Using Big Data to Predict Infant Mortality in Allegheny County, Pennsylvania

Fuchiang (Rich) Tsui, PhD.
Associate Professor of Biomedical Informatics, Intelligent Systems Prog., Bioengineering
Department of Biomedical Informatics
University of Pittsburgh

56th NAWRS Workshop

July 31, 2017
Disclosures

• Inventor
  System for Hospital Adaptive Readmission Prediction and management (SHARP)
• Funder & Consultant: General Biodefense, LLC
• Funding support: RK Mellon Foundation, NIH, Coulter foundation, Innovation Works, Local foundation in Pittsburgh, Children’s Hospital of Pittsburgh of UPMC
• Fan of learning health system, clinical/public health informatics, machine learning, big data science
Outline

• Background
  – Problem
  – Project Vision
  – Data collaborators

• Project approach and Current progress
  – Predictive modeling concept
  – Data collection and integration
  – Pilot study: MOMI data analysis and results

• Discussion and Conclusion
U.S. lags behind other wealthy nations on infant mortality

Infant mortality rates in selected OECD countries, 2010

<table>
<thead>
<tr>
<th>Country</th>
<th>Mortality Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finland</td>
<td>2.3</td>
</tr>
<tr>
<td>Japan</td>
<td>2.3</td>
</tr>
<tr>
<td>Portugal</td>
<td>2.5</td>
</tr>
<tr>
<td>Sweden</td>
<td>2.5</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>2.7</td>
</tr>
<tr>
<td>Norway</td>
<td>2.8</td>
</tr>
<tr>
<td>Korea</td>
<td>3.2</td>
</tr>
<tr>
<td>Spain</td>
<td>3.2</td>
</tr>
<tr>
<td>Denmark</td>
<td>3.4</td>
</tr>
<tr>
<td>Germany</td>
<td>3.4</td>
</tr>
<tr>
<td>Italy</td>
<td>3.4</td>
</tr>
<tr>
<td>Belgium</td>
<td>3.6</td>
</tr>
<tr>
<td>France</td>
<td>3.6</td>
</tr>
<tr>
<td>Israel</td>
<td>3.7</td>
</tr>
<tr>
<td>Greece</td>
<td>3.8</td>
</tr>
<tr>
<td>Ireland</td>
<td>3.8</td>
</tr>
<tr>
<td>Netherlands</td>
<td>3.8</td>
</tr>
<tr>
<td>Switzerland</td>
<td>3.8</td>
</tr>
<tr>
<td>Austria</td>
<td>3.9</td>
</tr>
<tr>
<td>Australia</td>
<td>4.1</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>4.2</td>
</tr>
<tr>
<td>Canada</td>
<td>4.9</td>
</tr>
<tr>
<td>Poland</td>
<td>5.0</td>
</tr>
<tr>
<td>Hungary</td>
<td>5.3</td>
</tr>
<tr>
<td>New Zealand</td>
<td>5.7</td>
</tr>
<tr>
<td>United States</td>
<td>6.1</td>
</tr>
</tbody>
</table>

Source: CDC

Note: Canada data from 2009
Current Approaches and Limitations for Infant Mortality Prediction

- Use of a handful of risk factors from a specific domain of data (e.g., demographics, social services, etc)
- Ad-hoc selection of risk factors for predictive modeling
- Lack of data merging or linking from multiple sources
- Lack of translation of research into practice to increase clinical impact
Project collects data to save babies' lives

A new collaboration of healthcare providers, researchers, public agencies and nonprofit foundations is aimed at preventing infant deaths and promoting good health among Allegheny County residents.

The Richard King Mellon Foundation made several grants in June to the Magee-Womens Research Institute, the University of Pittsburgh and Rand Corporation to collaborate with experts and the community to propose solutions using data and models to uncover factors that might predict the risk of a baby dying before its first birthday.
Project Vision

• Background
  – Between 2008 and 2012, 434 infants died within their first year of life in Allegheny County out of more than 65,310 births
  – Allegheny County’s infant mortality rate of 6.65 deaths per 1,000 live births is higher than the national average*
  – Racial disparity in infant mortality is 27 percent larger in Allegheny County than nationally despite a robust health and social services network

• Goal
  To accurately identify infant mortality risk using a real-time, adaptive, predictive model and reduce that risk through referrals to an optimized and personalized mix of health, social service, and community prevention strategies.

Data Collaborators

- Allegheny County Department of Human Services (DHS)
  - program services data
- Allegheny County Health Department (ACHD)
  - Birth and death certificates
- Magee Women’s Research Institute (MWRI)
  - The Magee Obstetric Medical and Infant (MOMI)
- University of Pittsburgh Medical Center (UPMC)
  - Electronic health records (labs, meds, demographics, etc)
- Healthy Start [in progress]
- Birth Circle [in progress]
- ...
Outline

• Background

• Project approach and Current progress
  – Predictive modeling concept
  – Data collection and integration
  – Pilot study: MOMI data analysis and results

• Discussion and Conclusion
Overall Project Approach

• Phase 1 (2016 - 2018)
  – Data collection and integration (DHS, ACHD, MWRI, UPMC)
  – **Predictive system development**: Use historical data to develop a real-time, adaptive, predictive model of infant mortality risk
  – **Feasibility study**: Assess implementation challenges for the system
  – **Prevention strategies**: Identify effective interventions for the underlying causes of infant mortality based on the identified risks (by RAND Corp.)

• Phase 2 (from late 2018)
  – Implementation of the predictive and intervention system in a clinic, embedded in an existing EHR system
  – Prospective evaluation
Predictive Modeling from (Big) Data

- EHR Data
- Behavior Data
- Social Services Data
- Image/Video Data
- Mobile/Sensor Data
- Weather Data
- Social Network Data
- Environment Data

Multi data sources

Predictive Model

Data integration & de-identification

Machine Learning

Predictions

\[ P(\text{Infant Mortality} | \text{Data}) = 0.9 \]
Predictive Model: Mortality Risk and Survival Time

Available in narrative reports (progress notes, social work reports, etc)

- Age
- Race
- Pre-term labor
- Congenital anomalies
- Labs
- Meds
- ...

Clinical EHR Data

- Divorce
- Mom education level
- Bankruptcy
- ...

Social Context

- Breastfeeding
- Drug abuse
- Alcoholic
- ...

Behavioral Data

Environment Data

- Air quality
- Water
- Lead
- ...

Structured data

Bayesian network

Infant Mortality Predictive Modeling

- $P(\text{Mortality} \mid \text{Data})$
- Median survival time

Collaboration with ACHD, DHS, MWRI, and RAND Corp
Why Natural Language Processing?

As much as 80% of clinical data are locked in narrative free-text reports*

Outline

• Background

• Project approach and Current progress
  – Predictive modeling concept
  – Data collection and integration
  – Pilot study: MOMI data analysis and results

• Discussion and Conclusion
Record integration across multiple data sources

MARS (UPMC EHR data)
- MARS-format text (2002-2015) with 7 data types
  - Encounter [MOM: 2,465,113; CHILD: 598,355]
  - Diagnosis [MOM: 3,443,011; CHILD: 765,962]
  - Procedure [MOM: 595,195; CHILD: 219,100]
  - Lab [MOM: 14,564,776; CHILD: 4,591,030]
  - Prescription med [MOM: 5,247,714; CHILD: 231,101]
  - Dispensed med [MOM: 4,322,357; CHILD: 430,418]
  - Clinical notes (NLP processed) [in processing]

MOMI
- 13 csv files for 2002-2014
  - 128,023 records
  - 85,477 moms, 118,130 children, 117,929 deliveries

ACHD
- 1 excel spreadsheet file (3 tabs)
  - infant mortality cases: 1021 birth certificates
    - DHS_ACHD orig/match summary
    - DHS service data
    - ACHD original file (death certificate data+birth certificate data)

DHS
- 5 excel spreadsheet files
  - All birth population: 194,043 birth certificates
    - DHS_ACHD orig/match summary
    - DHS program list
    - ACHD Birth certificate data
    - DHS service data 2003-2007
    - DHS service data 2008-2013

C
- Pulled 81,245 out of 85,477 MOMI mothers' MARS data
- Pulled 99,523 out of 118,130 MOMI children's MARS data

A
- Mapped 62,676 out of 140,663 ACHD mother entity

B
- Mapped 84,107 MOMI deliveries with 194,043 ACHD birth certificates

Performed by an Honest Data Broker
### Mother Entity Matching

#### Data Source

<table>
<thead>
<tr>
<th>Data Source</th>
<th># of Unique Moms</th>
<th>Matched count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOMI</td>
<td>85,477</td>
<td>60,903 (71.25%)</td>
</tr>
<tr>
<td>ACHD/DHS</td>
<td>140,663</td>
<td>60,903 (43.30%)</td>
</tr>
</tbody>
</table>

- Mother entities fully linked across the three data sources.
Outline

• Background

• Project approach and Current progress
  – Predictive modeling concept
  – Data collection and integration
  – Pilot study: MOMI data analysis and results

• Discussion and Conclusion
MOMI Data Quality Analysis

• Comprising 251 variables from (parents, infant) demographics, diseases, mom’s behavioral history (14 variables), healthcare utilization, etc from 2003 to 2013

• Data quality analysis
  – **Completeness**: check missing values prevalence, e.g., 0.5% (636) with missing mom’s SSN
  – **Validity**: variable’s value range profile, e.g.,
  – **Consistency**: compare derived variables with original variables (e.g., teenage_mom_flag is False but mom_age <20)
  – **Redundancy**: test duplicates to avoid bias

Pregnancy weeks histogram
MOMI Risk Factor Analysis

• **Univariate** logistic regression with adjusted p-values

• \( \ln \left( \frac{P(Y=1)}{1-P(Y=1)} \right) = \beta_0 + X\beta_1 \), where \( \beta_0 \) and \( \beta_1 \) are regression coefficients

• Use Benjamini & Hochberg (BH) method to control false discover rate (FDR) at 0.05
### Significant Risk Factors: Demographics

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Group v.s. reference group</th>
<th>Odds ratio (95% CI)</th>
<th>Adjusted P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal age</td>
<td>≤ 26 vs. &gt;26</td>
<td>2.0498 (1.7177, 2.4466)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Maternal education</td>
<td>High school or below vs. college education</td>
<td>2.6788 (2.1738, 3.3118)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Graduate school vs. college education</td>
<td>0.6676 (0.4564, 0.9508)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown vs. college education</td>
<td>4.6675 (3.5771, 6.0604)</td>
<td></td>
</tr>
<tr>
<td>Maternal marital status at delivery</td>
<td>Singe vs. married</td>
<td>3.5602 (2.9391, 4.3334)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Divorced, widowed, life partner, separated or</td>
<td>2.3781 (1.1702, 4.2863)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>committed relationship vs. married</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown, vs. married</td>
<td>1.2049 (0.2970, 3.1817)</td>
<td></td>
</tr>
<tr>
<td>Maternal race</td>
<td>African American vs. Caucasian</td>
<td>4.0065 (3.3481, 4.7946)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Native American, Hispanic, Asian, 2.1287 (0.9623, 4.0304)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mixed and other vs. Caucasian</td>
<td>2.1287 (0.9623, 4.0304)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown vs. Caucasian</td>
<td>0.7012 (0.2764, 1.4407)</td>
<td></td>
</tr>
</tbody>
</table>
## Significant Risk Factors: Behavior and Reproductive History

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Group vs. reference group</th>
<th>Odds ratio (95% CI)</th>
<th>Adjusted P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of pregnancies</td>
<td>&gt;3 vs. ≤3</td>
<td>1.7678 (1.4485, 2.1455)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Maternal substance abuse</td>
<td>alcohol vs. none</td>
<td>40.7918 (0.2956, 502.1857)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Hallucinogens vs. none</td>
<td>611.8701 (32.5999, 89330.6560)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Amphetamines vs. none</td>
<td>15.6890 (0.1202, 132.9522)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Barbiturates vs. none</td>
<td>5.2297 (0.0411, 38.1292)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cocaine vs. none</td>
<td>4.0142 (2.0284, 7.0598)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multiple diagnostic codes vs. none</td>
<td>4.3395 (0.0342, 31.2416)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other vs. none</td>
<td>4.3768 (0.9123, 12.5353)</td>
<td></td>
</tr>
<tr>
<td>Maternal marijuana use SMOKE</td>
<td>Yes vs. no</td>
<td>4.0619 (2.5617, 6.0976)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Yes vs. no</td>
<td>1.7885 (1.4477, 2.1932)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Total number of prenatal visits</td>
<td>&gt;7 and &lt;=11 vs. &gt;11</td>
<td>2.7018 (2.0874, 3.4526)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>&gt;5 and &lt;=7 vs. &gt;11</td>
<td>8.0349 (5.3528, 11.5970)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;=5 vs. &gt;11</td>
<td>29.8761 (22.6929, 38.8341)</td>
<td></td>
</tr>
</tbody>
</table>
Outline

• Background

• Project approach and Current progress
  – Predictive modeling concept
  – Data collection and integration
  – Pilot study: MOMI data analysis and results

• Discussion and Conclusion
Discussion and Conclusion

• Verified MOMI data and identified individual risk factors from MOMI dataset that are also reflected in literatures

• Using multiple linked data sources enables comprehensive research on infant mortality

• Univariate non-ad-hoc analysis is a necessary step for building predictive modeling

• Limitation: the use of a portion of AC infant deaths from MWRI data (~50% infant delivery coverage)

• Next steps: build a predictive model from all data sources, add intervention strategies, and conduct feasibility assessment
Acknowledgements

Jose Posada, Amie Barda, Victor Ruiz, Ye Ye

Helen Shi, Fan Mi, Diyang Xue, Howard Su

Funding source: Richard King Mellon Foundation

• 3 FTE
• 1 Postdoc fellow
• 6 PhD Students in
  • DBMI(3)
  • ISP(2)
  • Bioeng.(1)

Luge Yang, PhD
Any questions?

Rich Tsui, PhD
tsui2@pitt.edu