The Importance of (Big) Data for Healthcare Safety-Net Organizations

David Hartzband, D.Sc.
Research Affiliate
Sociotechnical Systems Research Center
Acknowledgements

• Financial & conceptual support from the RCHN Community Health Foundation
  – Special thanks to Feygele Jacobs, President & CEO

• Thanks to Srini Rao, Ph.D. for deployment & analytic support
  – Srini is Founder & CEO of Datycs, a healthcare analytics company
Path2Analytics Project - Goal

- The goal of the P2A Project is to introduce contemporary analytics into CHCs so that this capability may be used in the center’s strategic decision making
  - Contemporary analytics simply means the infrastructure, software & expertise to use new analytic capabilities
  - These capabilities are characterized by analysis of very large data sets or data sets with many different types & formats of data
  - This latter is generally more important in healthcare

© David Hartzband, 2016
P2A – Project Content

• This is accomplished by:
  - Deploying a Hadoop-based, open source analytic stack (Cloudera Express) into the CHC’s data center, &
  - Working with the CHC’s IT & analytic staff so that they understand & can use the deployment
  - Working with the CHC’s Executive Staff so that they understand the difference between the Business Intelligence (BI) & UDS analysis they currently do & the type of analysis made possible by the analytic stack, facilitation strategic inquiry by this staff
P2A – Project Status

• The Project is currently working with 3 large CHCs (2 urban, 1 rural) & a PCA (30 CHCs)
  – ~400 sites, 1,474,267 patients (2 full data years)
• This report will cover three data years (2012-2014) for 30 CHCs, ~300 sites with 1.2M patients
• Results, so far, are similar across all CHCs & sites
PCA – Component Schematic

Application User
EHR, PM...

Analytics User

CHC Infrastructure

Application Data

Analytics Server

Cloudera's Distribution for Hadoop

Analytics Stack

Data Warehouse

© David Hartzband, 2016
Initial Analysis – “Level-Up”

• Designed to determine aspects of data quality & to match the data in the EHR’s underlying data store with the data in the analytic stack (HDFS)

• Purpose is to normalize how data is defined & interpreted between EHR queries & P2A queries (done on analytic stack)

• Required several standard queries
  – Patients/year
  – Patients/diagnosis/year
    • Diagnoses: hypertension, diabetes, obesity, heart disease
    • UDS definitions used for both standard variables (patient, encounter, etc.), HRSA diagnosis lists used for definition of each condition
  – Prevalent comorbidities/population
## PCA – Percentages for Critical Diagnoses

<table>
<thead>
<tr>
<th></th>
<th>PCA</th>
<th>%</th>
<th>P2A</th>
<th>%</th>
<th></th>
<th>%</th>
<th></th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Year</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hypertension</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20.7</td>
<td></td>
<td>22.5</td>
<td>22.8</td>
<td></td>
<td>20.5</td>
<td>23</td>
<td>23.2</td>
<td></td>
</tr>
<tr>
<td><strong>Diabetes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.6</td>
<td></td>
<td>6.6</td>
<td>6.6</td>
<td></td>
<td>6.5</td>
<td>7.3</td>
<td>6.7</td>
<td></td>
</tr>
<tr>
<td><strong>Obesity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td></td>
<td>10</td>
<td>11.7</td>
<td></td>
<td>4.4</td>
<td>10.2</td>
<td>11.8</td>
<td></td>
</tr>
<tr>
<td><strong>Heart Disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td></td>
<td>1.6</td>
<td>1.6</td>
<td></td>
<td>1.4</td>
<td>1.6</td>
<td>1.6</td>
<td></td>
</tr>
</tbody>
</table>

Delta: PCA-P2a in 1% range
PCA-Percentage Comparisons with CDC U.S. Data

Hypertension

<table>
<thead>
<tr>
<th>PCC</th>
<th>P2A</th>
<th>CDC U.S.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>22.2</td>
<td>29</td>
</tr>
</tbody>
</table>

