Big Data Infrastructure and Analytics for Healthcare Applications

Arjun Shankar, Ph.D.
shankarm@ornl.gov
Computational Sciences and Engineering Division
Outline

• Current “big data” context in healthcare

• Data processing and analysis for healthcare delivery and effectiveness
  – Big data infrastructure to support scale, analytics, and visualization
  – Innovative data representations and applications
  – Population health and analysis of geographic variation

• Bridging clinical data and claims data
  – Data formats for clinical data
  – Linking clinical and claims data
Setting the Big Data Context
Individual Data Scales: Human Biology and Medicine

Genotype to Phenotype Spectrum

Petabytes, Terabytes, Giga-bases/second

Gigabytes

Megabytes
Population Data Scales: Healthcare Delivery and Outcomes Focus

Individual to Population

Megabytes  Gigabytes  Petabytes

(300 Million Individuals 5B Medicare-related events/year)
Big Data Buzz?

• More “sensors”: connected devices, social media, literature,
• Better data management and processing: scale, interactive, networked, named, mobile, etc.
• Improved analytics: machine learning, semantics, visualization, ...

Big data infrastructures and analytics: use of emerging tools and data-driven techniques to go beyond your traditional data-capacity, data-processing, and data-analysis capabilities.
Payment Reforms Will Motivate and Reward Innovation at a Whole New Level

- Patient Centered Medical Homes: Organized outpatient care, coordination and team-based approaches
- Accountable Care Organizations: Shared savings; redesigned care processes for high quality, efficient delivery
- Bundled Payments: Pilot program for episodes of care; incentivizes reduced costs around eight conditions
- Readmission Reduction Programs: Motivates hospitals to engage with care coordinators and better organize delivery systems

Innovations Needed:
- Data Mining/Analytics
- Care Integration Tools
- Timely Clinical Data, Decision Support
- Technology to Extend Physician Reach
- Consumer Engagement Tools/Platforms/Apps

From Todd Park, SXSW 2011 Presentation
Transformation Challenges for National Healthcare Environment

Person Centric | Value Based Payment | Outcome Driven | Continuous Improvement
---|---|---|---

1. How is the Beneficiary Doing? (better care, better health, lower cost)
2. Collect Performance Measures
3. Outcomes Based Analysis
4. Make Payment

Line of Sight to Benefices
Investments & Tools for Improvement
Line of Sight to Providers

Outcome Measures
Outcomes Based

States
Providers

Value Based Payment
Health Knowledge

Continuous Improvement
Users of Healthcare Big Data Processing

- Providers and healthcare delivery systems
  - Readmission prediction and reduction
  - Care-coordination to reduce cost
  - Measure outcomes

- Insurers/Payers
  - Actuarial calculations by age, region, etc.
  - Bundling payments

- Policy Makers
  - Cost to quality ratios – what works, what to cover, etc.

- Efficiency analysts – Fraud Detection, Clinical Standards, System efficiencies
  - Streamlining processes, clinical standards adherence, mitigating over payments, double charging, upcoding
Claims Streams Reflect Healthcare

Employer-paid Plans Example

Taxpayer-paid Plans Example

Payment Models Encoded in Claims and Codes
Data Vignettes

- Example of Medicare claim data-format elements

<table>
<thead>
<tr>
<th>Short SAS Name</th>
<th>Long SAS Name</th>
<th>Short Description</th>
<th>Type</th>
<th>Length</th>
<th>IP</th>
<th>OP</th>
<th>SNF</th>
<th>Hospic</th>
<th>HHA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BENE_ID</td>
<td>BENE_ID</td>
<td>Encrypted 723 Beneficiary ID</td>
<td>CHAR</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CLM_ID</td>
<td>CLM_ID</td>
<td>Claim ID</td>
<td>CHAR</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>RIC_CD</td>
<td>NCH_NEAR_LINE_REC_IDENT_CD</td>
<td>NCH Near Line Record Identification Code</td>
<td>CHAR</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>CLM_TYPE</td>
<td>NCH_CLM_TYPE_CD</td>
<td>NCH Claim Type Code</td>
<td>CHAR</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>FROM_DT</td>
<td>CLM_FROM_DT</td>
<td>Claim From Date</td>
<td>DATE</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>THRU_DT</td>
<td>CLM_THRU_DT</td>
<td>Claim Through Date</td>
<td>DATE</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>WKLY_DT</td>
<td>NCH_WKLY_PROC_DT</td>
<td>NCH Weekly Claim Processing Date</td>
<td>DATE</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>FI_CLM_PROC_DT</td>
<td>FI_CLM_PROC_DT</td>
<td>FI Claim Process Date</td>
<td>DATE</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>QUERY_CD</td>
<td>CLAIM_QUERY_CODE</td>
<td>Claim Query Code</td>
<td>CHAR</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>PROVIDER</td>
<td>PRVDR_NUM</td>
<td>Provider Number</td>
<td>CHAR</td>
<td>6</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

