Final Report to the National Library of Medicine

RO1 LM-O6488

Informatics Tools and Perinatal Knowledge Building

1997-2000 (and no cost extension through May 2001)

Linda Goodwin, RN, C, PhD
Principal Investigator

Wm Edward Hammond, PhD
Co-Investigator

Research Team Members

Patrick Crockett, PhD – Analytika, Inc
Shayne Gad, PhD – Gad Consulting Services
Jerzy Grzymala-Busse, PhD – University of Kansas, Computer Science
Marvin Hage, MD – Duke University, Obstetrics and Gynecology
Mary Ann Iannacchione, RN – Duke University, Project Director
Sean Maher, MS – Discovery Corps
Gerard Nahum, MD - Duke University, Obstetrics and Gynecology
Lucila Ohno-Machado, MD, PhD – Harvard Univ, Decision Sciences Group
Jonathan Prather, PhD – Duke Univ, Medical Informatics Doctoral Student
Kaye Schlitz, RN, C – Duke University, Information Services
Mary Ellen Sievert, PhD – University of Missouri, Info and Library Sciences
The specific aims of RO1 LM-O6488 research used Duke’s TMR electronic patient record data, informatics tools, and expert practitioners to develop decision support tools for assessment of preterm birth risk.

1. **(MET)** Using an extensive perinatal clinical data repository, explore data filtering techniques that optimize data accuracy.

2. *(See sub-aims)* Transform perinatal data into information using statistical and machine learning software programs.
   - 2a. **(MET)** Generate statistical models of preterm birth risk.
   - 2b. **(MET)** Generate a knowledge base using inductive machine learning techniques that generate production rules.
   - 2c. **(MET)** Generate a second knowledge base using connectionist machine learning techniques that generate neural networks.

3. **(PARTIALLY MET)** Verify information for preterm birth risk knowledge bases using certified expert clinicians and content validity procedures.

4. *(See sub-aims)* Validate the decision support provided by the preterm birth risk knowledge bases.
   - 4a. **(MET)** Compare the predictive accuracy of inductive versus connectionist preterm birth risk knowledge bases with statistical analyses.
   - 4b. **(MET)** Analyze verified knowledge base terms for mapping into Unified Medical Language System metathesaurus vocabularies where automatic links to library and information resources could be built.

Our findings are in agreement with Lovell\(^1\) that data mining results depend on the quality and content of the data. Our RO1 team spent nearly 16 months extracting, cleaning, and filtering TMR™ prenatal clinical data to yield the highest possible quality for conducting research. Given the fact that the data were collected as clinical data (not for research purposes), earlier work that found TMR™ contained 8.74% unusable data suggests a relatively high data quality for a clinical database\(^2\). Our research team consisted of both clinical and technical experts who spent many hours reviewing data sets, the data dictionary, and outlier reports. Our process used both computerized and manual techniques for optimizing data accuracy, which remains an imprecise skill. However, our research team’s attention to detail, rigor in data quality checking, and commitment to excellence leaves us confident that the research data is of high quality for the purposes of mining for predictors of preterm birth.


A growing issue in data mining research is both the ethical and the regulatory need to protect patient privacy in mining large clinical databases. Procedures used in our RO1 data mining research were recently compared with HIPAA Privacy regulations issued December 28, 2000 and found that compliance with new regulations would need only minor tweaking for county codes; this work is currently under review for a paper presentation at the fall 2001 AMIA symposium. We believe our work in this area is leading edge and has important implications as a benchmark for other data mining research and privacy protection issues in health care.

Encouraging results with LERS inductive rule generation, from our previous study, were not replicated in the current RO1. LERS is combinatorially explosive using exponential calculations to achieve rough set analysis. It has not yet been possible to adapt the software to analyze 1,622 variables and we cannot find reasonable processing power with more than about 50 variables. Thus analysis with LERS experimented with pre-selected variables that yielded disappointing results. We also were unable to persuade our clinical experts to complete content validity indexing (CVI) procedures for the multiple exploratory models in our iterative data analyses procedures. CVI is tedious and time consuming, and our experts suggested this will be more appropriate at a hypothesis and model-confirming stage of the research rather than the exploratory phase in which we are working now.

