overall vitamin and calorie consumption.

Table 10: Food Frequency Filtering

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
GL	food frequency analysis		0	mean	0	Glycemic Load
DT_KCAL	food frequency analysis		0	mean	0	KCAL Total
DT_THEOB	food frequency analysis		0	mean	0	Theobromine, mg, average daily
FOL_DFE	food frequency analysis		0	mean	0	Average daily Dietary Folate Equivalents, mcg
DT_METHI	food frequency analysis		0	mean	0	Methionine (S-containing), grams
DT_CALC	food frequency analysis		0	mean	0	Calcium, mg
DT_IRON	food frequency analysis		0	mean	0	Iron, mg
DT_MAGN	food frequency analysis		0	mean	0	Magnesium, mg
DT_PHOS	food frequency analysis		0	mean	0	Phosphorus, mg
DT_POTA	food frequency analysis		0	mean	0	Potassium, mg
DT_SODI	food frequency analysis		0	mean	0	Sodium, mg
DT_ZINC	food frequency analysis		0	mean	0	Zinc, mg
DT_COPP	food frequency analysis		0	mean	0	Copper, mg
DT_SEL	food frequency analysis		0	mean	0	Selenium, mcg
DT_VARAE	food frequency analysis		0	mean	0	Vitamin A Retinol Activity Equivalents (RAE), mcg
DT_RET	food frequency analysis		0	mean	0	Retinol, mcg
DT_ACARO	food frequency analysis		0	mean	0	Alpha-carotene, mcg
DT_BCARO	food frequency analysis		0	mean	0	Beta-carotene, mcg
DT_CRYPT	food frequency analysis		0	mean	0	Cryptoxanthin, mcg
DT_LYCO	food frequency analysis		0	mean	0	Lycopene, mcg
DT_LUTZE	food frequency analysis		0	mean	0	Lutein-Zeaxanthin, mcg
DT_ATOC	food frequency analysis		0	mean	0	Vitamin E as alpha-tocopherol, mg
DT_VITC	food frequency analysis		0	mean	0	Vitamin C, mg
DT_THIA	food frequency analysis		0	mean	0	Thiamin (Vitamin B1), mg
DT_RIBO	food frequency analysis		0	mean	0	Riboflavin (Vitamin B2), mg
DT_NIAC	food frequency analysis		0	mean	0	Niacin, mg
DT_VITB6	food frequency analysis		0	mean	0	Vitamin B6, mg
DT_VB12	food frequency analysis		0	mean	0	Vitamin B-12, mcg
DT_VITK	food frequency analysis		0	mean	0	Vitamin K as phylloquinone, mcg
DT_FA182	food frequency analysis		0	mean	0	Dietary PUFA (∼N-6) 18:2, grams
DT_FA183	food frequency analysis		0	mean	0	Dietary PUFA (∼N-3) 18:3, grams
DT_FA184	food frequency analysis		0	mean	0	Dietary PUFA 18:4, grams
DT_FA204	food frequency analysis		0	mean	0	Dietary PUFA (∼N-6) 20:4, grams
DT_FA205	food frequency analysis		0	mean	0	Dietary N-3 PUFA 20:5 (EPA), grams
DT_FA225	food frequency analysis		0	mean	0	Dietary N-3 PUFA 22:5 (DPA), grams
DT_FA226	food frequency analysis		0	mean	0	Dietary N-3 PUFA 22:6 (DHA), grams
DT_TOTN6	food frequency analysis		0	mean	0	Avg. daily omega-6 FA, grams
DT_TOTN3	food frequency analysis		0	mean	0	Avg. daily omega-3 FA, grams

6.12 Sleep Substudy

The sleep substudy filtering includes sleep quantity and quality by calculating the average hours slept per night as well as sleep apnea diagnoses.

Total Features: 6
Layer 1 Features: 4
Layer 2 Features: 6
Relevant Files: V1L, V3L
Dropped Features: Due to sleep being a separate substudy, only a few features were selected to be included. The rest were dropped as of writing.

Table 11: Sleep Substudy Filtering

NAME	FILE	RULE	TEMPORAL	IMPUTE	MISSING	DESCRIPTION
V1LF01a	V1L		1	Onehot	0.00514	Have you ever been told by a doctor or other health professionals that you have sleep apnea or obstructive sleep apnea?
V3LF01a	V3L		3	Onehot	0.00464	Have you ever been told by a doctor or other health professionals that you have sleep apnea or obstructive sleep apnea?
V1A_SLEEP	V1L	SLEEP_AVG	1	Mean		
V1LA02a	V1L		1		0.00463	How many hours of sleep do you usually get per night: On weekdays or workdays?
V1LA02b	V1L		1		0.00463	How many hours of sleep do you usually get per night: On weekends?
V3A_SLEEP	V3L	SLEEP_AVG	3	Mean		
V3LA02a	V3L		3		0.00357	How many hours of sleep do you usually get per night: On weekdays or workdays?
V3LA02b	V3L		3		0.00357	How many hours of sleep do you usually get per night: On weekends?

