High Risk Pregnancy Prediction from Clinical Text

Rebecca Knowles[a], **Mark Dredze**[abc], **Kathleen Evans**[d], **Elyse Lasser**[c], **Tom Richards**[c]
Jonathan Weiner[c], **Hadi Kharrazi**[c]
[a]Center for Language and Speech Processing
[b]Human Language Technology Center of Excellence
[c]Center for Population Health Information Technology
[d]Johns Hopkins HealthCare LLC
Johns Hopkins University
Baltimore, MD, USA

Abstract

Patients with high risk pregnancies can benefit from case management and additional care. In order to provide patients with those services, they may be enrolled in case management programs. Machine learning and natural language processing (NLP) methods can be leveraged to automatically detect patients for referral to such programs. We describe initial experiments in predicting high risk pregnancies from clinical text data collected from electronic medical records (EMRs). We show that simple text features from unstructured records outperform baseline classification.

1 Introduction

Patients with high risk pregnancies often benefit from extra care. This includes support like assistance in identifying and managing conditions like gestational diabetes. Additionally, the complications from high risk pregnancies (such as premature births or infant admission to the NICU) can be both costly and pose serious dangers to the mother and child. Identifying such patients allows healthcare teams to direct them to sources of additional assistance, particularly free assistance of which patients may not be aware. The population considered in our study contains a particularly high proportion of patients whose pregnancies are considered high risk, and some pregnancies that are originally considered low risk may become high risk during the pregnancy.

In addition to the particular issues faced by patients in our population, our sample comes from a low-income (Medicaid) population in Maryland, which historically has had a higher infant mortality rate than the U.S. as a whole. In 2007, Maryland reported infant mortality rates of 8.0 deaths per 1,000 live births (as compared to a national average of 6.8). By 2011, the discrepancy had narrowed somewhat, with Maryland reporting a rate of 6.7 and the U.S. reporting a rate of 6.0.[1] In particular, Baltimore city continues to have a higher infant mortality rate than the state as a whole, and the black infant mortality rate remains higher than the white infant mortality rate (though the black infant mortality rate has seen a 24% decline between 2009 and 2012). Low birth weight and maternal complications are major contributors to the infant mortality rate.

All of these issues motivate providing additional care to patients with high risk pregnancies. However, enrolling patients in such programs can be a challenge. Of the high risk patients in the population, only about one third of them are enrolled in the high risk management program available to our population. Some of their risk factors may be easy to extract from structured records, allowing for automatic flagging of patient records and enrollment in programs. Patients may identify themselves as high risk, or their doctors may refer them to such a program. In other cases, though, the only evidence of high risk status is found in free form clinical notes. Manual discovery of high risk status from clinical text is costly; it may take considerable time to read through a patient's electronic medical record (EMR) and label them high risk.

The goal of this project is to reduce time spent reading patient records and more quickly and easily identify at-risk pregnancies by building a classifier that predicts whether a patient's pregnancy is high risk. By building a high-recall classifier, we can automatically classify a subset of the patients as high risk and flag them for further review by clinicians. Prior work [4] has shown that it is possible to classify patient smoking status from free form EMR text, as well as predict heart failure using machine learning techniques. [5] However, the specific task of high risk pregnancy identification is unexplored.

We present initial results of data exploration and simple machine learning techniques for high risk pregnancy classification from EMR text notes (see Table 1 for the types of notes available in the data).

2 Task

Our task is to predict high risk pregnancies based on the data contained within EMR records for a patient. Our focus in this paper is on the free text associated with each patient. This clinical text is comprised of notes written by doctors and hospital staff and can range from early pregnancy check-ups through labor and delivery to postpartum care, though not all patients may have data covering the entire pregnancy.

For the purposes of this project, we define a pregnancy to be "high risk" if the mother and/or fetus has an increased chance of morbidity, mortality, or both during the prenatal period. This includes patients with diagnoses of health issues such as hypertension, diabetes, or HIV (among others), substance abuse, or histories of preterm birth or low birth weight babies.

We formulate the problem as a supervised binary classification task. The label of "high risk" or "not high risk" is determined on a per-patient basis. In order to do this, each patient's notes are concatenated to form a single document per patient. Gold-standard labels were obtained by a domain expert reviewing ICD-9 codes, claims, and patient records.

Predicting high risk pregnancy has some additional challenges beyond those seen in related efforts, such as predicting smoking status. [4] While smoking status is indicated by a single cause, tobacco use, there are a number of factors and conditions that can cause a pregnancy to be considered high risk. These include history from previous pregnancies, maternal medical conditions, substance use, and much more. This makes it challenging for a model trained on a small dataset to determine that a patient is definitively low risk. For the time being, we focus on building high-precision classifiers which we can use to flag patients who are definitively high risk, leaving a smaller set of records to be manually examined for risk status. We leave for our long-term goals the creation of a high-recall system and the exploration of how our system can work with decision support algorithms already in place to find cases that are currently missed or to classify cases that are unclear to the decision support algorithms.

3 Data

The dataset is comprised of records for patients over a 4 month period. The full dataset consists of 15,028 records. Each record contains a freeform text note (from doctors, nurses, etc.). The records correspond to 202 patients, with patients having a mean of 74.4 records (median 58, minimum 4, maximum 475). All records are associated with a patient ID number, which is a unique identifier for the patient. The records are also timestamped. The full time span for each patient ranges between 0 days and 316 days, with a median and mean of 207.70 and 192.59 respectively (standard deviation 67.14). However, the records are not distributed evenly over that time span; they tend to be clustered around the birth, with a large number of records produced during labor, as well as during and immediately after childbirth.[1] This clustering is borne out in observing the pairwise time between records; the minimum pairwise time between records is 0 days, the maximum is 154, but the median

[1] This is to be expected; early pregnancy doctor's visits are more spread out; they become more frequent toward the end of pregnancy.

