BioLINCC and the NHLBI’s Vision for Data Indexing

Sean Coady, Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute

Cross-Cohort Collaboration Consortium Investigator Meeting
March 24, 2018
Background


- BioLINCC established in 2009 to coordinate activities of Data Repository and Biorepository
Eligible Studies

Observational studies and clinical trials funded by NHLBI

- Supported by Contract (all)

- Supported by grant
  - > $500K and of high programmatic interest
  - Ancillary studies conducted on a majority of participants in an eligible parent study

- Eligible data consists of all de-identified laboratory, procedural data (exercise test, Echocardiography, Computed Tomography, etc.), questionnaires and outcome data.

- All appropriate documentation (protocols, annotated forms, manuals of procedures, data dictionaries, algorithms for calculated variables, etc.)
Timing for Release

- **Clinical Trials:**
  - ✓ 3 years after the final visit of the participants to their clinical trial sites, or
  - ✓ 2 years after publication of primary outcome paper

- **Observational Studies:**
  - ✓ 3 years after the completion of each examination or follow-up cycle, or
  - ✓ 2 years after the baseline, follow-up, genetic, ancillary study, or other data set is finalized within the study for analysis

Intent: study investigators receive a 2 year protected period of time
Current portfolio: 43 Observational studies (342,000 participants), 126 trials (417,000 participants)
Cumulative access requests for trial and observational study data (thru Mar 15, 2018)
Annual average number of data access requests for study databases per study per year available (3 year rates)

Average number of data access requests per study per year

- General Population OS
- Primary/Sec. Prevention CT
- Heart Failure CT
- Race/Ethnic Population OS
- Respiratory Distress CT
- Emergency Medicine CT
- Other CT
- Other OS
- Cardiovascular CT
- Disorder Population OS
- Sickle Cell CT
- Other Lung CT
- Asthma CT
- Transplant/Transfusion CT

2012-15: [chart data]
2015-18: [chart data]
Data Access Requests: Percent Requesting Single Study or 2+ Studies, 2012-Present
Primary Reason for Access Request for a Single Study (top panel) or Multiple Studies (bottom panel)

Percent of Access Request for 1 Study

Percent of Access Request for 2+ Studies

- New Question
- Meta-analysis
- Statistical Methods
- Clinical Trial Methods
- Clinical Prediction/Risk Prediction
- Comparison Group
- Other (Pilot, Simulation)

Years: 2012 to 2018
Need for accessible, searchable data index

- Currently repository has 169 clinical studies across the spectrum HLBS phenotypes, adding approximately 20+ new studies each year.

- Approximately 1/3 of data access requests involve more than one study.

Current search is “study” based.
Pilot study

- “Come as you are” repository limits resources for indexing

- Pilot Index using 11 studies: ARIC, CARDIA, CHS, Framingham, MESA, HCHS-SOL, AVID, DASH, ROC-PRIMED, ROMICAT II, and TAAG.

- Covers 1,277 datasets and 113,437 variables
Pilot results – search “sex” / “gender”

- Basic study information
- Basic data element info
- Data distribution info
Pilot results – “sex” / ”gender”

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>study_filter2</td>
</tr>
<tr>
<td>acronym</td>
</tr>
<tr>
<td>study_name_noacronym</td>
</tr>
<tr>
<td>current_accession</td>
</tr>
<tr>
<td>specimens</td>
</tr>
<tr>
<td>name</td>
</tr>
<tr>
<td>label</td>
</tr>
<tr>
<td>dataset</td>
</tr>
<tr>
<td>folder</td>
</tr>
<tr>
<td>levels</td>
</tr>
<tr>
<td>n</td>
</tr>
<tr>
<td>mean</td>
</tr>
<tr>
<td>std</td>
</tr>
<tr>
<td>median</td>
</tr>
</tbody>
</table>

*S denotes suppressed field due to small counts*
Pilot results – potential filters

<table>
<thead>
<tr>
<th>Description</th>
<th>StudyCounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular Trial</td>
<td></td>
</tr>
<tr>
<td>AVID</td>
<td>39</td>
</tr>
<tr>
<td>General Population</td>
<td></td>
</tr>
<tr>
<td>ARIC</td>
<td>19</td>
</tr>
<tr>
<td>CARDIA</td>
<td>47</td>
</tr>
<tr>
<td>CHS</td>
<td>11</td>
</tr>
<tr>
<td>FHS-Cohort</td>
<td>4</td>
</tr>
<tr>
<td>FHS-Gen III</td>
<td>8</td>
</tr>
<tr>
<td>FHS-OS</td>
<td>8</td>
</tr>
<tr>
<td>MESA</td>
<td>13</td>
</tr>
<tr>
<td>Other Trial</td>
<td></td>
</tr>
<tr>
<td>DASH</td>
<td>1</td>
</tr>
<tr>
<td>ROMICAT-II</td>
<td>2</td>
</tr>
<tr>
<td>TAAG</td>
<td>24</td>
</tr>
<tr>
<td>Race/Ethnic Specific Population</td>
<td></td>
</tr>
<tr>
<td>HCHS-SOL</td>
<td>39</td>
</tr>
</tbody>
</table>

Search on “sex” or “gender” yielded 215 hits

Filters based on domain or study name relatively easy to element
Pilot results - “FEV” / ”spirometry” search

Search on “fev” or “spirometry” yielded 303 hits

Other terms: forced expiratory, pulmonary function, pft, lung function, vital capacity
To be truly usable, need filters based on mapping to an ontology applicable to clinical research.

Hierarchy would need to be narrow enough to be useful, but not so narrow as to overwhelm.
Summary

- Clear need to develop an accessible, searchable and usable index to enable discovery of available data elements
- Feasible to create searchable index using straightforward programming
- To be usable, need mapping to an ontology designed for clinical research data