Predicting Preterm Birth Is Not Elusive: Machine Learning Paves the Way to Individual Wellness

Ilia Vovsha[*], Ashwath Rajan[*], Ansaf Salleb-Aouissi[*], Anita Raja[**], Axinia Radeva[*], Hatim Diab[*], Ashish Tomar[*] and Ronald Wapner[***]

[*] Columbia University Center for Computational Learning Systems (CCLS) New York, NY 10115, USA
{iv2121, anr2121, ansafsalleb, aradeva, hdiab, ast2124}@columbia.edu
[**]UNC Charlotte, Software and Information Systems, Charlotte, NC 28223. anraja@uncc.edu
[***]Columbia University Medical Center. Department of Obstetrics and Gynecology. rw2191@cumc.columbia.edu

Abstract

Preterm birth is a major public health problem with profound implications on society, there would be extreme value in being able to identify women at risk of preterm birth during the course of their pregnancy. Previous research has largely focused on individual risk factors correlated with preterm birth and less on combining these factors in a way to understand the complex etiologies of preterm birth. In this paper, we use the "Preterm Prediction Study," a clinical trial dataset collected by the National Institute of Child Health and Human Development (NICHD) – Maternal-Fetal Medicine Units Network (MFMU). We summarize two years of efforts to collect, prepare and process this dataset with a special emphasis to solve a so far elusive problem of predicting preterm birth in nulliparous (first time) mothers. Our approach includes comparison of two approaches for deriving predictive models: an SVM approach with linear and non-linear kernels and logistic regression with different model selection procedures. We demonstrate significant improvement compared to past work on this dataset while stressing the challenges we faced in data preparation and analysis.

1 Problem Description

Premature or preterm birth (PTB) is defined as birth before 37 completed weeks of gestation. It is a major long-lasting public health problem with heavy emotional and financial consequences to families and society. Over 26 billion dollars are spent annually on the delivery and care of the 12-13% of infants who are born preterm in the United States (Behrman et al. 2007). PTB is the leading cause of neonatal mortality and, long-term disabilities. These range from visual/hearing impairment, cerebral palsy, mental retardation as well as an increased likelihood of cardiovascular disease, hypertension, and diabetes later in life (Zerhouni 2008). The lower the gestational age at birth, the longer the hospital stay and cost and the greater the risk of long term sequelae.

The World Health Organization, the National Institute of Health (Zerhouni 2008) and particularly the NICHD (NICHD 2012), have recognized the need to discover the multifactorial etiologies of PTB and to use these to accurately predict at-risk pregnancies.

A crucial challenge is to identify women who are at the highest risk for very early preterm birth and to develop interventions. Equally important, would be the ability to identify women at the lowest risk to avoid unnecessary and costly interventions. Because of the multifactorial etiology of preterm birth, understanding and predicting this adverse pregnancy outcome remains unresolved despite immense efforts.

Previous research has largely focused on individual risk factors correlated with preterm birth and less on combining these factors in a way to understand the complex etiologies of preterm birth. These factors include prior preterm birth, black race, multiple gestations, and infection. Among these, the strongest and widely used risk factor among clinicians is having prior preterm births. Those judged at risk for PTB are typically treated by prenatal administration of progesterone (Flood and Malone 2012). To date, progesterone is considered as the most common treatment option (Acog 2008). History of PTB is obviously not available information for women who will give birth for the first time (nulliparas) – these represent 40% of pregnant women in the US (Zerhouni 2008). Nulliparas often go untreated due to a lack of predictability given the unknown combination of other factors involved. If it were possible to develop a reliable risk predictor of PTB for first-time mothers, it could substantially reduce the incidence of PTB and its consequences. Today, there is no widely tested prediction system that combines well-known factors (Davey et al. 2011) with a good prediction technique to provide actionable decisions within a clinical environment (Mercer et al. 1996).

The high-level problem we are interested in solving is how we can derive powerful prediction models from large medical data that would be relevant to a particular individual. Specifically, how can we reliably determine the risk of a nulliparas mother giving birth prematurely. In the process of solving this problem, we have developed an approach that also improves prediction of PTB occurrence in the general case as well.

In this paper, we describe this approach as an application of Machine Learning toward the problem of predicting preterm birth. We use the "Preterm Prediction Study," a clinical trial dataset collected by the National Institute of child Health and Human Development (NICHD) – Maternal-Fetal Medicine Units Network (MFMU). This paper summarizes two years of efforts by our multidisciplinary team to collect, prepare and process this dataset. Specifically, we compare two approaches for deriving predictive models: an SVM approach with linear and non-linear kernels and a logistic regression approaches with different model selection. We focus our attention on predicting (1) any kind of preterm birth, (2) spontaneous preterm birth, and (3) predicting preterm birth for nulliparous (first time) mothers. We also derive models at different time points representing the three main prenatal visits in the preterm prediction study.

This paper is organized as follows: We first provide an overview of the risk factors and state-of-the-art systems for predicting PTB. We then describe the Preterm Prediction study dataset, our approach and our empirical evaluation. We finally conclude with the significance and impact of this study along with future work.

2 Background

In this section, we first describe the known risk factors for PTB. We then present state-of-the-art approaches to devise a risk-scoring system for PTB.