- Example of Medicare claim form

```
21. Diagnosis or nature of illness or injury (Relate items 1, 2, 3 or 4 to item 24E by line)
1. ______.
2. ______.
3. ______._
4. ______.
```

```
22. Medicaid resubmission code
```

```
23. Prior authorization number
```

```
24. Date(s) of service
From MM DD YY To MM DD YY
Place of Service: EMG
```

```
D. Procedures, services, or supplies (Explain unusual circumstances)
```

```
E. Diagnosis pointer
```

```
F. $ Charges
```

```
G. Days
```

```
H. ESGT Family Plan
```

```
I. ID Qual
```

```
J. Rendering provider ID #
```

UPPLIER INFORMATION

NPI

NPI

NPI
Setting up and Deploying a Big Data Infrastructure
ORNL Knowledge Discovery Infrastructure for Big Data Processing and Analysis

Applications

Analytics

Infrastructure

Data-Feeds, Extract-Transform-Load

Analytics

Statistics Analysis

Social Network Analysis

Geospatial-Temporal

Machine Learning

Analytics

Key Value Datastore

BI Solution

Visualization Libraries (JS)

Online User

Hosted User

Analytics

Hive

MPP DB

Loader - Sqoop, bulk-loader

XenCenter - VM, Storage and Compute Resource Management

Knowledge Discovery Infrastructure - ORNL Knowledge Discovery Infrastructure for Big Data Processing and Analysis

Data-Feeds, Extract-Transform-Load
A Use-Pattern for Big Data over Hadoop

1. User submits SQL to Hive
2. Hive transforms SQL to MapReduce
3. MapReduce is executed in Hadoop cluster
4. Data is loaded into Hadoop filesystem (HDFS)
Typical Recipe for Data Analysis Workflow

• Process using scalable infrastructure (Extract-Transform-Load)
• Data cleaning and organization
• Input into relational (Hive, MySQL, Column-databases, etc.)
• Run queries to extract relevant information
  – SQL-based queries, SAS-based programs, etc.
  – Learn correlations, patterns, causal behaviors, data-dependencies
• Visualize
• Re-run extractions of relevant data
Software Interconnections View

Push as much of the initial data into HDFS and Hive (reliability, low cost)

Hadoop
- Hive
- HDFS

Custom - Transformation and Analytics
- Java - ETL
- Java - Map/Reduce
- Python - CPI

HTTP(S)
- Node.js
- CouchDB

Data traffic over HTTP is either user interaction or RESTful calls

HTTP(S)
- Tomcat
  - Servlet
  - OD4A4J
  - JDBC
- MPP DB
- Neo4J

TCP/IP (SSH)
Customized Population Health Views
Analyzer for Duals
CMS_DUALS_CROSSWALK_NORTHCAROLINA

Select Service
Inpatient

Select Top N Diagnostic Codes
10

Medicaid Costs for top 10 diagnostic Codes

Medicare Costs for top 10 diagnostic codes

Top Medicaid Diagnostic Codes
- Routine medical
- Chest pain NEC
- Hb-SS disease w.
- Hyp kid NOS w.
- Acute kidney fail.
- Pneumonia, orga.
- Septicemia NOS
- Obs chr bronc w.
- Rehabilitation pr.

Top Medicare Diagnostic Codes
- Septicemia NOS
- Hb-SS disease w.
- Rehabilitation pr.
- Cmry athscinat.
- Pneumonia, orga.
- Obs chr bronc w.
- Acute kidney fail.
- Chest pain NEC

Comments
Emerging Big Data Representation Approach – Graphs and Linked Data

Example 1: healthdata.gov, Hospital Compare

An instance of the class
http://health.data.gov/def/hospital/Hospital
(or hosp:Hospital for short), like
http://health.data.gov/id/hospital/393303
returns all the info on the left

Its hosp:site property links to
/id/hospital/393303/site/1

Its org:siteAddress property
links to all other VCard info in
/393303/site/1/vcard/1

Slide courtesy George Thomas, Clinical Quality Linked Data, 2011-06-09
Example 2: Using Network Analysis to Infer Suspicious Activity

(By following a link from a known indicted Medicare provider)

Many different names associated with the ID in a TX-provider file

19 distinct people using the same address and phone in TX provider file.

Exploratory matches by address (practice, corresponding, billing etc.) and phone
Example 3: Semantic Inference from Literature

Population Health Analysis Case Study
Knowledge Discovery Process in Big Data

• Many methods and tools available
  – statistical methods and packages
  – visual analytics tools and frameworks
  – textbooks and courses on data mining

• Most analysts specialize in a small number of techniques

• Few sources offer comprehensive and systematic approach to healthcare knowledge discovery in large scale

What data and what recipe can an analyst use?
Online Public Health Data Sources

CMS has recently announced a new data initiative. This initiative creates a new Office of Information Products and Data Analysis (OIPDA), the goal of which is to maximize CMS data and information products and serve as the single point of accountability and coordination across the implementation of policies related to data release and dissemination. The OIPDA will update the suite of CMS data and information products and serve as the authoritative source for data requests.
About the HIW

Purpose
Access to high quality data improves understanding of a community’s health status and determinants, and facilitates the prioritization of interventions. The purpose of the HIW is to:

- Provide a single, user-friendly, source for national, state, and community health indicators
- Meet needs of multiple population health initiatives
- Facilitate harmonization of indicators across initiatives
- Link indicators with evidence-based interventions
- Serve as the data hub for the HHS Community Health Data Initiative, a flagship HHS open government initiative to release data; encourage innovative application development; and catalyze change to improve community health

Partners
The HIW is a collaboration of many Agencies and Offices within the Department of Health and Human Services. The HIW is maintained by the CDC’s National Center for Health Statistics. Data support and funding are provided by the following:

- Centers for Medicare & Medicaid Services
- Department of Health and Human Services
- Health Resources and Services Administration

HIW v1.0
- Pre-constructed national, state, and local level indicators including:
  - Healthy People 2020 indicators
  - County Health Rankings indicators
  - Community Health Status Indicators (CHSI) indicators
  - Medicare quality and utilization indicators
- Ability to map, chart, graph, and trend indicators (to the extent data allow)
- Links from indicators to evidence-based interventions
- Web services—peer-to-peer business application capability
- Supporting descriptive indicator definitions, methods, data sources and other descriptive data needed to facilitate appropriate use of indicators
Categories of CMS Indicators

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Patient Safety Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Prevalence</td>
<td>Physician Evaluation &amp; Management</td>
</tr>
<tr>
<td>Home Health</td>
<td>Physician Procedures</td>
</tr>
<tr>
<td>Hospital Compare_ Congestive Heart Failure Measures</td>
<td>Prevention Quality Indicators</td>
</tr>
<tr>
<td>Hospital Compare_ Disease Specific Readmissions and Mortality Rate</td>
<td>Readmissions &amp; ED Visits</td>
</tr>
<tr>
<td>Hospital Compare_ Heart Attack Measures</td>
<td>Skilled Nursing Facilities</td>
</tr>
<tr>
<td>Hospital Compare_ Pneumonia Care Measures</td>
<td>Utilization</td>
</tr>
<tr>
<td>Hospital Compare_ Surgical Care Measures</td>
<td>PPS</td>
</tr>
<tr>
<td>Hospital Inpatient</td>
<td>ASC</td>
</tr>
<tr>
<td>Imaging</td>
<td>CAH</td>
</tr>
<tr>
<td>Inpatient Rehabilitation Facility</td>
<td>DME</td>
</tr>
<tr>
<td>Laboratory Tests</td>
<td>Hospice</td>
</tr>
<tr>
<td>Long-Term Care Hospital</td>
<td>Other hospital IP</td>
</tr>
<tr>
<td>Other Tests (Non-laboratory)</td>
<td>Part B</td>
</tr>
<tr>
<td>Outpatient</td>
<td>Post acute care</td>
</tr>
<tr>
<td>Other Tests (Non-laboratory)</td>
<td></td>
</tr>
</tbody>
</table>

CMS Indicators are expressed at State and Hospital Referral Region (HRR) Levels
<table>
<thead>
<tr>
<th>No.</th>
<th>Health Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Breast cancer deaths per 100,000 females</td>
</tr>
<tr>
<td>2</td>
<td>Colorectal cancer deaths per 100,000 pop.</td>
</tr>
<tr>
<td>3</td>
<td>Deaths all causes per 100,000 pop.</td>
</tr>
<tr>
<td>4</td>
<td>Diabetes %</td>
</tr>
<tr>
<td>5</td>
<td>Few fruits vegetables %</td>
</tr>
<tr>
<td>6</td>
<td>Flu vaccination adults %</td>
</tr>
<tr>
<td>7</td>
<td>Heart disease deaths per 100,000 pop.</td>
</tr>
<tr>
<td>8</td>
<td>High blood pressure %</td>
</tr>
<tr>
<td>9</td>
<td>Unintentional Injury deaths No MVA per 100,000 pop.</td>
</tr>
<tr>
<td>10</td>
<td>Lung cancer deaths per 100,000 pop.</td>
</tr>
<tr>
<td>11</td>
<td>Mammogram %</td>
</tr>
<tr>
<td>12</td>
<td>No exercise %</td>
</tr>
<tr>
<td>13</td>
<td>Papsmear %</td>
</tr>
<tr>
<td>14</td>
<td>Physically or mentally unhealthy days %</td>
</tr>
<tr>
<td>15</td>
<td>Pneumococcal vaccination %</td>
</tr>
<tr>
<td>16</td>
<td>Proctoscopy %</td>
</tr>
<tr>
<td>17</td>
<td>Population density per square mile</td>
</tr>
</tbody>
</table>
A Collaborative HIW Analysis Platform

1. Correlation Views

http://cda.ornl.gov/heat/heatmap.html
<table>
<thead>
<tr>
<th>Emergency Department Visits per 1000 Beneficiaries</th>
<th>Heart attack patients given aspirin at hospital arrival</th>
<th>Heart attack patients with aspirin prescribed at hospital discharge</th>
<th>Heart attack patients prescribed angiotensin converting enzyme inhibitor or</th>
<th>Heart attack patients with smoking cessation counseling during hospital stay</th>
<th>Heart attack patients with beta blocker prescribed at hospital discharge</th>
<th>Heart attack patients with fibrinolytic received within 30 minutes of hospital admission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Non-Hispanic White</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analyzing HIW via Partial Correlation

• By adjusting for the effect of High Blood Pressure %, the correlation between No Exercise % Pop. and CHF Admission Rate is reduced from +.622 to +.202.

• What is the true causal model?
Analyst's Workflow

Start

Select indicators → Transform indicators

Generate descriptive statistics

- Univarlate plots?
  - Yes → Display Boxplots
  - No → Display mean structures

- Display heatmap of correlation matrix
- Display Scatterplot Matrix
- Reorder indicators
- Visualize outliers?
  - Yes → Display partial correlations
  - No → Display Bag plots

Winnow indicator list?

Perform model-based analytics

- Too many dimensions?
  - Yes → Display sorted bar-heatmap
  - No → Multivariate outlier analysis

Predictive analytics

- Reduce data to fewer dimensions
  - Yes → Causal analytics
  - No → SEM, confirmatory factor analysis, bayesian networks

PCA, factor analysis, LLE, Isomap, cluster analysis

Multiple regression; supervised learning; support vector machines
Integrating Claims and Clinical Data
Benefit to integrating clinical and claims data

• Can measure quality and cost relationship accurately
  Currently quality is derived from diagnosis codes, recurrence, preventability, etc.

• Longitudinal analysis
  What interventions are working in what delivery system?

• Fraud detection
  Are the clinical procedures necessary and being followed accurately?
Clinical data integration obstacles

• Data formats concerns
  – Unified Medical Language System (UMLS)
  – Systematized NOmenclature of Human and Veterinary MEDicine – Clinical Terms (SNOMED-CT)
  – Logical Observation Identifiers Names and Codes (LOINC)
  – Digital Imaging and Communications in Medicine (DICOM)
  – Health Level 7 (HL7)

• Data coding progression
  – ICD-9, ICD-10 etc. for CPT, HCPCS codes.

• Privacy concerns
Linking Strategies

Example Steps for Linking Inpatient Registry Data to Medicare Inpatient Claims Data

• **Subset registry data**
  - Limit to records for patients aged 65 years or older.

• **Subset Medicare data**
  - Limit to records for patients aged 65 years or older.
  - Limit using broadly defined inclusion criteria to match registry entry criteria.

• **Link hospital identifiers**
  - Link records between data sources on exact values of admission date, discharge date, patient date of birth or age, and patient sex.
  - Use resulting number of matches to inform links.

• **Determine specifications for hospitalization-level linking**
  - Choose rules to apply.
  - Decide whether Medicare records should be allowed to link to multiple registry records.
  - Link hospitalization records

From *Linking Inpatient Clinical Registry Data to Medicare Claims Data Using Indirect Identifiers*, Bradley G. Hammill, MS,a Adrian F. Hernandez, MD, MHS,a,b Eric D. Peterson, MD, MPH,a,b Gregg C. Fonarow, MD, Kevin A. Schulman, MD,a,b and Lesley H. Curtis, PhDa,b; [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732025/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732025/)
Thank you! Questions?