Predictors of Preterm Birth in Birth Certificate Data

Karen L. COURTNEY a1, Sara STEWART b, Mihail POPESCU c, and Linda K. GOODWIN d

a University of Pittsburgh, USA, b Independent Researcher, USA, c University of Missouri – Columbia, USA, d Duke University, USA

Abstract. Demographic factors have been shown to be moderate predictors of preterm birth in prior studies which used hospital databases and epidemiologic sample surveys. This retrospective study used de-identified 2003 North Carolina birth certificate data (n=73,040) and replicated the statistical and computational methods used in a prior study of an academic medical center’s data warehouse. Receiver Operating Characteristics (ROC) curves were used to compare results across methods. Due to differences between the data collected for birth certificates and the original clinical database, five of the seven demographic variables in the clinical database model were available for model testing (maternal age, marital status, race/ethnicity, education and county). Even with a reduced model, multiple methods of statistical and computational modeling supported the earlier findings of demographic predictors for preterm birth. The reduced model AUC results were acceptable (logistic regression = 0.605, neural networks = 0.57, SVM = 0.57, Bayesian classifiers = 0.59, and CART = 0.56), but lower than in the prior study as might be expected for a reduced model. On a population level, these results support a prior demographic predictor preterm birth model generated from a clinical database and the use of computational methods for model formation. Additional testing for stronger predictor models within birth certificate data is suggested as birth certificate data is a parsimonious population dataset already routinely collected.

Key Words. Premature Birth, Health Services Research, Modeling, Obstetrics & Gynecology, Nursing

Introduction

Worldwide, preterm birth and/or low birth weight contribute to 24% of neonatal deaths [1]. In the United States, preterm births (birth prior to 37 weeks gestation) are the leading cause of neonatal deaths not related to congenital birth defects [2]. Rates of preterm birth have been increasing in the United States and in 2003, the national average was 12.7% of all births within the United States [3], but rates among some racial and ethnic populations such as African Americans (18.3%) in comparison to among non-Hispanic white mothers (11.4%) [3].

1 Corresponding Author: Karen Courtney, School of Nursing, University of Pittsburgh, 415 Victoria Building, Pittsburgh, PA 15261, USA; Email: Karen.courtney@alumni.duke.edu
Preterm infants who survive are at increased risk for developmental disabilities, cerebral palsy, mental retardation, blindness, chronic lung problems, and other lifelong conditions which add to the pain and suffering families may experience with the birth of a preterm baby. The annual financial toll for initial hospital care of preterm infants is estimated to be US$15.5 billion [2] each year, with additional costs for children with special needs going well beyond this already astronomical amount. Small increases in gestational age are also associated with lower mean hospital charges. In one study, the mean hospital charges for full term infants was US$4,788 compared to US$10,561 (33 – 36 weeks), US$55,792 (29-32 weeks), and US$239,749 (26 – 28 weeks) for preterm infants [4]. Clearly, strategies are needed to find solutions to this important preterm birth problem. Complicating the search for effective preterm birth prevention interventions is a lack of an adequate and reliable preterm birth prediction model.

Exact etiology of spontaneous preterm birth is not known at this time and is likely to be the interaction of multiple risk factors [2, 3]. The literature is replete with contradictory research findings that yield minimal guidance for perinatal practitioners, and there is no scientifically validated framework for preterm risk factors. As an example, despite the prior widespread use of a preterm risk scoring tool, this tool was later found to be ineffective in accurately identifying most preterm births [5].

More recent modeling for preterm birth prediction has used medical records to investigate the role of personal health behaviors, history and demographics in preterm birth. However this approach has also resulted in inconsistent results. Goodwin, et al. [6] used multiple data mining techniques to study 19,970 racially diverse pregnant women and found seven demographic variables produced 0.72 area under the curve (AUC). It is interesting to note that in contrast to their results using seven variables, using all 1,622 variables available in the dataset produced only a slight increase in AUC (0.04) when predicting birth outcomes. In using receiver operating characteristic (ROC) techniques, chance prediction would be 0.50 AUC and perfect prediction would be 1.0, so 0.72 is a respectable prediction that suggests demographic data are important but, obviously, improved predictive accuracy is needed for clinical decision support.

Other studies have supported several of the demographic variables noted in the Goodwin model including: maternal age; marital status race and/or ethnicity [7]. Other research has also found personal income and neighborhood income characteristics to be important demographic predictors of preterm birth [8].

Sampling in these prior studies potentially could have affected the resulting preterm prediction models. Researchers have noted considerable variation in socio-demographic variables as well as preterm birth outcomes in comparison of a study sample drawn from local clinics and the local population characteristics [9]. As the Goodwin et al. [6] model was based on the medical records of academic medical center patients during years 1986 to 1995, this sample may have also affected the resulting prediction model. Patients receiving care in an academic medical center may represent higher acuity patients and not be representative of the population as a whole. Further testing of this prior model with either a broader sample or with population records was needed.
1. Methods

1.1. Research Design

This study was a retrospective, secondary analysis of North Carolina birth record data. The purpose of this study was to examine if a demographic preterm prediction model previously generated by data from academic medical center’s clinical database could be applied to a population database. Following approval by the Health Sciences Institutional Review Board, we attempted to replicate a prior demographic preterm birth prediction model [6] with birth certificate data. Variables within the birth certificate data were as closely matched with the variable definitions from the prior model as possible. Two variables, mother’s religion and payor source were not available within the birth certificate file and therefore omitted from the replication model. Preterm birth for this study was a dichotomous variable defined as a birth prior to 37 weeks gestation.