2.1 Risk factors for preterm birth

Approximately 30% of preterm deliveries are indicated based on maternal or fetal conditions such as mother's preeclampsia, and intra uterine growth restriction. The remaining 70%, spontaneous PTB (SPTB) occur following the onset of spontaneous preterm labor, prelabor Premature Rupture Of the Membranes (pPROM) or cervical insufficiency (Goldenberg et al. 2008). Spontaneous preterm labor is a heterogeneous condition, the final common product of numerous biologic pathways that include immune, inflammatory, neuroendocrine, and vascular processes (Behrman et al. 2007). Those judged at risk for PTB are typically treated by prenatal administration of progesterone 17 OHPC (IM progesterone) (Acog 2008; Flood and Malone 2012). Epidemiological investigations have largely associated single factors with PTB. Of the many risk factors for preterm labor, a prior history of preterm delivery is the most predictive with a recurrence risk as high as 50% depending on the number and gestational age of previous deliveries (Goldenberg et al. 2008). In the United States, women classified as black, African-American, and Afro-Caribbean are at an increased risk (Goldenberg et al. 2008). Other risk factors include low socioeconomic status, extremes in age, single marital status (Smith et al. 2007; Thompson et al. 2006); low prepregnancy body mass index, (Hendler et al. 2005); and high-risk behaviors during pregnancy (e.g. tobacco, cocaine and heroin use). Psychological factors that have been proposed to increase the risk of PTB include depression (Grote et al. 2010), and high levels of stress during pregnancy (Copper et al. 1996). Obstetrical conditions associated with the onset of spontaneous preterm labor are heterogeneous and include intrauterine infection, uteroplacental ischemia and hemorrhage, vascular lesions of the placenta, uterine over distention and cervical insufficiency (Goldenberg, Hauth, and Andrews 2000). Additional pregnancy factors associated with preterm delivery include closely spaced gestations (Conde-Agudelo, Rosas-Bermudez, and Kafury-Goeta 2006), and multiple gestations (Goldenberg et al. 2008). Assisted reproductive technologies may also constitute a risk factor, for unknown reasons (Allen, Wilson, and Cheung 2006). Finally, environmental toxins have been suggested as risk factors, though results are conflicting (Behrman et al. 2007). Exposure to tobacco smoke appears to carry the greatest risk (Kharrazi et al. 2004; Jaakkola, Jaakkola, and Zahlsen 2001). Besides these, there also seems to be a genetic contribution to preterm birth (Porter et al. 1997). There are other critical factors that have been used for predicting PTB by some practices. It has been shown (Goldenberg et al. 1998) that the odds ratio of SPTB was highest for a positive fetal fibronectin test followed by short cervix and history of prior PTB. In fact these factors along with bacterial vaginosis were more strongly associated with early than late SPTB. Systematic studies of multiple published studies (Crane and Hutchens 2008) showed that cervical length (threshold of 25 mm) measured by transvaginal ultrasonography in asymptomatic (no uterine tightening/contractions) high-risk women is highly predictive of spontaneous preterm birth before 35 weeks. However, in practice, prior history of preterm delivery is used as the most predictive indicator of PTB in most clinical settings.

2.2 Risk scoring systems for predicting preterm birth

Papiernik proposed in the late sixties an empirical method for estimating the risk of premature delivery denoted "The Coefficient of Premature Delivery Risk (CPDR)" (Papiernik-Berkhauer 1969). In this approach, maternal characteristics are grouped into four series of comparable variables (social status, obstetric history, work conditions, pregnancy characteristics) in a two-dimensional table. A number of points varying from 1 to 5 according to the degree of their importance is assigned to each characteristic. The sum of the points gives the risk of Premature delivery. Papiernik's risk table was later modified by Creasy et al. and used in the risk of preterm delivery (RPD) system proposed in (Creasy, Gummer, and Liggins 1980). Further assessment of the prediction performance of Creasy's table (Edenfield et al. 1995) on another population has shown low performance. A graded risk system was proposed by (Mercer et al. 1996) in the context of the NICHD MFMU preterm prediction study. The results of a multivariate logistic regression were modest (sensitivity of 24.2% and 18.2%; Specificity of 28.6% and 33.3%, respectively for nulliparous and multiparous women).

This study was a first step toward combining factors and showed promise that using more sophisticated techniques could better identify women at risk of PTB. Today, there is no widely tested risk scoring/prediction system that combines PTB factors (Davey et al. 2011).

Goodwin and her colleagues have explored the use of data mining techniques to predict preterm labor. In early work (Woolery and Grzymala-Busse 1994), they show that it is feasible to use machine learning to generate expert system (knowledge-base) rules for prediction of preterm delivery. In subsequent work (Goodwin et al. 2001), data from the Duke University medical center in North Carolina consisting of 19,970 patient records and 1,622 variables was studied using data mining techniques. In this paper, the authors claim to have identified a parsimonious set of seven demographic variables (maternal age, marital status, race, education, patient insurance category, county and religion) that play a more critical role as predictors of preterm birth for their data set. While these results are interesting, there are concerns whether the sampling of a particular demographic (academic medial center) would be representative of more general population data. Furthermore, their experiment procedure is unclear – for example, the AUC could have been obtained on a validation set or an unseen test set; consequently it is difficult to reproduce their results.