6.13 Genetic data

The genotyped cohort comprises 9,757 nulliparous women from the nuMoM2b study who had adequate samples and agreed to be genotyped. DNA extractions from whole blood, which had been frozen at $-80°$, were carried out on a Qiasymphony instrument at the Center for Bioinformatics and Genomics (Indiana University). Genotyping was done at the Van Andel Institute (Grand Rapids, MI, USA) using the Infinium Multi-Ethnic Global D2 BeadChip (Illumina, Miami, USA). We imposed standard filters for quality control of loci at this stage (cluster separation < 0.3, AA R Mean < 0.2, AB R Mean < 0.2, BB R Mean < 0.2, 10% GC < 0.3) using GenomeStudio v2.4 (Illumina). Genotype calls (in .GCT format) for the 1,748,280 loci that passed initial quality control were made with Beeline autoconvert (Ilumina).

7 Lessons Learned and Discussion

The main goals of the data preprocessing of the nuMoM2b dataset were to get a deeper understanding of the data, extract the most relevant features, and reduce the dimensionality of the data. Additionally, we wanted to make the fewest assumptions when filling in unknown values, and to understand the intricacies behind the dependencies present in the data.

The original ratio of features to patients of around 4,600 to 10,000 is too high to create reliable models because of the high possibility of over-fitting in a high dimensional space. Therefore, it was crucial for the interdisciplinary team to work together to leverage medical expertise, computing and statistics skills in order to gain a good grasp

of the wealth of information in the nuMoM2b dataset. The data preprocessing team worked to make sure that any medical assumptions, such as grouping related conditions together or deciding to drop select details, were approved by the medical experts.

It was equally crucial to understand the protocol that doctors follow in the administration of interventions. Debates about definitions, classification of conditions, and treatments exist and were taken into consideration. A notable example is the debate around the usage of progesterone as an intervention for preterm birth. Another is simply the definitions of spontaneous and indicated preterm birth and how the events that fall under those categories may have changed over time.

With regards to the data itself, the most significant challenges were understanding the dependencies between different features and their medical relevance to preterm birth. The data includes both patient interviews and abstractions from the patients' charts. Sometimes the questions overlapped, other times they were parallel, or had features that combined information from both. Each source has a different level of relevance in building the models.

Standardization of data formatting was another obstacle. Much of the original labeling and information provided in the codebooks was not sufficient to directly begin the modeling process. For instance, labels such as variable type had to be manually added. It was also important to determine when specific data were relevant and when they were collected. For example, for data that were not collected exactly within the strict time designations for Visits 1 through 4, the associated dates were computed relative to the estimated date for the start of pregnancy. Manual inspection was also required when coded values of responses did not match the information shown in the data collection forms. Codings such as *Don't Know*, or *Refused response* did not exist in the data even though they were mentioned in the codebooks.

Throughout the preprocessing, we aimed to make the filtering and imputation steps as systematic as possible, while organizing the data into medically homogeneous groups. The discussion here is merely a glimpse of the intended goals and actions taken during the preprocessing. As our team delves into the PTB-related research goals, we anticipate that more preprocessing will be required on specific data.

8 Institutional review board statement and funding sources

Human subjects approval for this study, titled "SCH: Prediction of Preterm Birth in Nulliparous Women", was obtained following review by Columbia University Human Subjects Institutional Review Board under number IRB-AAAR9413, and the City University of New York CUNY HRPP/IRB review number 2019-0855. Human subjects training requirements were completed by all authors of this study.

This work is supported by NIH/NLM (www.nlm.nih.gov) grant R01LM013327.

References

[1] `http://www.nichd.nih.gov/research/supported/Pages/nuMoM2b.aspx`.

[2] Clinical Informatics Group. Data pre-processing for the preterm prediction study MFMU dataset. `http://www.cs.columbia.edu/~ansaf/cing/CCLS-13-04.pdf`, 2013.

[3] David Haas et al. A description of the methods of the nulliparous pregnancy outcomes study: monitoring mothers-to-be (numom2b). *American Journal of Obstetrics and Gynecology*, 212(4):539.e1–539.e24, 2015.

[4] Ilia Vovsha, Ashwath Rajan, Ansaf Salleb-Aouissi, Anita Raja, Axinia Radeva, Hatim Diab, Ashish Tomar, and Ronald Wapner. Predicting preterm birth is not elusive: Machine learning paves the way to individual wellness, 2014.

[5] Ilia Vovsha, Ansaf Salleb-Aouissi, Anita Raja, Thomas Koch, Alex Rybchuk, Axinia Radeva, Ashwath Rajan, Yiwen Huang, Hatim Diab, Ashish Tomar, and Ronald Wapner. Using kernel methods and model selection for prediction of preterm birth. In Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 55–72, Children's Hospital LA, Los Angeles, CA, USA, 18–19 Aug 2016. PMLR.