Integrative Biomedical Informatics and Clinical Phenotype Characterization

Joel Saltz MD, PhD
Director Center for Comprehensive Informatics
Disclosures

• Member of Appistry Scientific Advisory Board
INTEGRATIVE BIOMEDICAL INFORMATICS ANALYSIS

• Reproducible anatomic/functional characterization at fine level (Pathology) and gross level (Radiology)
• Integration of anatomic/functional characterization with multiple types of “omic” information
• Create categories of jointly classified data to describe pathophysiology, predict prognosis, response to treatment
• Characterize clinical phenotype
Integration of heterogeneous multiscale information

- Coordinated initiatives: Pathology, Radiology, "omics"
- Exploit synergies between all initiatives to improve ability to forecast survival & response.
Unsupervised Morphological Clustering – Outcome and Molecular Correlates
Extreme Spatio-Temporal Sensor Data Analytics

- Leverage exascale data and computer resources to squeeze the most out of image, sensor or simulation data
- Run lots of different algorithms to derive same features
- Run lots of algorithms to derive complementary features
- Data models and data management infrastructure to manage data products, feature sets and results from classification and machine learning algorithms
Clinical Phenotype Characterization and the Emory Analytic Information Warehouse

- Example Project: Find hot spots in readmissions within 30 days
  - What fraction of patients with a given principal diagnosis will be readmitted within 30 days?
  - What fraction of patients with a given set of diseases will be readmitted within 30 days?
  - How does severity and time course of co-morbidities affect readmissions?
  - Geographic analyses

- Compare and contrast with UHC Clinical Data Base
  - Repeat analyses across all UHC hospitals
  - Are we performing the same?
  - How are UHC-curated groupings of patients (e.g., product lines) useful?

- Need a repeatable process that we can apply identically to both local and UHC data
5-year Datasets from Emory and University Healthcare Consortium

- EUH, EUHM and WW (inpatient encounters)
- Removed encounter pairs with chemotherapy and radiation therapy readmit encounters (CDW data)
- Encounter location (down to unit for Emory)
- Providers (Emory only)
- Discharge disposition
- Primary and secondary ICD9 codes
- Procedure codes
- DRGs
- Medication orders (Emory only)
- Labs (Emory only)
- Vitals (Emory only)
- Geographic information (CDW only + US Census and American Community Survey)

Analytic Information Warehouse
Using Emory & UHC Data to Find Associations With 30-day Readmits

- **Problem:** “Raw” clinical and administrative variables are difficult to use for associative data mining
  - Too many diagnosis codes, procedure codes
  - Continuous variables (e.g., labs) require interpretation
  - Temporal relationships between variables are implicit

- **Solution:** Transform the data into a much smaller set of variables using heuristic knowledge
  - Categorize diagnosis and procedure codes using code hierarchies
  - Classify continuous variables using standard interpretations (e.g., high, normal, low)
  - Identify temporal patterns (e.g., frequency, duration, sequence)
  - Apply standard data mining techniques
**Derived Variables**

30-day readmit

The 9 Emory Enhanced Risk Assessment Tool diagnosis categories

UHC product lines

Variables derived from a combination of codes and/or laboratory test results

- Obesity
- Diabetes/uncontrolled diabetes
- End-stage renal disease (ESRD)
- Pressure ulcer
- Sickle cell disease/sickle cell crisis

Temporal variables derived over multiple encounters

- Multiple MI
- Multiple 30-day readmissions
- Chemotherapy within 180 (or 365) days before surgery
- Previous encounter within the last 90 (or 180) days
# 30-Day Readmission Rates for Derived Variables

Emory Health Care

<table>
<thead>
<tr>
<th>Patient Population</th>
<th>Number of Encounters</th>
<th>Number of Readmissions</th>
<th>Readmission Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-Emory</td>
<td>202181</td>
<td>36734</td>
<td>15%</td>
</tr>
<tr>
<td>Multiple MI</td>
<td>4414</td>
<td>1506</td>
<td>36% (Single MI 15%)</td>
</tr>
<tr>
<td>ESRD</td>
<td>18445</td>
<td>5036</td>
<td>27% (CKD 23%)</td>
</tr>
<tr>
<td>&gt;=4 readmissions</td>
<td>19510</td>
<td>10707</td>
<td>55%</td>
</tr>
<tr>
<td>Multiple MI <em>and</em> &gt;= 4 readmissions</td>
<td>997</td>
<td>520</td>
<td>52%</td>
</tr>
<tr>
<td>CKD <em>and</em> &gt;=4 readmissions</td>
<td>7865</td>
<td>4110</td>
<td>52%</td>
</tr>
<tr>
<td>Uncontrolled diabetes</td>
<td>12219</td>
<td>2573</td>
<td>21% (Diabetes 19%)</td>
</tr>
<tr>
<td>Uncontrolled diabetes &amp; pressure ulcer</td>
<td>648</td>
<td>201</td>
<td>31%</td>
</tr>
<tr>
<td>Uncontrolled diabetes &amp; ESRD</td>
<td>1645</td>
<td>531</td>
<td>32%</td>
</tr>
<tr>
<td>Sickle cell crisis</td>
<td>1809</td>
<td>663</td>
<td>37% (Sickle cell anemia 34%)</td>
</tr>
<tr>
<td>MRSA</td>
<td>1565</td>
<td>410</td>
<td>26%</td>
</tr>
<tr>
<td>Stroke and MRSA</td>
<td>42</td>
<td>16</td>
<td>38% (Stroke 24%)</td>
</tr>
<tr>
<td>MI and MRSA</td>
<td>140</td>
<td>43</td>
<td>31% (MI 15%)</td>
</tr>
</tbody>
</table>
Geographic Analyses
UHC Medicine General Product Line (#15)

Readmission $O:E = \frac{\text{# of 30-day readmits in the census tract}}{\text{# of 30-day readmits overall}}$

$\frac{\text{# of encounters in the census tract}}{\text{# of encounters overall}}$

Income Levels
- $< 25000$
- $25000 - 50000$
- $50000 - 75000$
- $75000 - 100000$
- $> 100000$
Predictive Modeling for Readmission

• Random forests (ensemble of decision trees)
  – Create a decision tree using a random subset of the variables in the dataset
  – Generate a large number of such trees
  – All trees vote to classify each test example in a training dataset
  – Generate a patient-specific readmission risk for each encounter

• Rank the encounters by risk for a subsequent 30-day readmission
# Emory Readmission Rates for High and Low Risk Groups Generated with Random Forest

<table>
<thead>
<tr>
<th>ERAT Category</th>
<th>30-Day Readmission Rate (%)</th>
<th>Predictive Modeling High Risk Group</th>
<th>Predictive Modeling Low Risk Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>18.46</td>
<td>41.59</td>
<td>4.45</td>
</tr>
<tr>
<td>CKD</td>
<td>22.03</td>
<td>47.32</td>
<td>7.04</td>
</tr>
<tr>
<td>COPD</td>
<td>17.48</td>
<td>45.52</td>
<td>7.02</td>
</tr>
<tr>
<td>Diabetes</td>
<td>16.87</td>
<td>42.47</td>
<td>8.72</td>
</tr>
<tr>
<td>Transplant</td>
<td>22.77</td>
<td>41.86</td>
<td>11.51</td>
</tr>
<tr>
<td>MI</td>
<td>13.76</td>
<td>40.41</td>
<td>2.97</td>
</tr>
<tr>
<td>Pulmonary Hypertension</td>
<td>21.06</td>
<td>47.83</td>
<td>3.72</td>
</tr>
<tr>
<td>Sickle Cell Anemia</td>
<td>34.03</td>
<td>46.36</td>
<td>19.42</td>
</tr>
<tr>
<td>Stroke</td>
<td>12.28</td>
<td>48.40</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Graphics (non-programmer!) Interface
Display and Integrative Analysis of Combined Physiologic and EHR Data

Numerics and Waveforms (240 Hz)

Excel Medical Server
BedMaster XA

~ 2 sec latency

Slide from Timothy Buchman
Status of Clinical Phenotype Analysis and Characterization