Field	Num.	Max	Mean	St. Dev.
Progress	2068	222	26.28	31.29
Flowsheet	10223	496	29.10	39.61
Additional	975	355	74.54	43.85
Attending	341	358	28.54	42.53
HP Assess.	295	191	57.00	27.63
LD	644	248	63.86	36.79
LD Attending	235	255	11.00	30.22
Delivery	247	312	87.36	46.18

Table 1: Number of tokens per note type.

and mean are 0.03 and 2.62 respectively (standard deviation 8.4). Shown in Table 1 are the token count values for the various types of notes.

In addition to the unstructured text, patient records contain structured information. This includes information about family history, current medications used by the patient, past medical history, and past pregnancies. At the moment, the only structured data that we have access to is the estimated gestation weeks field.

Approximately 11% of the records contain estimated weeks of gestation, a numerical attribute. The minimum value is 5 weeks, and the maximum value is 40. The distribution of estimated gestation weeks provides additional evidence of the clustering of records around birth; the mean and median are 30.50 and 28.18 weeks, respectively, while the standard deviation is 9.0.

4 Features

As the data consists of unstructured text, we begin by extracting simple unigram features. Each token consists of one or more alphanumeric characters (punctuation is ignored). We experimented with several values of n, but did not find consistent improvements across models using $n > 1$, so we present results on unigram features only. Larger contexts will likely prove useful as more data becomes available. Experiments also did not show improvement by removing stopwords or removing extremely common or extremely uncommon tokens from the feature set.

We experiment with three versions of the features: count (the count of occurrences of a type within a document), TF-IDF (term frequency-inverse document frequency weighted features), and binary (all non-zero counts set to 1).

As the dataset grows, we plan to expand our experiments to include structured data on patient medical history, family history, and medications. This includes both categorical and numerical data, rather than the unstructured text of the notes. Additionally, to extract further information from the text we plan to utilize the Apache clinical Text Analysis and Knowledge Extraction system (cTAKES), [3] an open-source NLP system for EMR text. This will allow us to consider named entity recognition (NER) for extracting information on diseases, medications, etc. It also contains a morphological normalization component for normalizing lexical variations. We have also noted typos in the data, and may wish to use normalization to address those as a data preprocessing step.

5 Models

For learning we use standard machine learning classifiers: naive Bayes and support vector machines (SVMs). Scikit-Learn [2] provides several implementations of both models. Multinomial naive Bayes is commonly used for text classification, with either count or TF-IDF features. The parameter α is used for smoothing. The Bernoulli naive Bayes implementation uses only binary-valued features. We also present results with two SVMs, one with a linear kernel and the other with a radial basis function kernel. The baseline is a simple majority classifier (which labels all patients as high risk).

Table 2 shows results for each classifier across the three feature sets (binary, count, and TF-IDF).

	Binary	Count	TF-IDF
Majority	56.8	-	-
Bernoulli NB	59.8	-	-
Mult. NB, $\alpha = 0.01$	62.0	58.9	62.0
Linear SVM	57.2	53.9	54.2
RBF SVM	56.8	61.4	56.8

Table 2: Results for 25-fold cross validation using unigrams.

6 Results

Due to the small data size, we present the results of 25-fold cross validation on our labeled dataset. Table 2 presents results using unigram features. The baseline classification accuracy (majority classifier) performs at 56.8%. The best model performs at 62.0% correct, while some versions of the SVM classifiers perform below the majority baseline.

7 Future work

Much of our future work depends on the ongoing collection of additional data. We will combine the structured and unstructured data to improve prediction, as well as exploring how the unstructured data can provide predictive power beyond that of the structured data.

After improving our high risk detection models through additional feature engineering and data collection, we plan to focus on early detection. In particular, how early can we detect high risk pregnancies? Early risk detection should provide more time for providers to enroll patients in case management programs, with the goal of improving patient outcomes. However, early detection faces the challenge of limited data; patients may be classified on the basis of just a few notes.

Since there are many potential reasons for a pregnancy to be classified as high risk, we may choose to explore clustering of patients by their risk factors. Certain types of risk may be easier to predict from the unstructured data, so we may wish to build separate models for subclasses of high risk pregnancies. Additionally, this would serve to give clinicians information about why a patient is labeled as high risk. Finally, we plan to explore the tradeoff between high accuracy and high recall.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1232825.

References

[1] Maryland Department of Health and Mental Hygiene. 2013. *Maryland vital statistics: Infant mortality in maryland, 2012.* Retrieved October 6, 2014, from http://www.healthybabiesbaltimore.com/uploads/file/State Infant Mortality Report 2012 (2).pdf

[2] F. Pedregosa et al. 2011. *Scikit-learn: Machine Learning in Python*. JMLR 12, pp. 2825-2830.

[3] Guergana Savova, James Masanz, Philip Ogren, Jiaping Zheng, Sunghwan Sohn, Karin Kipper-Schuler, and Christopher Chute. 2010. *Mayo Clinic Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*. JAMIA.

[4] Sunghwan Sohn and Guergana Savova. 2009. *Mayo Clinic smoking status classification system*. Proc. AMIA.

[5] Jionglin Wu, Jason Roy, and Walter F. Stewart. 2010. *Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches.* Medical Care, vol 48, issue 6.