Achieving the Promise of Interoperability & Big Data: Real World-Examples of Successful Health Data Analytics Leveraging the U.S.’s Largest & Longest Running HIE

Shaun Grannis MD, MS, FAAFP, FACMI
Director, The Regenstrief Center for Biomedical Informatics

Prepared for the Int’l Symposium on Data Driven Healthcare, 2017
About Regenstrief Institute

Regenstrief Institute was established in 1969 by philanthropist Samuel Regenstrief with the goal of improving the quality and efficiency of healthcare delivery, preventing medical errors, and enhancing patient safety.
EHR Integration: The Indiana Network for Patient Care (INPC)

**Data Management**
- Hospital
- Payers
- Labs
- Outpatient RX
- Physician Office
- Ambulatory Centers
- Public Health

**Data Access & Use**
- Hospitals
  - Results delivery
  - Secure document transfer
  - Shared EMR
  - Credentialing
  - Eligibility checking

- Physicians
  - Results delivery
  - Secure document transfer
  - Shared EMR
  - CPOE
  - Credentialing
  - Eligibility checking

- Labs
  - Results delivery

- Public Health
  - Surveillance
  - Reportable conditions
  - Results delivery
  - De-identified, longitudinal clinical data

- Payer
  - Secure document transfer
  - Quality Reporting

- Researcher
  - De-identified, longitudinal clinical data (OHDSI CDM, i2b2)
  - Subject Recruitment
  - Clinical Trials
HIE Statistics

Source: http://www.ihie.org
Secondary Use of Clinical Data

• Any use of health data not specifically intended or anticipated when data is initially obtained.
• Secondary use is defined in the complementary sense (i.e., NOT primary use.)
Machine learning’s ability to produce actionable results from unstructured data is clearly demonstrated in a study published in
Example: Identifying Cancer from Free-Text

Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection

Suranga N. Kasthurirathne a,*, Brian E. Dixon b,c, Judy Gichoya d, Huiping Xu c, Yuni Xia d, Burke Mamlrn b,d, Shaun J. Grannis b,d
### Surgical Pathology Report

**Clinical History:** Large Gastric Mass

**Specimen:** Gastric Mucosa

**Diagnosis:**
- **Malignant Epithelioid Gastrointestinal Stromal Tumor**
  - Tumor size: 10 x 9 x 8 cm
  - Cell Type: Epithelioid and Spindle
  - High cellularity; present
  - Muscular invasion: Focally present adjacent to ulceration
  - Necrosis: present
  - Mitotic Count: <10/50 HPF
  - Myxoid background: Focally present
  - Focal of necrosis present
  - CD117, Vimentin, and CD34: Uniformly positive

**Gross Description:**
The specimen consists of an approximately 5 x 7 cm portion of gastric mucosa that is surrounded and underlying by a lobulated mass which is 10 x 9 x 8 cm. The central portion of the mass appears to have an approximately 1.5-cm ulcer. The mucosa away from the area of ulceration is partially removed from the underlying tumor. The underlying mass appears encapsulated and lobular. Gross sections show the lesion to consist of several different patterns. A single area has a gray to gray-tan pattern with an area of central necrosis showing a fairly uniform appearance whereas other regions of the tumor are gray-white and somewhat lobular in appearance. Areas of yellow necrosis are scattered through the tumor. Representative portions submitted.

**Microscopic Description:**
Sections through the neoplasm show it to be primarily a high cellular neoplasm. The cells are in part arranged in fascicles and sheets, with antrum-dominant nuclei having relatively few nucleoli. In some areas, the fascicles have an interwoven appearance. Mitotic figure up to 10/50 HPF. A few areas show foci of necrosis with the cells appearing to be surrounded by somewhat myxoid stroma. Foci of displayed necrosis are present. The lesions appear circumscribed, although not specifically encapsulated. It focally involved the mucosa and shows full thickness ulceration. The tumor immediately beneath the mucosal area of ulceration has a nearly lobular somewhat spindled growth pattern. Some areas of the tumor have a slightly more rounded nuclei and somewhat epithelioid appearance. The cells appear to be arranged in groups and clusters. Some of the cells have crystalloids in vacuoles. These areas also show a prominent mitotic activity.

Some mitotic figures are abnormal and atypical. The tumor contains numerous relatively open vascular channels which appear to be part of the neoplasm. The tumor has a pseudo capsule and in some areas appear to be nearly covered.

**Immunostains:** Strongly positive for CD117 (C-kit), CD34, and Vimentin, Smooth muscle actin, Desmin, Synaptophysin, S-100, and Cdx1/2 are negative.

**Comment:**
Immunostains were performed on the core biopsy and demonstrate that the tumor cells are positive for CD117. The findings are consistent with the above diagnosis.
Overall Process

1. INPC Data
   - Extract INPC Plaintext Pathology Reports
   - Review gold Standard
   - Create master feature vector

2. Feature selection (FS) approach
   - Manual Selection
   - Informed
   - Automated

3. Feature subset size
   - 5
   - 10
   - 15
   - 20

4. Decision model
   - SLR
   - NB
   - KNN
   - RF
   - J48

5. Evaluation
   - Sensitivity
   - Specificity
   - Accuracy
   - PPV
   - ROC
Results

- Sensitivity: 85-90%
- Positive Predictive Value: 95-97%
- Specificity: 97-99%
- Random Forest and Logistic Regression exhibited highest AUC
Notifiable Condition Detector

Realtime

Compare to NCMT

Abnormal flag, Organism name in NCMT, Value above threshold

Inbound Message

Potentially Reportable

Reportable Condition

Reportable Conditions Databases

E-mail Summary

Daily Batch

To Public Health

To Infection Control

Reports

Record Count as denominator
Table 1: Salmonella detection accuracy stratified by decision model and the number of tokens used.

<table>
<thead>
<tr>
<th>Model accuracy</th>
<th>Classifier:</th>
<th></th>
<th>Classifier:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens used</td>
<td>Logistic: Sensitivity</td>
<td>specificity</td>
<td>ROC</td>
<td>Random forest: Sensitivity</td>
</tr>
<tr>
<td>First 4</td>
<td>0.951</td>
<td>0.926</td>
<td>0.927</td>
<td>0.951</td>
</tr>
<tr>
<td>First 10</td>
<td>0.962</td>
<td>0.941</td>
<td>0.962</td>
<td>0.965</td>
</tr>
<tr>
<td>First 19</td>
<td>0.955</td>
<td>0.928</td>
<td>0.979</td>
<td>0.958</td>
</tr>
<tr>
<td>First 38</td>
<td>0.944</td>
<td>0.913</td>
<td>0.970</td>
<td>0.962</td>
</tr>
</tbody>
</table>

Population Analytics
A total of 96 emergency departments contributed data to this analysis.
Neuro Event

![Graph showing ED Visit Count from 30-Nov-05 to 30-Dec-05 with a peak on 24-Dec-05. A horizontal line at 14 indicates a threshold above which an event is considered.]
Distribution of patients stratified by the total number of ED visits. Note that six patients visited the ED more than 300 times and a single patient accumulated 385 visits for the 3-year study period.
Leveraging Analytics to Develop Behavioral Phenotypes

A network diagram illustrating the connectedness among Indiana EDs that participate in PHESS. Circular nodes represent EDs; node size indicates the visit volume; node color indicates the centrality of the ED. The gray edges connecting nodes indicate where patient crossover occurs. EDs that share proportionally larger number of patients are clustered together. While general clusters of “medical trading areas” emerge, the myriad gray edges clearly illustrate how interconnected all EDs are to one another.

- Patients receive healthcare from multiple providers and across organizations
- More than 40% of ED visits are for patients having data at multiple institutions
# The Results

<table>
<thead>
<tr>
<th>Model No.</th>
<th>Multivariable Logistic Regression Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>No. of visits constituting ‘frequent use’</td>
<td>&gt;= 8</td>
</tr>
<tr>
<td>Area under ROC curve (AUC)</td>
<td>0.84</td>
</tr>
<tr>
<td>With sensitivity &lt;=25 %, probability &gt; 0.5</td>
<td></td>
</tr>
<tr>
<td>PPV (%)</td>
<td>64.5</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>99.5</td>
</tr>
<tr>
<td>False positive patients</td>
<td>Total No.</td>
</tr>
<tr>
<td></td>
<td>5883</td>
</tr>
<tr>
<td></td>
<td>&gt;= 8 visits (No.)</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Adjusted PPV for patients with &gt;=8 visits in subsequent two years (%)</td>
<td>64.5</td>
</tr>
</tbody>
</table>