Diabetes

<table>
<thead>
<tr>
<th>PCC</th>
<th>P2A</th>
<th>CDC U.S.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>6.8</td>
<td>9</td>
</tr>
</tbody>
</table>

Obesity

<table>
<thead>
<tr>
<th>PCC</th>
<th>P2A</th>
<th>CDC U.S.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.7</td>
<td>8.8</td>
<td>35</td>
</tr>
</tbody>
</table>

Heart Disease

<table>
<thead>
<tr>
<th>PCC</th>
<th>P2A</th>
<th>CDC U.S.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>1.5</td>
<td>11</td>
</tr>
</tbody>
</table>

~+400%

~+700%

*U.S. cumulative % from CDC FastStat

© David Hartzband, 2016
PCA – Comorbidity Analysis 2014

Please Note:
Low % figures for Obesity & Heart Disease make comorbidity figures inaccurate… although %s may be indicative...

© David Hartzband, 2016
PCA – Data Issues 1

• Initially PCA & P2A patient numbers & percentage figures were 10%-30% off
  – PCA was using many different definitions for patient, more than half of the participating health centers were using non-UDS definitions
  – The same was true of diagnoses, the PCA was using non-UDS definitions for the ICD-9 codes for diagnoses
  – All definitions were adjusted to be compliant with UDS (HRSA) definitions

• Obesity & Heart Disease population percentages were 400% & 700% respectively below CDC figures for U.S. population
  – We expect health center populations to be “sicker” than the general population
PCA – Data Issues 2

• We also found that the EHR in use (NextGen) treated query specifications differently
  • e.g. 250-ICD-9 general diabetes code:
    • 250, 250.0, 250.00, 250* produced different results, as did 250.5 & 250.50
    • This meant that we often had to compare SQL code to determine how a diagnosis was specified in order to match results
• The PCA had used a contractor to design & deploy a data warehouse which supplied the data for both the PCA query & the analytic stack extract to HDFS
  • The schema of this warehouse was so complex that very complex queries had to be written to specify relatively simple results
    • This increased the possibility of errors in coding
Interpretation – EHR Structure & Function

• Navigation in NextGen for anything but the simplest workflows is difficult enough to discourage use
  – Acute workflows not conducive to treatment of multiple conditions

• Little support for multiple diagnoses/encounter
  – Use case informally tested at HIMSS15 with 6 EHR vendors – none provided adequate support

• Many deleterious effects of migration from one EHR to another & unsuccessful integration in general
Interpretation – Population Percentages

• Low obesity diagnoses discussed with many CMOs & CHC staff
  – Most agree that this is not surprising, although most CMOs thought their population was in the range of 40-45%
  – Many CMOs say their providers do not use the full range of UDS specified ICD-9 codes (V-codes for specification of overweight, obese etc.)
  – Many CMOs, providers & staff members believe that there is a social bias against making this diagnosis

• Heart Disease percentages (<5% when the U.S. population mean is 11%) are harder to interpret
  – Most CMOs felt that their population might be in the range of 20%-30%
  – Cause still under investigation
A recent paper in PLoS ONE* compared ICD-9 data reported in the U.S. Inpatient Reporting Sample (NIS, HRSA) to interview data reported in the Behavioral Risk Factor Surveillance System (BRFSS, CDC) for 2011 data & found that in the NIS data, that is hospital discharge data, the percentage of obesity reported was ~9%, & that in the BRFSS or interview data, the percentage of obesity was 27%. This is a significant difference between the recorded data & the observed data.

