Predictors of Preterm Birth in Birth Certificate Data

K. Courtney\textsuperscript{(1)}, S. Stewart\textsuperscript{(2)}, \textbf{M. Popescu}\textsuperscript{(3)}, L. Goodwin\textsuperscript{(4)}

\textsuperscript{(1)} University of Pittsburgh
\textsuperscript{(2)} Epic Systems
\textsuperscript{(3)} University of Missouri
\textsuperscript{(4)} Duke University
Preterm Birth (PB) Problem

- **Definition**: birth prior 37 weeks of gestation
- 24% of neonatal deaths are preterm
- Survivors have increased risk for developmental disabilities.
- US PB Rates have increased in recent years
  - 12.7% (2003) at a cost of ~$15 billion
  - 11.4% of white mothers
  - 18.3% of African American mothers
Study aim

- Create a model to predict premature birth using existing birth certificate data
- More effective preterm birth (PB) intervention
- Compare results of various computational modeling approaches
- Compare results of this state wide study to the results of an academic medical center study.
  - Is the demographic model generalizable?
  - Can the model be used in any small hospital that does not have EMR?
Data sources

- North Carolina birth certificate records for 2003 and 2004 (present study)
  - Contained approximately 120,000 live births
  - Publicly available and de-identified
  - Racially and ethnically diverse population
  - North Carolina preterm birthrate (13.3%) is consistently higher than national average (12.4%)
  - Maternal, paternal, infant and health care system variables

- Duke Medical Center medical records data (previous study)
  - ~20,000 live births, from 1986-1995
Research design

- Retrospective study
- Tried to use the same variables as in the previous study
  - Match the birth certificate data with medical records
- 2 variables could not be matched
  - Mother’s religion
  - Payor source
Data set preprocessing

- Filtered out
  - Out of state births
  - Induced or stimulated labor
  - Multiple births
- Approximately 73,000 birth records left
- 43 variables
  - Demographic
  - Medical History
  - Prenatal care
- Except age, all variables were categorical (match the previous study)
Data analysis

- Logistic regression:
  - Output: 0- no preterm, 1-preterm
- Support Vector Machines (SVM)
- Neural Networks (NN)
  - Nx2Nx2, N=43
  - 2 class output: preterm, not preterm
- Bayesian Classifiers
  - Used only the mean of each class (term, preterm) nearest prototype
Results evaluation

- Used are under the ROC curve (AUC)
- ROC = true positive rate (sensitivity) vs. false positive rate (1-specificity)

true_positive_rate = \frac{\# \_ predicted \_ preterm}{\# \_ total \_ preterm}

false_positive_rate = \frac{\# \_ falsely \_ predicted \_ preterm}{\# \_ total \_ term}

- AUC values:
  - 0.5 – results can be explained by chance
  - 1 - perfect prediction
## Results (AUC values)

<table>
<thead>
<tr>
<th>Method</th>
<th>Original Study (Goodwin, et al., 2001) (n = 19,970)</th>
<th>Current Study (n = 73,040)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.66</td>
<td>0.605</td>
</tr>
<tr>
<td>Neural networks (NN)</td>
<td>0.64</td>
<td>0.57</td>
</tr>
<tr>
<td>CART-based Custom Classifier Software (Goodwin, et al., 2001)</td>
<td>0.72</td>
<td>0.56</td>
</tr>
<tr>
<td>Rule induction</td>
<td>0.67</td>
<td>Not used</td>
</tr>
<tr>
<td>Bayesian classifiers</td>
<td>Not used</td>
<td>0.59</td>
</tr>
<tr>
<td>Support Vector Machines (SVM)</td>
<td>Not used</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Discussion

- The demographic variables from the birth certificate data may be used to predict preterm births.
- The birth certificate (state wide) models were not as accurate as the medical records (local) based one.

Reasons:
- Two of the variables (religion and source of payment) were not available.
- The hospital sample was biased (affluent, mainly white population).
- Methodological (modeling): proprietary CART software vs. SAS implementation.
- Possibly: local (medical records) models are all the time more accurate than the population based models.
The birth certificate data contains variables not available in the hospital records.

Some of those might be good PB predictors:
- Kotelchuck index – APNCU (utilization of prenatal care)
- Kessner index (adequacy of prenatal care)
  - Low, medium, High
Conclusions

- Preterm births predictors can be used in clinical decision support systems that may reduce the preterm pregnancies.

- The relatively consistent results across models imply that more factors need to be identified to increase the accuracy of the models.
Thank you!

Questions?