Courtney et al. (Courtney et al. 2008) describe a secondary analysis showing that the demographic preterm prediction model generated by (Goodwin et al. 2001) generalizes to a broader population with a modest accuracy. We believe our models are deeper and more powerful than this previous work for several reasons: (1) the dataset under study represents a diverse population from ten medical centers across the US, (2) we derive predictive models at different stages in pregnancy, and (3) we derive models for nulliparous women and also for spontaneous PTB.

3 The Preterm Prediction Study Data

We have obtained the released data set for the *Preterm Prediction Study*, performed by the NICHD Maternal Fetal Medicine Units (MFMU) Network between 1992 to 1994. This study is an observational prospective study of 3,073 women with singleton pregnancies recruited at less than 24 weeks gestational age. Of the women enrolled, 2,929 participated in the study at the 10 participating MFMU centers between 10/1992 and 07/1994. There were 1,711 multiparous and 1,218 nulliparous women. The incidence of spontaneous preterm birth was 10.3% overall 8.2% for nulliparous and 11.9% for multiparous women (Mercer et al. 1996). Henceforth we will refer to this data as the *MFMU data*. Participating women in this study have been followed by research nurses during four visits at 24 (T0), 26 (T1), 28 (T3) and 30 (T4) weeks gestation for screening tests. The MFMU data timeline is illustrated in Figure 1. The data collected at all visits altogether (Maternal Fetal Medicine Units Network 1994) includes over 400 variables: demographic, behavioral, medical history, previous and current pregnancy history, digital cervical examination, vaginal ultrasound, cervical and vaginal fetal fibronectin, KOH prep for yeast tests and a psychosocial questionnaire. The detailed outcome is shown in Table 1. We have grouped the features into categories as depicted in Figure 1. At each visit, a set of feature groups is collected. We focus our study on the three major visits at time T0, T1 and T3. T0 is composed of 3,002 examples described by 50 features. T1 dataset contains data for 2,929 examples with 205 features. Finally, T3 is composed of 2,549 examples with 316 features. This includes previous information and data collected during the second visit (minor) and the third visit (major).

Outcome	N	%
Spontaneous PTB <32 weeks	50	2
Spont PTD <35 weeks	129	4
Spont PTD <37 weeks	309	10
Indicated PTD <37 weeks	124	4
Fetal growth retardation	163	5
Low birth weight	361	12

Table 1: Outcome measures in the MFMU dataset.

This dataset is unique in many ways: it is a well-curated dataset of women with singleton pregnancy from ten centers. Remarkably, current PTB treatments were not in use in the early nineties at the time where this data was collected. This makes this dataset [1] compelling as we have natural incidence of PTB independent of any treatment.

4 Methods and Technical Solutions

In this section, we first describe data preparation, then we present the methods we used for prediction of PTB, Specifically, we consider SVMs and regression methods with model selection.

4.1 Data preparation

The MFMU dataset was compiled from a set of tests and a detailed questionnaire administered to mothers (patients) over the entire duration of the pregnancy. As a result, multiple processing steps are required to harness this very rich and highly structured data. We face three specific challenges,

1. *Missing Data:* A substantial number of features is missing, both randomly (a particular question was not completed) and structurally (dependency between questions).
2. *Varying Sample Size:* The number of examples changes over time as patients give birth, withdraw, or miss a visit.
3. *Skewed Class Distributions:* By the nature of the preterm birth problem, class sizes are significantly uneven. This fact is relevant to subsets of the data as well (e.g., when we consider nulliparous mothers only).

We handle the complexity of the data by organizing features into groups (according to the original questionnaire and the definition). In addition, we keep track of the number and type of missing values, and the processing steps taken to complete or delete features. Our main objective is to retain as many features and examples as possible, while converting the data into a standard format suitable for off-the-shelf machine learning algorithms. We believe that the MFMU data (if properly processed) is an exciting application from a machine learning perspective. We describe our preprocessing steps in (Clinical Informatics Group 2013).

4.2 Method for prediction of PTB

In this paper, we consider PTB prediction as a binary classification problem, where patients who deliver a baby preterm (full-term) are assigned the positive (negative) class respectively. At every tick $(0, 1, 3)$, each patient (example) is described by a complete feature vector (see section 3) and a label (x_i, y_i), $y_i \in \{+1, -1\}$.

We apply various logistic regression and Support Vector Machines (SVM) methods and compare their error rates.

[1] We plan on releasing raw and preprocessed MFMU datasets as a benchmark in the UCI repository by end of 2013.

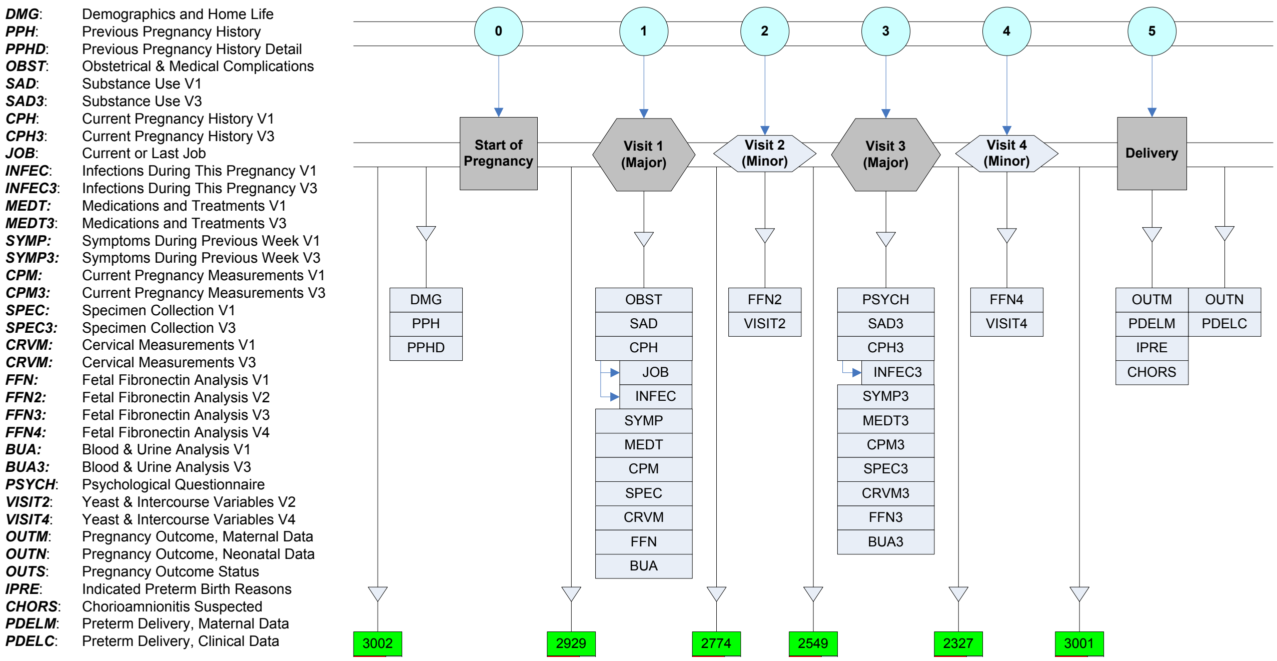
DMG: Demographics and Home Life
PPH: Previous Pregnancy History
PPHD: Previous Pregnancy History Detail
OBST: Obstetrical & Medical Complications
SAD: Substance Use V1
SAD3: Substance Use V3
CPH: Current Pregnancy History V1
CPH3: Current Pregnancy History V3
JOB: Current or Last Job
INFEC: Infections During This Pregnancy V1
INFEC3: Infections During This Pregnancy V3
MEDT: Medications and Treatments V1
MEDT3: Medications and Treatments V3
SYMP: Symptoms During Previous Week V1
SYMP3: Symptoms During Previous Week V3
CPM: Current Pregnancy Measurements V1
CPM3: Current Pregnancy Measurements V3
SPEC: Specimen Collection V1
SPEC3: Specimen Collection V3
CRVM: Cervical Measurements V1
CRVM: Cervical Measurements V3
FFN: Fetal Fibronectin Analysis V1
FFN2: Fetal Fibronectin Analysis V2
FFN3: Fetal Fibronectin Analysis V3
FFN4: Fetal Fibronectin Analysis V4
BUA: Blood & Urine Analysis V1
BUA3: Blood & Urine Analysis V3
PSYCH: Psychological Questionnaire
VISIT2: Yeast & Intercourse Variables V2
VISIT4: Yeast & Intercourse Variables V4
OUTM: Pregnancy Outcome, Maternal Data
OUTN: Pregnancy Outcome, Neonatal Data
OUTS: Pregnancy Outcome Status
IPRE: Indicated Preterm Birth Reasons
CHORS: Chorioamnionitis Suspected
PDELM: Preterm Delivery, Maternal Data
PDELC: Preterm Delivery, Clinical Data
0
1
2
3
4
5
Start of Pregnancy
Visit 1 (Major)
Visit 2 (Minor)
Visit 3 (Major)
Visit 4 (Minor)
Delivery
DMG
PPH
PPHD
OBST
SAD
CPH
JOB
INFEC
SYMP
MEDT
CPM
SPEC
CRVM
FFN
BUA
FFN2
VISIT2
PSYCH
SAD3
CPH3
INFEC3
SYMP3
MEDT3
CPM3
SPEC3
CRVM3
FFN3
BUA3
FFN4
VISIT4
OUTM
PDELM
IPRE
CHORS
OUTN
PDELC
3002
2929
2774
2549
2327
3001

Figure 1: Illustration of MFMU data timeline and description of the set of feature groups. The numbers at the bottom of the diagram indicate the number of patients that reached that point in time of the study. These numbers decrease with time for several possible reasons including: patients withdrawing from study/delivered/lost to follow up/skipped major visit etc.

To make results more robust, we repeat the following procedure throughout: each dataset is randomly divided into train and test sets with an 80/20 ratio, and each class is split proportionally between the sets. We then apply 5-fold cross-validation (CV) to the train set, to determine the best model and optimal parameters (if any). The best model is tested on the (unseen) test set, and confusion matrices for various subsets of the data are recorded.

Support Vector Machines We use a support vector machine with a linear and Radial Basis function (RBF) kernels. The standard approach for binary problems is to use the soft-margin dual formulation (Boser, Guyon, and Vapnik 1992; Cortes and Vapnik 1995).

