Using a Systems Approach to Developing Analytics Capability in the Healthcare Safety Net

David Hartzband. D.Sc.
Engineering Systems Division
Path2Analytics: What is It?

- Path2Analytics is a project developed by the RCHN Community Health Foundation with the goal of working with community health centers to integrate the capacity to productively use contemporary analytics capabilities into the work of the health center.

- The project will work with both rural & urban health centers to meet this goal:
  - The first two centers have been selected & the project is in progress at both.
  - The project will consist of both a process & a set of software to assist in meeting this goal.
FQHCs

• Health centers started operation under federal grants in 1975
  – Funding consolidated under Section 330 (Public Health Act) in 1996 (42 USC §254b Section 330)

• 1131 FQHCs (2012) with ~6700 sites
  – ~100 “Look-Alikes” with ~ 500 sites

• Serve mainly uninsured & disadvantaged populations
  – Mainly Medicaid & uninsured
  – Served ~23M patients in 2012
Path2Analytics: Assumptions 1

• Much of current healthcare is increasingly an information management effort.

• Analytics (i.e. the systematic analysis of data focused on answering a specific question or set of questions) must be an integral part of a healthcare organization’s administrative & clinical plans.

• Analytics is not an IT function, a software package or a technical methodology.
  – It is a way of thinking about an organization’s goals that makes use of IT, analytic (software) packages & technical methodologies.
• Analytics is not necessarily about “Big Data”.
  – But it must be about “right data”
  – It can be very effective making use of “little data”, that is the data that an organization already has.
  – Not all organizations have multiple petabytes of data, but even multiple gigabytes of data can produce productive results.

• An analysis is only as good as:
  – The quality of the data used, &
  – The quality of the questions asked.

• Questions asked are only as good as their alignment with:
  – The quality & content of the available data, &
  – The strategy & goals of the organization.
Path2Analytics: Process

- The project consists of five units:
  - Landscape – what are other healthcare organizations doing with respect to analytics? how does this relate to CHCs?
  - Analysis Planning – what resources are available? what questions to ask? how to address them analytically? what tools to use, setting goals & metrics
  - Data Organization – what data is available, extraction & aggregation, ETL & normalization
  - Analysis & Interpretation – operation of analytic process
  - Outcome & Outreach – how to organize & convey results, who to tell, communication process etc.
LANDSCAPE
Landscape - 1

• Data Warehousing
  – Almost all projects require some data extraction or warehousing

• Point-of-Care Recommendations
  – Mayo Clinic - Large-scale warehousing: semantic normalization, data dictionary, 5M clinical records covering 15 years, 15-25PB of data (2500x the size of the content of the Library of Congress)
    • AWARE “bedside consulting”: delivering diagnosis & best-practice treatment for individual patients based on analysis of anonymized data set
  – BIDMC - Clinical Query: system used directly by providers to analyze 2.2M clinical records to determine best-practice treatment in real-time
  – Kaiser - Clinical, pharma & lab data on 9.1M patients over 10 years
    – Natural Language query system allows providers to get recommendations for best-practice treatment
  – Partners – combined clinical, ops, financial for real-time PoC data & best practice recommendations (Queriable Inference Patient Dossier)
• Outcome & Population Characterizations
  – Intermountain (Deloitte as partner) – 90M patient records, Outcome Miner = factors affecting outcome, Population Miner = relationship between treatment & outcomes
  – McKinsey (BeyondCore) – “next 5% analysis”, 30M claims characterize “microsegments for next 5% of patients wrt cost, assign care managers

• Clinical & Financial Optimization
  – Many organizations are aggregating clinical & financial data to be able to do analyses such as:
    • Cost/service (/provider, /location)
    • Cost/outcome (/provider, /location)
    • Cost trends
Landscape 3

• Best Practice Guidance
  – Geisinger – ProvenCare: point-of-care system that interactively provides best practice elements for specific treatment areas (congestive heart failure, strike, etc.)

• Prediction
  – Express Scripts – 1.5B prescriptions/year, modeling to predict which patients most likely to modify drug use behavior, proactive intervention

• Research
  – Mt Sinai Medical Center (Ayasdi) – topological data analysis, analysis of entire e. coli (1M DNA variants) to determine bacteria’s response to antibiotics
Landscape 3 - What’s Missing

• Operations? – very few projects identified that were strictly non-financial optimizations, opportunities for:
  – Schedule optimization including analysis of no-shows
  – Patient flow for reduced waiting
  – Readmission rates – analysis of multiple dimensions: demographic criteria, location, diagnosis etc.
  – Predictive modeling of changes in:
    • Patient number & type
    • Service utilization
    • Locational utilization

• Others?
ANALYSIS PLANNING
Types of Analysis

- **Visualization** - Use of visualization tools to characterize relationships
- **DB Query (SQL)** – Conventional query of relational databases
- **MapReduce** - Analytic framework using large-scale distributed cluster(s) for storage & processing of ultra-large data sets (Hadoop)
- **Algorithmic** – Use of programming languages such as R, Python, Pig Latin… to create algorithms for custom analysis (often used with Hadoop)

- All but algorithmic used in project
This is Different!

• It must be emphasized that this is not statistical analysis in the classical sense.
  – It’s also not “report generation” as is required for HRSA, etc.
• It is the empirical characterization of data in the context of specific inquiries
• If a Point-of-Care recommendation system finds 9,274 cases that match a specific set of patient parameters & classifies the outcomes according to treatment plan, that is not a statistical prediction of the best treatment plan. It is an exact characterization based on the data!
  – When we say that a specific treatment plan was associated with a specific outcome in 72% of the cases, this is not a probability, but an exact proportion!
  – The provider may choose not to follow this characterization
It Must Also Be Emphasized...

- Project goal is to produce initial analyses & leave health center with hardware & software that can be used by their Staff to do additional analytic inquiry, BUT...
- Perhaps a more important goal is to introduce the concept of “data as an asset” & to embed this through discussion & example so that data-driven decision making becomes a tool that the health center can routinely use.
- This is a way of thinking as much as anything
Analysis Planning

- Identification of areas of inquiry & actual query building most effectively approached as a systems problem
- Inquiry must be aligned with health center strategy in order to be relevant
- System includes not just clinical & financial health center data but also at least: local & regional demographics, social & cultural determinants, local & regional economics, regulatory environment...
- Ontology (Owl) developed & used to guide discussions of inquiry planning
- Discussions held with Estaff, Clinical & Admion/Support Staffs at health center(s)
P2A Ontology (native org.) - OWL 4.1.2

This work was conducted using the Protégé resource which is supported by grant GM10331601 from the National Institute of General Medical Sciences, NIH
This work was conducted using the Protégé resource which is supported by grant GM10331601 from the National Institute of General Medical Sciences, NIH
Model for Developing Inquiry

Discussion of Health Center Strategy

Discussion of Analytic Ontology

Initial Inquiry Topics

Refinement of Strategic Criteria

Evaluation of Existing Data

Data Normalization & Quality Checks

Review of External Data & other Criteria

Refinement of Inquiry Topics

Final Inquiry Topics

Summary Review

Review of

HealthcareAnalyticsintheSafetyNet
Use of Systems View

• The ontology was used as a guide to have Health Center participants focus on larger scale & strategic issues for inquiry
  – Prior to use of ontology, most ideas about analysis were highly specific & tactical; related to current reporting issues
  – After the ontology was introduced & the larger-scale system was discussed, health center participants were more likely to focus on strategic issues for inquiry
Inquiry Topics CHC 1

• Cost/visit increasing but revenue declining in dental practice?
  – Characterize trends in both cost/visit & revenue, identify additional expense in dental practice (by location, by provider), compare with trends in medical practice

• Overall patient encounters declining? Are our patients getting care elsewhere?
  – Use external (State, County) data to correlate demographic trends by location with encounter data, Use data from area hospitals & other clinics to correlate with patient care outside of of CHC
    • Just acquired State hospital admittance/discharge detail data

• What patient populations are more “non-compliant” in managing their health & healthcare?
  – Characterize patterns in declining clinical outcomes, determine if there is a pattern wrt diagnosis, location, demographic data
Demographics for three counties from State Records - Encounters in Counties 1 & 2 have declined at ~12% rate, but have increased ~57% in County 3.

Used as an example to CHC Staff for the relevance of external data.
Inquiry Topics CHC 2

• Service utilization, especially enabling services,... by location
  – Analysis of enabling services provided by cost, location, outcome (especially interested in interpretation)

• Overall number of patients steady, but 200-300 new patients per month
  – Characterization of: single visits, episodic visits

• Workflow bottlenecks?
  – Analysis of duration of encounter by treatment & diagnosis codes, patterns in duration? By provider? By location?
DATA & SOFTWARE DEPLOYMENT & ORGANIZATION
Need for Analytic Stack (Hadoop,...)

• There are three characteristics of data that may require the use of a contemporary analytic approach – the three V’s:
  – Volume – Some organizations actually do have petabytes (1024 terabytes) of data... such ultra-large data sets are not easily managed or analyzed by conventional means
  – Variety – Many of the types of inquiries that are relevant to current decision making require that data from multiple different sources & in different formats be analyzed together... *e.g.* clinical & financial data
  – Velocity – particularly data from sensor nets (NASA is the best example)... 1,000,000 sensors sampled 3x/second

• In healthcare, the issues usually is variety...
2013 Poll - HealthCatalyst

- At what level does your organization consistently and reliably function?
- 305 Respondents

% Function per Maturity Level

| Level 8 | Personalized Medicine & Prescriptive Analytics |
| Level 7 | Clinical Risk Intervention & Predictive Analytics |
| Level 6 | Population Health Management & Suggestive Analytics |
| Level 5 | Waste & Care Variability Reduction |
| Level 4 | Automated External Reporting |
| Level 3 | Automated Internal Reporting |
| Level 2 | Standardized Vocabulary & Patient Registries |
| Level 1 | Enterprise Data Warehouse |
| Level 0 | Fragmented Point Solutions |