The “high ED user” phenotype

**Model A**

\[
\text{Log (Odds)} = -5.57 + 0.28 \times \text{Age(<5)} + 1.09 \times \text{Age(15-24)} + 0.97 \times \text{Age(25-44)} + 0.71 \times \text{Age(45-64)} + 0.50 \times \text{Age(>=65)} + 0.32 \times \text{Female} + 0.34 \times \text{Distance (<=5)} - 0.24 \times \text{Distance (>20)} + 0.58 \times \text{Visits}_2008 - 0.0075 \times (\text{Visits}_2008)^2 + 0.08 \times \text{CC_GI} - 0.10 \times \text{CC_Skin} + 0.42 \times \text{CC_RESP} + 0.11 \times \text{CC_NEURO} + 0.001 \times \text{CC_UDI} - 0.20 \times \text{CC_ILI} + 0.08 \times \text{CC_Pain} + 0.44 \times \text{CC_Dental} - 0.16 \times \text{CC_MUSC} + 0.07 \times \text{CC_Lymphatic} + 0.23 \times \text{CC_Alcohol} - 0.10 \times \text{CC_Unclassified}
\]

**Model B**

\[
\text{Log (Odds)} = -7.93 + 0.08 \times \text{Age(<5)} + 1.77 \times \text{Age(15-24)} + 1.78 \times \text{Age(25-44)} + 1.47 \times \text{Age(45-64)} + 0.68 \times \text{Age(>=65)} + 0.27 \times \text{Female} + 0.33 \times \text{Distance (<=5)} - 0.069 \times \text{Distance (>20)} + 0.50 \times \text{Visits}_2008 - 0.0062 \times (\text{Visits}_2008)^2 + 0.12 \times \text{CC_GI} - 0.42 \times \text{CC_Skin} + 0.45 \times \text{CC_RESP} + 0.15 \times \text{CC_NEURO} - 0.10 \times \text{CC_UDI} - 0.34 \times \text{CC_ILI} + 0.21 \times \text{CC_Pain} + 0.37 \times \text{CC_Dental} - 0.26 \times \text{CC_MUSC} - 0.08 \times \text{CC_Lymphatic} + 0.48 \times \text{CC_Alcohol} - 0.08 \times \text{CC_Unclassified}
\]

**Fig. 2**

Equations for model predicting frequent emergency department (ED) use as defined as 8 or more visits (a) and model predicting frequent ED use as defined as 16 or more visits in the subsequent two years (b). Distance (<=5): the straight-line distances between geographic points from patients’ home to hospital less than 5 miles; Distance (>20): the straight-line distances between geographic points from patients’ home to hospital greater than 20 miles; CC: chief complaints; GI: gastrointestinal; RESP: respiratory; NEURO: neurological; UDI: undifferentiated infection; ILI: influenza-like illness and MUSC: musculoskeletal.

Predicting ED High Utilizers

6mo Prediction ROC

- Lasso (0.75)
- Lasso – squared interactions (0.84)
- Random Forest (0.98)
Flu ICD9 HIE
Flu CC PHESS
Pneumonia ICD9 HIE
Pneumonia CC PHESS
ILI ICD9 HIE
ILI CC PHESS
All Flu Tests HIE
Positive Flu Tests NCD
Positive Rate ALL
Environmental – Linking Clinical Data to the Environment
What Makes Us Healthy versus What We Spend on Being Healthy.

Families Living in Poverty

By Counties

By Census Tracts (Marion County)

By Townships

By Block Groups (Marion County)

Source: SAVI Community Information System and 2000 Census
- Labs
- Diagnoses
- Medications

- SNP’s
- Whole Exome
- Whole Genome

- Pollution
- Radon
- Temperature
- SES
- Crime

- ED usage
- Outpatient visits
- Inpatient admissions
- Online behavior
- Quantified self
- etc.
Achieving the Promise of Interoperability & Big Data: Real World-Examples of Successful Health Data Analytics Leveraging the U.S.’s Largest & Longest Running HIE

Shaun Grannis MD, MS, FAAFP, FACMI
Director, The Regenstrief Center for Biomedical Informatics

Prepared for the Int’l Symposium on Data Driven Healthcare, 2017