In order to allow the algorithm to handle the skewness in the dataset, we scale the hinge loss penalty from the cost function proportionally to the size of each class. The cost function is thus a slightly modified version of the classical SVM cost:

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}||w||^2 + C_- \sum_{y_i=-1} \xi_i + C_+ \sum_{y_j=+1} \xi_j \\ \text{s.t.} \quad & y_k[w^\top x_k + b] \geq 1 - \xi_k, \quad \forall k, \end{aligned}$$

where x's are the feature vectors, y's are the labels, w is learned weight vector, the ξ values are slack variables used during optimization, and C_+ and C_- are the regularization parameters.

By assigning different misclassification costs, we can give equal overall weight to each class in measuring performance. In order to avoid having to tune two cost parameters, we set $C_+n_+ = C_-n_-$ where n_+, n_- are the number of positive (negative) examples (Ben-Hur and Weston 2010).

To be able to construct nonlinear surfaces we use the radial basis function (RBF) kernel,

$$K(x, y) = exp(-\gamma \|x - y\|^2).$$

Logistic and Lasso regression Our regression study is motivated by the desire to create a meaningful baseline model to evaluate the performance of linear models in this problem space. Regression methodologies are widely used within the biostatistics and medical domain for the purpose of prediction.

To deal with the skewed dataset, we use the oversampling techniques (e.g. Adasyn (He et al. 2008)) to achieve 1:1 levels of negative to positive examples.

We consider four logistic regression model selection methodologies: forward selection, stepwise selection, l_1 lasso regression, and elastic net regression (Zou and Hastie 2005; Tibshirani 1996). Forward selection is a greedy algorithm, which at each step, selects the covariate that best improves the fit, until adding covariates is no longer productive. In forward selection, once a predictor is added, it is never removed. Stepwise selection is very similar, but at each step, also considers removing each already added covariate. Lasso regression uses an (l_1 norm) penalty to encourage sparse solutions and perform a level of feature selection. The lasso is well-documented methodology in biostatistics , and has been shown to perform well compared to forward/stepwise selection models.

Because of the nature of the structure we have imposed on the data, there is a chance that there will be correlation amongst groups of the predictor variables, both in an incidental manner, where certain predictors are intrinsically but non-obviously correlated (e.g., drugs and cigarettes), and in a structural manner, where certain predictors are grouped, and obviously connected. It is with this in mind that we also construct an elastic net model, which combines the sparsity induction of the l_1 norm to eliminate the trivial covariates, whilst using the ridge regression l_2 norm to automatically include whole groups of collinear predictors once a single covariate in added(Zou and Hastie 2005).

With these four regression models, we get a good picture of the nature of the problem: the two subset selection methodologies are prone to high variance in comparing selected models from sampled datasets, but provide us with a baseline model to which we may compare the lasso and elastic net methodologies. This gives us great perspective, as all of the models produce highly interpretable models; what is more, the relative accuracies of the models will grant us information about the nature of the most influential predictors. For example, having the lasso under perform the stepwise/forward selection models would suggest that the most important predictors are often collinear/grouped, with lasso driving to 0 many of the grouped predictors.

Finally, we use the receiver operating characteristic (ROC) to help us choose a suitable threshold value for the logistic regression cut-off. We use the metric of threshold value that provides the closest value to the top left of the ROC curve space, which corresponds to sensitivity=specificity= 1.

The problem of diagnosing preterm birth is an important one. Within the medical domain, it is intelligent to ask the relative benefits of high sensitivity vs. high specificity; in our case, higher sensitivity is valuable, as greater detection of true positives would allow doctors to intervene in the course of the mother's pregnancy. Because the cost of intervention is small, and the benefits large, we designed our ROC threshold search for two situations: one where the relative weights of the sensitivity and specificity were equal, and one were sensitivity was valued twice more than specificity in the objective function:

$$\min((1 - \text{sensitivites})^2) + r \times (1 - \text{specificities})^2)$$

where r is the weighting factor.

We use the glmnet package, which uses coordinate descent to train the elastic net and lasso models, and the built-in function STEP to train the subset selection models (Friedman, Hastie, and Tibshirani 2010). We use the step function within R to train the forward and stepwise models, which in turn use the AIC to base the next step of subset selection.

5 Empirical Evaluation

Recent work has acknowledged the unclear boundary between spontaneous and indicated PTB as they share substantial etiologic overlap (Klebanoff and Keim 2011). Hence, we decided to focus our study on distinguishing between any kind of preterm birth and full term birth. In addition, we considered the more difficult problems of spontaneous versus full term births, and finally preterm birth in nulliparous (mothers-to-be) versus full term. For each of the three problems above, we derive prediction models at different time points (ticks). Each tick (T0,T1,T3) represents a critical point (major visit) at which information is collected.

We present averaged sensitivity, specificity, and g-mean rates for each algorithm.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$
$$\text{G-Mean} = \sqrt{\text{Sensitivity} * \text{Specificity}}$$

Since the negative class is the majority class, it is not difficult to obtain high specificity rates at any tick. However, we would like to have a fair balance between sensitivity and specificity, hence our choice to use the geometric mean metric as a well suited performance measure.

We present our results in Table 2. Each cell shows the specified measure for the unseen test set, averaged over five runs.

Table 3 lists the most important features for classification obtained by running the linear SVM algorithm. The features are ranked according to their overall importance in all five runs.