MIT ESD
Engineering Systems Division
Proposed Technical Approach

• Preliminary characterization of data using existing reporting/BI system
  – Can be used to determine if normalization/ETL is needed

• Use of existing data extracts/warehouse if available
  – Analysis of data quality needs to be done

• Layering of existing data into open source analytic stack (Hadoop-based)
  – Data from multiple sources & of different types can be loaded
  – Normalization can be done entirely in Hadoop

• Use of Yarn/MapReduce to generate results, &/or

• Use of Hbase for data management

• Use of Hive/Impala to do query

• Use of Tableau to visualize raw data
Deployed Analytic Stack

- **Tableau**
- **Cluster access:** Hue
- **Workflow:** Oozie
- **Inquiry:** Impala (SQL)
- **Search:** Soir

- **Zookeeper**

- **HBase**
- **YARN**
- **HDFS**

- Visualization
- Group Services & Synch
- Map Reduce 2
- Dist. File System
Cloudera Manager running at CHC1
Hbase Table Browser running at CHC1

<table>
<thead>
<tr>
<th>row_key, row_prefix</th>
<th>scan_len</th>
<th>col1, family:col2, fam3</th>
<th>col4</th>
<th>col5</th>
<th>col6</th>
<th>col7</th>
<th>col8</th>
<th>col9</th>
<th>col10</th>
</tr>
</thead>
<tbody>
<tr>
<td>000F2371-742C-4CS3-B730-D329A763251</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>location_mstr</td>
<td>loc_tax_id</td>
<td>location_mstr</td>
<td>fac_entity_id</td>
<td>location_mstr</td>
<td>clia_id</td>
<td>location_mstr</td>
<td>optical_location_ind</td>
<td>location_mstr</td>
<td>tax_rate_library_id</td>
</tr>
<tr>
<td>77</td>
<td>22D0069727</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>row_key, row_prefix</th>
<th>scan_len</th>
<th>col1, family:col2, fam3</th>
<th>col4</th>
<th>col5</th>
<th>col6</th>
<th>col7</th>
<th>col8</th>
<th>col9</th>
<th>col10</th>
</tr>
</thead>
<tbody>
<tr>
<td>01020D82-E9F3-4FCA-308B-819929D1E152E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>location_mstr</td>
<td>loc_tax_id</td>
<td>location_mstr</td>
<td>fac_entity_id</td>
<td>location_mstr</td>
<td>clia_id</td>
<td>location_mstr</td>
<td>optical_location_ind</td>
<td>location_mstr</td>
<td>tax_rate_library_id</td>
</tr>
<tr>
<td>77</td>
<td>22D0069727</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fetched 10 entries starting from null in 2.511 seconds.
Hive Query running at CHC1
Next Steps - 1

• Once a goal & rationale for an inquiry is decided on, several steps have to take place before analysis can be performed:
  – Preliminary queries/explorations have to be decided on & developed
    • Generally done by consensus with stakeholders
    • Quality & normalization issues should be done in the context of preliminary queries so that the entire universe of data does not need to be addressed (infeasible to impossible)
  – Data issues & Data quality need to be assessed
  – Semantic normalization needs to be addressed
Next Steps – 2: Data Quality

• The issues of data quality with healthcare information, especially EHR data are well known

• Previous work with the data warehouse of a PCA & several associated CHCs showed that:
  – Usage patterns both of different providers & different CHCs created quality issues
  – Many Providers use the EHR (regardless of which one) in idiosyncratic ways (filling in Notes, but not on forms, etc.) that affect the ability use the data in standardized ways
  – Many CHCs have idiosyncratic coding practices that make comparisons across health centers infeasible or impossible
Next Steps – 3: Normalization

• Semantic normalization is rarely addressed
  – e.g. Who is a provider? MDs?, D.Psy.? D.NPs? ANPs? RNs? Social Workers? Some of these sometimes?
  – What is a service?
  – What is an encounter?
  – What is an outcome?

• A recent exercise at a medium-size, urban CHC produced the following definitions:
  – Provider = anyone with a license to access & modify the EHR
  – Service = Medical, Dental, Behavioral Health, Medical Specialities, Social Services
  – Encounter = a checked-in visit with an encounter number (exclusions by patient type)
  – Outcome – deferred – a (semi)recent ONC working group agreed unanimously on one outcome... mortality!
ANALYSIS & INTERPRETATION
OUTCOME & OUTREACH

both in progress
Initial Lessons Learned

- Semantic & Syntactic normalization of data essential for analysis
  - Even if normalization is done at time of analysis (as in Hadoop), there must be standards for the data to be normalized against

- Deployment of Analytic Stack takes time & extreme attention to detail

- Substantial time is needed for selection of analysis areas & inquiry development

- Large opportunity for operational optimization if data is available
Takeaways

• This technology is VERY powerful, but it CANNOT solve a health center’s issues by itself
• Inquiry is empty unless it is closely aligned with the needs & strategy of the health center
• Projects like this are not IT projects – they require input & participation from all areas of the health center
• Knowing your “data” is key... not just the EHR data, but all of the data that is critical to the health center
  – Focus on the idea of “data as an asset”,.. All data
• Perspective is important – No analysis is a panacea