Iterative analyses focused on optimizing sensitivity metrics and comparing results across multiple analysis (both data mining and statistical) methods using Receiver Operating Characteristic (ROC) curves that provide graphical as well as statistical comparisons. Results found small differences in performance between four different modeling techniques that used neural networks, logistic regression, CART, and new experimental software that combined case-based and CART algorithms. We find similar ROC curves for preterm birth prediction for all models/methods tested.

Given the volume of data available for study, copious experiments were run using a variety of statistical analysis procedures. We can produce massive volumes of statistical output and graphic displays, however we believe that doing so would be misleading from a clinical standpoint. Interpreting an individual P value is easy, however our analyses generate thousands of P values which are often difficult to interpret, especially when considering that thousands of hypotheses can also be generated for testing in this large data set. Given the large sample size, statistical significance is often achieved but expert interpretation finds the significance has no clinical merit or validation. Rather than hypothesis testing, this phase of our research has been exploratory and hypothesis generating in its aims to develop knowledge-based models of preterm birth predictors using a variety of statistical and machine learning methods. Statistical significance does not adequately address whether the results in a given study will replicate. Thus, in our preterm domain and exploratory phase, replicability of results is of greater interest than statistical significance. Procedures for splitting data into training versus testing sets and ten-fold cross-validation were used to analyze replicability of results and future studies will add new data sets for further analysis and tests of replicability.

3 Health Insurance Portability and Accountability Act of 1996 (HIPAA) http://aspe.os.dhhs.gov/admnsimp/
We believe that the most important results of our current exploratory work find that of 1,622 variables extracted for this study, 7 demographic variables (binary coding resulted in 30 variables for analysis) yielded the most promising results (.72 area under the curve using Receiver Operating Characteristic analyses). This finding supports an important concept in data mining that indicates machine learning techniques can be used for dimensionality reduction. In a broad sense, data mining research seeks to reduce redundancy in data and thereby reduce its complexity and dimensionality; the larger the data set, the more complex is its dimensionality.

Hardware and software limitations have historically forced data analysts to "pre-select" those variables they believed were best/most relevant. Data mining researchers tended to use experts for dimensionality reduction through a process called “feature selection” to reduce the number of attributes (variables) for analysis. Feature selection is frequently a daunting task when deciding which attributes are important, resulting in finding many "good" features that result in a combinatorial problem space. Using expert heuristics, Creasy developed a manual preterm risk-scoring tool that was widely used in clinical practice for nearly a decade, but later evaluated as ineffective for accurate identification of most preterm births5. Dalal6 described that human pre-selection bias for Challenger space shuttle data ultimately ended in disaster, and that when analyses with all available data were conducted, the O-ring failure was accurately predicted. Thus, based on Creasy's manual tool failure and human pre-selection problems reported by Dalal, our RO1 work tried to avoid feature (variable) selection by humans.

Our work was, in fact, highly successful in dimensionality reduction where 1,622 variables were reduced to seven demographic variables with a .72 area under the curve (ROC) ability to predict birth outcomes (see Figure 1). However using a data volume of 1,622 variables resulted in noise-to-signal problems that we believe masks the effect of other potentially informative variables. Continued research will apply new algorithms for dimensionality reduction within the data to search for other informative variables that improve the sensitivity and positive predictive value of the models.

The seven demographic variables that produced .72 area under the curve (ROC) results include maternal age, race, education, marital status, payor category, county of residence, and religion; collection of these demographic variables is non-invasive and should be collected on all patients anyway (low cost). Analyses were conducted using maternal age as a continuous variable and binary coding resulted in 30 variables for analysis for the other demographic variables (e.g. black race = yes or no; white race = yes or no; high school education = yes or no, Durham county = yes or no, etc.)