5.1 Observations

We can make several general observations based on the experiments:

1. A graded risk system was proposed by (Mercer et al. 1996) in the context of the preterm prediction study. As stated earlier, the results of a multivariate logistic regression were modest with a sensitivity of 24.2% and 18.2%; Specificity of 28.6% and 33.3%, respectively for nulliparous and multiparous women. Our study demonstrates that model selection and non linear kernels are promising approaches for prediction of PTB.
2. Linear SVM provides a robust baseline for the quality of performance one can expect from algorithms applied to this data. Sensitivity rate of 40% and G-Mean of 57-60% are obtained. While Creasy's table (Creasy, Gummer, and Liggins 1980), which is a "hand-picked" linear model, wasn't very effective (Edenfield et al. 1995), we have showed that a "machine-picked" linear model is reasonable. We have attempted to use Creasy's table on this data but the feature mapping turned out to be difficult.
3. Comparing our work to Goodwin et al. (Goodwin et al. 2001), we consider our methodology to be more transparent and the results reproducible. Furthermore, they claim that the outcome is more or less consistent across the timeline (ticks). We actually see an improvement with linear/RBF SVM from T0 to T3 when we consider the full data. Finally, they mention seven important demographic variables. If we look at the top weights for linear-SVM in Table 3, we can see that four of them are present: Marital status, Race, Mom-Age and Insurance. However, we also see that previous pregnancy data can be even more important.

Preterm vs. Full term, All data

	Sensitivity			**Specificity**			**G-mean**		
	T0	T1	T3	T0	T1	T3	T0	T1	T3
Lasso	0.591	0.515	0.505	0.59	0.672	0.727	0.589	0.587	0.604
Elastic Net	0.591	0.512	0.502	0.587	0.672	0.731	0.588	0.586	0.604
Linear SVM	0.404	0.427	0.454	0.825	0.821	0.84	0.575	0.591	0.616
RBF SVM	0.576	0.548	0.594	0.621	0.724	0.719	0.596	0.63	0.652

Preterm vs. Fullterm, Spontaneous only

	Sensitivity			**Specificity**			**G-mean**		
	T0	T1	T3	T0	T1	T3	T0	T1	T3
Lasso	0.528	0.349	0.356	0.54	0.658	0.666	0.533	0.477	0.483
Elastic Net	0.521	0.357	0.356	0.547	0.652	0.674	0.533	0.481	0.486
Linear SVM	0.498	0.534	0.468	0.498	0.514	0.569	0.49	0.523	0.515
RBF SVM	0.403	0.403	0.428	0.594	0.601	0.584	0.49	0.489	0.5

Preterm vs. Fullterm, Nulliparous only

	Sensitivity			**Specificity**			**G-mean**		
	T0	T1	T3	T0	T1	T3	T0	T1	T3
Lasso	0.361	0.347	0.310	0.577	0.682	0.75	0.455	0.483	0.473
Elastic Net	0.354	0.347	0.302	0.583	0.686	0.755	0.452	0.485	0.47
Linear SVM	0.395	0.4	0.417	0.588	0.604	0.657	0.48	0.488	0.519
RBF SVM	0.406	0.341	0.424	0.637	0.643	0.679	0.5	0.46	0.533

Table 2: Average test rates for different populations at each tick.

4. The observed SVM performance is obtained without requiring any additional data processing (beyond what is described in (Clinical Informatics Group 2013)). We use the unbalanced version of SVM along with the G-Mean metric and do not resort to synthetic sampling techniques to artificially balance the classes.

5. When we consider the entire (full) dataset, we perform better with increasing ticks (as the pregnancy progresses). This reflects our intuition as we are getting more and more information (features) about each patient (example), and hence expect to better discriminate between them.

6. SVM with a non-linear (RBF) kernel outperforms linear SVM for the full data. For spontaneous only data, there is no improvement by using non-linear SVM or increasing the tick. This suggests that the data is not suitable (due to noise or some other reason) to discriminate between classes, and further feature processing may be required.

7. The high number of support vectors required for the SVM solution throughout the SVM runs (across ticks, kernels, data) suggests that preferable decision rule is approximately linear. In other words, under-fitting the data (small C value) generalizes better to unseen examples.

8. We consider the Nulliparous data only to be the most difficult of the three datasets. This is especially clear at tick 0 when most of the critical features come from previous pregnancy history which is not available for nulliparous women.

9. A nearest neighbor algorithm which we applied to the data suggests that the examples of the minority class (preterm) have few minority class neighbors. In other words, the minority examples are isolated among the majority class examples. As a result, it is difficult to obtain meaningful nonlinear classifiers that would generalize well (and hence the tendency to select a small C value in all SVM experiments).

10. This study is a step towards harnessing the wealth of information encapsulated in vast amount of health data to derive powerful prediction models and using this knowledge to improve personal health and wellness.

6 Significance and Impact

Developing a reliable risk prediction system for diseases is a critical step towards better personal health and wellness. This usually involves deriving powerful models from health-related big data. In this paper, we describe our efforts towards developing a preterm birth prediction system to be used clinically to reduce its incidence and consequences.

Dr. Creasy qualified PTB as a "vexing national problem" (Creasy 1993). It truly is, if we consider the increasing incidence of PTB in the US and worldwide despite the immense efforts spent in the last decades to solve it.

Today, there is no prediction system to identify women at risk of PTB to prevent this adverse pregnancy outcome. Specifically, nulliparous women (first time mothers-to-be) remain the most vulnerable population. We have demonstrated in this work that prediction of PTB is not an elusive goal to achieve. Our experiments with support vector machines, and regression techniques have shown that we significantly improve on the existing results. Moreover, this methodology is used off-the-shelf to handle very unbalanced classes of examples, large number of features, and different feature types. Consequently, it is not necessary to remove features or examples to produce meaningful results, and we discard virtually no data throughout our study. However, more work is needed to achieve models that can be

T0	T1	T3
Preterm delivery	Preterm delivery	Preterm delivery
#Term delivery	#Term delivery	Membranes protruding into cervix
# Yrs. since last pregnancy	Cervical consistency	Uterine contractions in last 2 wks
Marital status	Membranes protruding into cervix	Cervical consistency
Race	Vaginal bleeding, 1st or 2nd trim.	# Yrs. since last pregnancy
# Prev. preterm deliveries	#Term delivery	Cervical position
Mom age	Visit 1 FFN (wk 23-24)	Use of car
Use of car	Summary FFN (wk 23-24)	Activity restriction
Parity	Race	Visit 3 FFN (wk 27-28)
#Prev. PROM 20-36 weeks	Cervical FFN +/- (wks 23-24)	Major complications since Visit 1
Insurance	Uterine contractions in last 2 wks	Uterine contractions in last 2 wks
Total number of pregnancies	Marital status	Summary FFN (wk 27-28)
Income	#Prev. preterm deliveries	Marital status

Table 3: Rank order of the top features in predicting Preterm vs. Full term (all data) using Linear SVM.

used clinically to reliably identify women at risk.

Future work will address conducting larger scale experiments on two sources of data. First, data collected from electronic health records at a large urban hospital. Second, a new rich clinical trial data of 10,000 nulliparous women with a broad range of features including genetic data. We are also working on a novel decision theoretic approach for sequential action recommendation for PTB prevention that will account for dynamic temporal trends in patient history.

Preterm birth is a challenging and complex real world problem that pushes the boundary of state-of-the-art data mining methodologies. This is due to the inherent nature of pregnancy data that can be qualified as dynamic, noisy, containing missing data, groups of variables, probably missing important factors (e.g., genetic), skewed (due to incidence of PTB), and often with multiple overlapping classes (e.g. patient may experience a preterm labor and a pPROM). While SVM seem to handle unbalanced datasets well, and produced better results than past work on this dataset, several questions that are pertinent to PTB data as well as big data in general remain including: how can we avoid filling in missing values (e.g., it does not make sense to fill in history of PTB for nulliparous women), why is this problem so intractable? how to deal with overlapping classes? and how much data is needed for prediction?

Acknowledgment and Disclaimer

This research was supported in part by a Research Initiative in Science and Engineering (RISE) grant from Columbia University Office of Vice President for Research. We acknowledge the assistance of NICHD, the MFMU Network, and the Protocol Subcommittee in making the database available on behalf of the project. We are thankful to the physicians involved in this project for bringing in domain knowledge and valuable insights in interpreting the data. The contents of this paper represent the views of the authors and do not represent the views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network or the National Institutes of Health.

References

Acog. 2008. ACOG Committee Opinion. Use of progesterone to reduce preterm birth. (419 (Replaces No. 291, November 2003)).

Allen, V. M.; Wilson, R. D.; and Cheung, A. 2006. Pregnancy outcomes after assisted reproductive technology. *Journal of obstetrics and gynaecology Canada* 28(3):220 – 50.

Behrman, R.; Butler, A.; of Medicine (U.S.). Committee on Understanding Premature Birth, I.; and Outcomes, A. H. 2007. *Preterm birth: causes, consequences, and prevention*. National Academies Press.

Ben-Hur, A., and Weston, J. 2010. A user's guide to support vector machines. 609:223–239.

Boser, B.; Guyon, I.; and Vapnik, V. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, 144–152. New York, NY, USA: ACM.

Clinical Informatics Group. 2013. Data pre-processing for the preterm prediction study MFMU dataset.

Conde-Agudelo, A.; Rosas-Bermudez, A.; and Kafury-Goeta, A. C. 2006. Birth spacing and risk of adverse perinatal outcomes: a meta-analysis. *JAMA : the journal of the American Medical Association* 295(15):1809 – 23.

Copper, R. L.; Goldenberg, R. L.; Das, A.; Elder, N.; Swain, M.; Norman, G.; Ramsey, R.; Cotroneo, P.; Collins, B. A.; Johnson, F.; Jones, P.; and Meier, A. 1996. The preterm prediction study: Maternal stress is associated with spontaneous preterm birth at less than thirty-five weeks' gestation, ,. *American Journal of Obstetrics and Gynecology* 175(5):1286 – 1292.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20(3):273–297.

Courtney, K. L.; Stewart, S.; Popescu, M.; and Goodwin, L. K. 2008. Predictors of preterm birth in birth certificate data. In *MIE*, 555–560.

Crane, J., and Hutchens, D. 2008. Transvaginal sonographic measurement of cervical length to predict preterm birth in asymptomatic women at increased risk: a systematic review. *Ultrasound Obstet Gynecol* 31(5):579–87.

Creasy, R.; Gummer, B.; and Liggins, G. 1980. System for predicting spontaneous preterm birth. *Obstet Gynecol* 55(6):692–695.

Creasy, R. K. 1993. Preterm birth prevention: Where are we? *American journal of obstetrics and gynecology* 168(4):1223–1230.

Davey, M.; Watson, L.; Rayner, J.; and Rowlands, S. 2011. Risk scoring systems for predicting preterm birth with the aim of reducing associated adverse outcomes. *Cochrane Database Syst Rev* 11.

Edenfield, S.; Thomas, S.; Thompson, W.; and Marcotte, J. 1995. Validity of the creasy risk appraisal instrument for prediction of preterm labor. *Nursing research* 44(2).

Flood, K., and Malone, F. D. 2012. Prevention of preterm birth. *Seminars in Fetal and Neonatal Medicine* 17(1):58 – 63.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22.

Goldenberg, R.; Iams, J.; Mercer, B.; Meis, P.; Moawad, A.; Copper, R.; Das, A.; Thom, E.; Johnson, F.; McNellis, D.; Miodovnik, M.; Van Dorsten, J.; Caritis, S.; Thurnau, G.; and Bottoms, S. 1998. The preterm prediction study: the value of new vs standard risk factors in predicting early and all spontaneous preterm births. nichd mfmu network. *Am J Public Health* 88(2):233–8.

Goldenberg, R. L.; Culhane, J. F.; Iams, J. D.; and Romero, R. 2008. Epidemiology and causes of preterm birth. *The Lancet* 371(9606):75 – 84.

Goldenberg, R. L.; Hauth, J. C.; and Andrews, W. W. 2000. Intrauterine infection and preterm delivery. *New England Journal of Medicine* 342(20):1500–1507.

Goodwin, L.; Iannacchione, M.; Hammond, W.; Crockett, P.; Maher, S.; and Schlitz, K. 2001. Data mining methods find demographic predictors of preterm birth. *Nursing Research* 50(6):340–5.

Grote, N. K.; Bridge, J. A.; Gavin, A. R.; Melville, J. L.; Iyengar, S.; and Katon, W. J. 2010. A meta-analysis of depression during pregnancy and the risk of preterm birth, low birth weight, and intrauterine growth restriction. *Arch Gen Psychiatry* 67(10):1012–1024.

He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*, 1322–1328.

Hendler, I.; Goldenberg, R. L.; Mercer, B. M.; Iams, J. D.; Meis, P. J.; Moawad, A. H.; MacPherson, C. A.; Caritis, S. N.; Miodovnik, M.; Menard, K. M.; Thurnau, G. R.; and Sorokin, Y. 2005. The preterm prediction study: Association between maternal body mass index and spontaneous and indicated preterm birth. *American Journal of Obstetrics and Gynecology* 192(3):882 – 886.

Jaakkola, J. J. K.; Jaakkola, N.; and Zahlsen, K. 2001. Fetal growth and length of gestation in relation to prenatal exposure to environmental tobacco smoke assessed by hair nicotine concentration. *Environmental Health Perspectives* 109(6):pp. 557–561.

Kharrazi, M.; DeLorenze, G. N.; Kaufman, F. L.; Eskenazi, B.; Bernert, J. T., J.; Graham, S.; Pearl, M.; and Pirkle, J. 2004. Environmental tobacco smoke and pregnancy outcome. *Epidemiology* 15(6):660 – 70.

Klebanoff, M. A., and Keim, S. A. 2011. Epidemiology: The Changing Face of Preterm Birth. *Clinics in Perinatology* 38(3):339–50.

Maternal Fetal Medicine Units Network. 1994. Screening for Risk Factors for Spontaneous Preterm Birth – manual of operations.

Mercer, B.; Goldenberg, R.; Das, A.; Moawad, A.; Iams, J.; Meis, P.; Copper, R.; Johnson, F.; Thom, E.; McNellis, D.; Miodovnik, M.; Menard, M.; Caritis, S.; Thurnau, G.; Bottoms, S.; and Roberts, J. 1996. The preterm prediction study: A clinical risk assessment system. *American Journal of Obstetrics and Gynecology* 174(6):1885 – 1895.

NICHD. 2012. Scientific Vision: The Next Decade (13-7940). Washington, DC: U.S. Government Printing Office.

Papiernik-Berkhauer, E. 1969. [coefficient of premature delivery risk (c.p.d.r)]. *Presse Med* 77(21):793–4.

Porter, T. F.; Fraser, A. M.; Hunter, C. Y.; Ward, R. H.; and Varner, M. W. 1997. The risk of preterm birth across generations. *Obstetrics and gynecology* 90(1):63–7.

Smith, L. K.; Draper, E. S.; Manktelow, B. N.; Dorling, J. S.; and Field, D. J. 2007. Socioeconomic inequalities in very preterm birth rates. *Archives of Disease in Childhood - Fetal and Neonatal Edition* 92(1):F11–F14.

Thompson, J. M. D.; Irgens, L. M.; Rasmussen, S.; and Daltveit, A. K. 2006. Secular trends in socio-economic status and the implications for preterm birth. *Paediatric and Perinatal Epidemiology* 20(3):182–187.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)* 58:267–288.

Woolery, L. K., and Grzymala-Busse, J. 1994. Machine learning for an expert system to predict preterm birth risk. *Journal of the American Medical Informatics Association* 439–446.

Zerhouni, E. A. 2008. Prematuriy Research at the NIH.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67:301–320.