1.2. Sample and Data Source

The data source for this project was the publicly-available, de-identified, birth certificate records for North Carolina for 2003 [10]. These data were the most recent data available at the beginning of the study. The North Carolina birth records were chosen as the prior study was based on medical records from an academic medical center located in North Carolina, but this study’s dataset (2003) did not contain any records from the previous study’s birth dataset (1986-1995) [6]. This study’s data set contained 120,168 live births and included maternal, paternal, infant and health care system variables. As state law requires birth certificate completion for all births, this data set is estimated to represent 99% of all births [3]. The exact number of undocumented births is unknown though likely small in number. When filtered for out-of-state births, induced or stimulated labor or multiple births, the data set contained 73,040 birth records.

1.3. Analysis

This study replicated the statistical and computational methods used in the prior study: logistic regression, neural networks, classification and regression trees (CART) and also included Support Vector Machines (SVM) and Bayesian classifiers for model comparison. This project used the Matlab® Neuronal Networks package and the SVM classifier implemented in the Matlab® Statistical Package. The Bayes classifier was coded in Matlab® by one of the authors (MP). Receiver Operating Characteristics (ROC) curves were used to compare results across methods.

Logistic regression was chosen based on the primarily categorical data being used. The response variable was binary (preterm or not preterm). With the exception of age, all the data were categorical. From the birth certificate data, the variables were categorized in the same manner as the original study. For example, in the birth certificate data, education is listed as the number of years of education. We put these education data into categories to match the variables used by the previous study. Our raw data also had ethnicity and race as distinct categories, which was not done in Goodwin et al.’s study [6]. Therefore, we had to combine these two variables from the birth certificate dataset in order to match original study definitions.
The NN structure consisted in N input binary inputs, 2N hidden layers and 2 output nodes, one for each class (preterm or term). The number of inputs, N, varied according to the number of variables used and the discretization of the continuous variable such as maternal age. For the case when N was large (for example, the county variable) we used principal component analysis (PCA) to reduce the dimensionality of the input space. The training of the NN was performed on 2,000 records from the 2004 North Carolina birth certificate data.

The Bayes classifier was implemented in a simple fashion by using the class mean. An unknown pattern was assigned to the class with the closest (using Euclidean distance) mean. More sophisticated Bayesian approaches did not seem to improve the classification result.

ROC curves are used to give a graphical representation of results of prediction. It is a graph of sensitivity versus specificity (true positives vs. true negatives) given a binary classifier. The area under the curve (AUC) represents predictive accuracy of the model. At 0.5, it means the results can be explained by chance. The ROC curve was useful for comparison in this study, as we had a binary classifier (preterm vs. not preterm) and all the methods gave us an AUC value. Even with the AUC values, it is helpful to compare the visual representation of the curves, which is done in Figure 1.

2. Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Original Study (Goodwin, et al., 2001) AUC Results (n = 19,970)</th>
<th>Current Study AUC Results (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</td>
<td>0.72</td>
<td>0.56</td>
</tr>
<tr>
<td>(Goodwin, et al.; 2001)/Classification &amp;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression Trees (CART)</td>
<td></td>
<td></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>

Due to differences between the data collected for birth certificates and the original clinical database, five of the seven demographic variables in the clinical database model were available for model testing (maternal age, marital status, race/ethnicity, education and county). Even with a reduced model, multiple methods of statistical and computational modeling supported the earlier findings of demographic predictors for preterm birth. The reduced model AUC results were acceptable (logistic regression = 0.605, neural networks = 0.57, SVM = 0.57, Bayesian classifier = 0.59, and CART = 0.56), but lower than in the original study.
Figure 1 compares the ROC curve results among the different statistical and computational analyses. As illustrated, the results of this model are consistent across multiple statistical and computational models.

3. Discussion

The results of this study yielded lower AUC results than the original work; however this was not unexpected as two of the original variables in the model were unavailable in the new dataset. Although the direction of these results do support the original demographic model proposed by Goodwin et al. [6], these results also suggest that further refinement of this socio-demographic model is needed. The birth certificate data set contains other variables not available to the original research team in the clinical database. These variables include paternal demographic factors, maternal health behaviors and maternal medical history which may be important and should be incorporated into future socio-demographic model research. Of particular note, the
birth certificate data set contains measures of prenatal care adequacy: the Kotelchuck Index and the Kessner Index which were not available in the prior clinical dataset. Previous research has suggested that adequacy of prenatal care can affect preterm birth outcomes [11]. These measures should be tested in further refinement of a socio-demographic preterm prediction model.

4. Conclusion

The significance of demographic predictors in a pregnant population center around opportunities for intervention where socio-economic and geography factors can be mediated to prevent preterm birth and other pregnancy complications. Refinement of a socio-demographic prediction model using existing population data sources is the next step. Future models can be used in the design of clinical decision support systems. Additional research will need to examine how well subsequent socio-demographic models predict birth outcomes geographically in order to identify communities most at risk and target interventions effectively.

Acknowledgements:

This work was supported in part by the National Library of Medicine Biomedical and Health Informatics Research Training Grant T15-LM07089-14.

References: