CohortGen: Promoting Collaboration Through Automatic Cohort Generation
Billy Barnett MA
Charisse Madlock-Brown PhD
Have an EHR warehouse...what next?

• What is a Data Warehouse?
  – An integrated, granular, historical single point of reference for data
  – Data warehousing in healthcare
    • Millions of patient records collected
    • Billions of data points
  – Extracting data is challenging
    • Requires special skill
    • Many researchers are unclear on what is available
Promoting Research And Collaboration

• Need for cohort generation
  – Challenges associated with promoting collaboration through profiling systems
  – Difficulty identifying meaningful subsets from Big Data EHR data warehouses
Primary Data Source

• Health Facts Data Warehouse
  – Started in 2001 by Cerner
    • Over 70 million patient records
  – Types of data

  o Demographics  o Diagnoses
  o Prescriptions  o Procedures
  o Laboratory tests  o Clinic locations
  o Billing  o Hospitals
Elsevier Developer API

• REST API to retrieve research article metadata:
  – Title
  – Authors and affiliations
  – Abstract
  – Keywords

• With a developer key, we were able to get all UTHSC articles
cTakes

- Clinical Text Analysis and Knowledge Extraction System (cTAKES)
- Apache NLP system for extracting information from EMR
- Can be used to tag documents with UMLS concepts
Other Data Sources

• Unified Medical Language System (UMLS)
  – Produced by the National Library of Medicine
  – Metathesaurus database
  – Useful for informatics research

• Faculty extract from human resources
Sample Navigation of UTHSC Areas of Excellence
Finding researchers
Researcher Profile

• Sunburst chart summary of research areas
HEALTH Facts information for Area

- Gender breakdown summary
Health Facts information for Area

- Age group summary

<table>
<thead>
<tr>
<th>Age Groups</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 10</td>
<td>182409</td>
</tr>
<tr>
<td>0 - 100</td>
<td>358408</td>
</tr>
<tr>
<td>0 - 20</td>
<td>264158</td>
</tr>
<tr>
<td>0 - 30</td>
<td>414029</td>
</tr>
<tr>
<td>0 - 40</td>
<td>581594</td>
</tr>
<tr>
<td>0 - 50</td>
<td>1220720</td>
</tr>
<tr>
<td>0 - 60</td>
<td>2392245</td>
</tr>
<tr>
<td>0 - 70</td>
<td>2526285</td>
</tr>
<tr>
<td>0 - 80</td>
<td>1915430</td>
</tr>
<tr>
<td>0 - 90</td>
<td>957486</td>
</tr>
</tbody>
</table>
Health Facts information for Area

• Race breakdown summary
Small User Review

- Group of three researchers found our system helpful
- Need for search option
  - We have created solr indexes for scholarly articles and diagnostic codes.
- Want to view articles in system
- Not interested in drill-down view
- Want collaborator suggestions when reaching Health Facts page
- Need emphasis on current research
CohortGen Fundamentals

• Generating personalized cohorts based on automatically identified faculty interest
• Data visualization of aggregated data gives users simple view of cohorts
• Research data can be linked to team opportunities
References