- Integrative dataset analysis can leverage patient information gathered over many encounters
- Temporal analyses can generate derived variables that appear to correlate with readmissions
- Predictive modeling has promise of providing decision support
- Joint work with Kaiser on treatment patterns and efficacy for newly treated African American hypertensives
- Ongoing analyses involve characterization of clinical phenotype in GWAS, biomarker and quality improvement efforts
- Collaborations with Tim Buchman, Sharath Cholleti in Critical Care informatics
- Co-lead (with Bill Hersh) of CTSA CER Informatics taskforce dedicated informatics support for Pragmatic Clinical Trials
Summary and Perspective

- Large scale integrative data analytic methods and tools to integrate clinical, molecular, Pathology, Radiology data (happy to discuss Radiology aspects off line)
- Coordinated methods to analyze healthcare data and to define clinical phenotypes
- Generate and manage nuanced temporal summary of patients health status, co-morbidities, treatment, treatment response
Thanks to:

- In silico center team: Dan Brat (Science PI), Tahsin Kurc, Ashish Sharma, Tony Pan, David Gutman, Jun Kong, Sharath Cholleti, Carlos Moreno, Chad Holder, Erwin Van Meir, Daniel Rubin, Tom Mikkelsen, Adam Flanders, Joel Saltz (Director)
- caGrid Knowledge Center: Joel Saltz, Mike Caliguiri, Steve Langella co-Directors; Tahsin Kurc, Himanshu Rathod Emory leads
- caBIG In vivo imaging team: Eliot Siegel, Paul Mulhern, Adam Flanders, David Channon, Daniel Rubin, Fred Prior, Larry Tarbox and many others
- In vivo imaging Emory team: Tony Pan, Ashish Sharma, Joel Saltz
- Emory ATC Supplement team: Tim Fox, Ashish Sharma, Tony Pan, Edi Schreibmann, Paul Pantalone
- Digital Pathology R01: Foran and Saltz; Jun Kong, Sharath Cholleti, Fusheng Wang, Tony Pan, Tahsin Kurc, Ashish Sharma, David Gutman (Emory), Wenjin Chen, Vicky Chu, Jun Hu, Lin Yang, David J. Foran (Rutgers)
- NIH/in silico TCGA Imaging Group: Scott Hwang, Bob Clifford, Erich Huang, Dima Hammoud, Manal Jilwan, Prashant Raghavan, Max Wintermark, David Gutman, Carlos Moreno, Lee Cooper, John Freymann, Justin Kirby, Arun Krishnan, Seena Dehkharghani, Carl Jaffe
- ACTSI Biomedical Informatics Program: Marc Overcash, Tim Morris, Tahsin Kurc, Alexander Quarshie, Circe Tsui, Adam Davis, Sharon Mason, Andrew Post, Alfredo Tirado-Ramos
- NSF Scientific Workflow Collaboration: Vijay Kumar, Yolanda Gil, Mary Hall, Ewa Deelman, Tahsin Kurc, P. Sadayappan, Gaurang Mehta, Karan Vahi
The AIW/CVRG/MH-GRID Teams

• Stakeholders
  – Joel Saltz, MD, PhD – CCI Director and ACTSI BIP Director
  – William Bornstein, MD, PhD – Emory Healthcare Chief Quality Officer
  – Dee Cantrell, RN – Emory Healthcare CIO
  – Marc Overcash – Emory Deputy CIO of Research and Health Sciences IT

• Project Team
  – Andrew Post, MD, PhD – AIW Project Lead & CCI Clinical Informatics Architect
  – Terry Willey, RN – IS Director of Business Strategy/Planning
  – Richie Willard – Project Manager
  – Tahsin Kurc, PhD – CCI Chief Software Architect
  – Sharath Cholleti, PhD – Research Scientist
  – Jingjing Gao, PhD – Biostatistician
  – Michel Mansour – Software Engineer
  – Himanshu Rathod – Software Engineer
  – Mike Torian – Data Warehouse Engineer
  – Michael Brown – Software Engineer
  – Geoff Milton – Software Engineer
  – Akshatha Kalsanka Pai – Software Engineer
Thanks!