Diagnosis Errors 1

• The Institute of Medicine released a report in September 2015 titled *Improving Diagnosis in Healthcare*
  – The report stated that every person involved in our healthcare system would receive at least one error in diagnosis in their lifetime
  – Causes were listed as:
    • inadequate collaboration & communication
    • System & workflows not well designed to facilitate the diagnosis process
    • Limited feedback on diagnosis, &
    • A medical culture which is punitive & discourages transparency & disclosure of diagnostic errors
Diagnosis Errors 2

• To these I would add:
  – A reimbursement system that still emphasizes pay-for-procedure & does not facilitate multiple diagnoses or preventive efforts
  – EHRs & other HIT systems not designed to align with providers actual workflow & that do not facilitate recording of multiple diagnosis
  – A clinical environment strongly influenced by regulatory requirements (e.g. meaningful use evolving to MIPS, Merit-Based Incentive Payment System)
Error Types…

• **Normalization – semantic & otherwise**
  – Results are very sensitive to initial definitions of core terms such as: patient, provider, encounter, location &c.

• **Straight errors**
  – Omission: in data entry or through data import, often clinical data is entered into notes fields which are not searchable (import from 3rd party system that are not fully integrated)
  – Commission: in data entry through lack of knowledge or time. e.g. BMI=900 etc.

• **EHR structure/function**
  – Issues with workflow alignment, complexity, data interpretation already discussed

• **Sociocultural & organizational issues**
  – Many CMOs felt that providers were less likely to record a diagnosis of obesity in certain populations
  – Many providers felt that it was more important to provide adequate notes for the next provider to review than to provide searchable clinical data
  – Many organizations had no emphasis on data awareness or “data as an asset” so data issues were not a priority
Big Data?

• “Big” data not necessary to do analytics!
  – Many strategic questions can be addressed with the data at hand
    • Each CHC we worked with has 5-10 GBs of EHR & up to 5 GBs of other data (financial, cost accounting, registry, pharmacy etc.)
    • PCA has ~250GBs of EHR data & 100GBs of other data (as above)
    • Each expected to double in the next 2-3 years plus need to include an additional 20-25GBs of external data (public health, macrodemographic (State & Federal), macroeconomic (State & Federal) data
  – Largest healthcare organizations already have ~50 PBs of data (Kaiser)

• Asking the right questions is more important than having the most data!
Right Questions?

• How do we ask the “right” questions?
  – Being “data aware” allows us to focus on strategy rather than analysis
  – Need to know:
    • What data is available (not just EHR, PM)
    • Quality of data & how that may limit use
    • Types of analysis available (including modeling & prediction)

• Involving people from multiple parts of the CHC is important in developing analyses: Executive, Admin, Clinical all must be represented

• Essential to understand the culture of providing care in developing questions & interpreting analytic results (e.g. see Obesity results)
Change is Good, Right?...

- The information management environment for CHCs (& healthcare organizations in general) is changing
- In the next 3-5 years:
  - CHCs will have double to triple the amount of data they current use including considerable external data from many sources
  - New organizational models (HIEs, ACOs…) new regulatory requirements (Meaningful Use 3) will require CHCs to make much more use of all of this data to be able to provide care & operate in the new environment
  - Technology is changing quickly & current best practices for information storage, management analysis & usage will be obsolete shortly (2-3 years)
- Beginning the exploration of analytics will position the health center to be able to evolve with the environment
What is to be Done?*

• Healthcare organizations, of all sizes & capabilities, must develop & emphasize the idea of data awareness so that the quality of data, clinical & otherwise, that is captured, analyzed & used for decision making is given the highest priority

• Simplified design of data storage & analytic systems must also be emphasized

• Organizations should work with vendors to ensure that the HIT they use is as well aligned with their work process as possible & that the data that is captured & available for analysis is correct & useful

• Making analytics a core part of decision making will both start the technology evolution process & ads well as develop the data awareness necessary for the organization to be successful in the future

*With apologies to V.I. Lenin
David Hartzband, D.Sc.

Research Affiliate
Sociotechnical Systems Research Center

dhartz@mit.edu
617.501.4611

Blog: http://posttechnical.blogspot.com/