Figure 1. Seven TMR Demographic variables produced .72 area under the curve

Our results are in direct contradiction with work reported by Wildschut, however our TMR data was more racially diverse (see Table 1) than that analyzed by Wildschut. Our findings include maternal age and (black) race amongst those reported by Norwitz et al., who provided an excellent August 1999 review article and reported demographic variables of very young or older maternal age, black race, low weight before pregnancy, and low socioeconomic status as predictive of preterm birth outcomes; they also reported non-demographic variables (biochemical markers and biophysical conditions) as predictive of preterm birth, and their review will help guide our future research.

Table 1. Minority Representation

<table>
<thead>
<tr>
<th>American Indian or Alaskan Native</th>
<th>Asian or Pacific Islander</th>
<th>Black, not of Hispanic Origin</th>
<th>Hispanic</th>
<th>White, not of Hispanic Origin</th>
<th>Other or Unknown</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>105 (1%)</td>
<td>116 (1%)</td>
<td>10901 (55%)</td>
<td>519 (3%)</td>
<td>7837 (39%)</td>
<td>492 (2%)</td>
<td>19,970</td>
</tr>
</tbody>
</table>

Contrary to clinical wisdom and earlier studies, most of the predictive power in Duke’s TMR database used for this study was found in 7 demographic variables in a racially

---

A diverse sample. Where seven demographic variables produced results of 0.72 area under the curve (ROC), the addition of approximately a thousand additional variables added only .03 area under the curve (ROC). We believe there are other informative variables in those thousand additional variables, but filtering informative variables from the noisy volume of variables is an unsolved problem that requires further research.

An experimental rule induction software program was used to identify sub-populations in this database that had more than twice the average risk of premature birth. The rules often included lengthy (attribute, FALSE) conditions. The (attribute, TRUE) conditions confirm many of the heuristic rules related to medical conditions of preterm risk used by clinicians and supported in the literature. Our TMR and clinical experts indicated that pervasive "unknown" data values usually occur in a TMR record when a patient arrives in the emergency room having had no prenatal care. Validation checks were run for individual patients and confirmed that, in most cases, this was true. The American Indian rule was a surprise that merits further study with larger samples. Hypothesis testing for the American Indians in the study was unable to find an explanation for their high rate of preterm births, and remains an area for further research.

### Table 2. Sub-populations at Higher Risk of Preterm Birth

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>% Premature</th>
<th>Agrees with findings by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple gestation</td>
<td>True</td>
<td>0.78</td>
<td>Ventura(^9) and others</td>
</tr>
<tr>
<td>Incompetent cervical OS</td>
<td>True</td>
<td>0.57</td>
<td>Norwitz (see footnote 8) and others</td>
</tr>
<tr>
<td>Uterine fibroids</td>
<td>True</td>
<td>0.56</td>
<td>Norwitz (see footnote 8g) and others</td>
</tr>
<tr>
<td>Previous Stillborn</td>
<td>True</td>
<td>0.53</td>
<td>Woolery(^10)</td>
</tr>
<tr>
<td>Education; unknown AND Marital status; unknown</td>
<td>True</td>
<td>0.52</td>
<td>Curry(^11)</td>
</tr>
<tr>
<td>Chronic hypertension</td>
<td>True</td>
<td>0.50</td>
<td>Samadi &amp; Mayberry(^12) and others</td>
</tr>
<tr>
<td>Cerclage</td>
<td>True</td>
<td>0.48</td>
<td>Iams(^13) and others</td>
</tr>
<tr>
<td>Urine screen; cocaine</td>
<td>True</td>
<td>0.47</td>
<td>Martinez, Larribee, &amp; Monga(^14) and others</td>
</tr>
<tr>
<td>American Indian</td>
<td>True</td>
<td>0.42</td>
<td>None found</td>
</tr>
<tr>
<td>Poor weight gain</td>
<td>True</td>
<td>0.42</td>
<td>Schieve(^15) and others</td>
</tr>
</tbody>
</table>

---

Long-term goal of the research:

Ultimately, the goal of this research should lead to clinical decision support system(s) for accurate identification and appropriate intervention in pregnant women who are at risk for preterm birth outcomes.

Summary of Progress, Current Status, and Next Steps

<table>
<thead>
<tr>
<th>Study</th>
<th>Data</th>
<th>N</th>
<th>Vars</th>
<th>Machine Learning</th>
<th>Output</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1R43 NR-02899</td>
<td>SL&amp;HD</td>
<td>18,890</td>
<td>214</td>
<td>LERS</td>
<td>520 production rules</td>
<td>53-89% accuracy in 3 different data sets</td>
</tr>
<tr>
<td>R01 LM-06488</td>
<td>TMR</td>
<td>19,970</td>
<td>1,622</td>
<td>Logistic regression, Neural Networks, CART, experimental software</td>
<td>Multiple predictive models</td>
<td>.72 Area under the curve (ROC) for 7 demographic variables</td>
</tr>
<tr>
<td>Next study</td>
<td>OB TraceVue, PDA database and forms</td>
<td>To be determined</td>
<td>30 prospective; retrospective to be determined</td>
<td>To be determined</td>
<td>To be determined</td>
<td></td>
</tr>
</tbody>
</table>

To date, review articles report that manual preterm risk scoring tools have yielded only 17-38%\(^{16}\) and 20-30%\(^{17}\) positive predictive values. Both the prior SBIR work and the RO1 work surpass the predictive rates that manual preterm risk scoring offers.

Holzman, Paneth, and Fisher\(^ {18} \) report that preterm factors could be categorized as antecedents, biophysical markers, and mediators of preterm birth. Our work is focused on assessment, prevention, and early detection, thus our studies analyze antecedents and biophysical markers. Table 3 shows results from review and research articles that will help guide our continued research.

Next Steps:
1. Add variables reported in the literature that are not included in previous data sets
2. Continue to conduct data cleaning and filtering procedures that meet or exceed compliance requirements and HIPAA privacy regulations
3. Continue exploratory data analysis with new machine learning programs and data
   a. Search for other informative variables in existing noisy data sets
   b. Apply machine learning programs to new data sets
4. Using steps 1-3 above, improve predictive accuracy and sensitivity of preterm birth prediction models above .72 area under the curve (ROC)
5. Continued systems lifecycle work for knowledge-based decision support

---


Table 3: Preterm birth predictive variables

<table>
<thead>
<tr>
<th>Antecedents</th>
<th>Biophysical Markers</th>
<th>Mediators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norwitz, Robinson, &amp; Challis (see footnote #4)</td>
<td>Previous preterm delivery; Multiple gestation; <strong>Bacterial vaginosis</strong>; Uterine anomalies; Hydramnios; Infection; Smoking; Demographic (maternal age, black race, low prepreg. weight, low SES)</td>
<td>Cervical length; <strong>Fetal fibronectin</strong>; Relaxin; Serum and salivary estriol; Corticotropin-releasing hormone</td>
</tr>
<tr>
<td>NICHD Maternal Fetal Medicine Unit¹⁹</td>
<td><strong>Stress</strong>; Upper genital tract infections; <strong>Bacterial vaginosis</strong>; Previous spontaneous preterm birth; Hypertension; Working; History of lung disease; Covariates: race, gestational age</td>
<td><strong>Fetal fibronectin</strong></td>
</tr>
<tr>
<td>(Woolery) Goodwin et al (prior work) see footnote 10</td>
<td>Weeks at referral; Previous Intrauterine fetal death; Hydramnios; 2nd trimester bleeding; Previous preterm labor; History of spontaneous abortions; Premature ruptured membranes</td>
<td>Cervical changes</td>
</tr>
<tr>
<td>Goodwin et al (RO1 LM-06488)</td>
<td>7 demographic variables (See Figure 1) 10 subgroup rules (See Table 3)</td>
<td>Numerous markers including cervical measures, vital signs, and lab results</td>
</tr>
</tbody>
</table>

Please note: variables not in bold were included in the TMR data with current work and may be suppressed as a result of noisy data. The next proposed studies will apply new machine learning programs to the same data to explore other potentially informative variables in existing data. In addition, we will prospectively collect items in bold that are not included in existing TMR data.

¹⁹ http://www.bsc.gwu.edu/mfmu/
RO1 LM-O6488 work resulted in twenty-one data-based and refereed publications (four manuscripts under review) including three book chapters and two posters with published abstracts:


