Data Harmonization & Data Sharing

Cross-Cohort Collaboration
March 5, 2016
Questions to Address

• What is the objective of data harmonization and sharing for the CCC?
• What are the steps involved in data harmonization?
• What are the resources required for successful data harmonization?
• What resources are already available?
• What type of data access and control is desirable?
• What ought to be the next steps in data harmonization for the CCC?
  • Short term
  • Long term
What is the Objective of Data Harmonization & Sharing for the CCC?
Objective of Data Harmonization & Sharing

- Facilitate opportunities to address questions that require larger sample sizes, and more diverse populations and/or expertise
  - E.g., studies of rare outcomes, complex interactions, and mediations

- Using inferentially similar, high quality data for valid synthesis
What Are the Steps Involved in Data Harmonization?
Steps in Data Harmonization

• Cataloguing the characteristics of each cohort study (e.g. design, sample size) in a systematic way.

• Defining the ‘target variables’ - set of variables to be generated from the harmonization process.
  • ‘target’ variables determine the data information content that is required from each study to generate compatible (i.e. harmonized) variables.
Steps in Data Harmonization

• Each cohort determining its potential for generating ‘target variables’
  • Each cohort evaluates study-specific questionnaires, standard operating procedures and data dictionaries, the potential for each cohort study to generate the target variables can be determined.

• Each cohort working with the data to transform their data locally into a common harmonized format.
  • Use standard algorithms
What Are the Resources Required for Data Harmonization?
## Resources Required for Data Harmonization

<table>
<thead>
<tr>
<th>Step/Process</th>
<th>Resources Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cataloguing characteristic of each cohort</td>
<td>Teams to develop standard data collection tools, complete data forms, and catalogue descriptive information</td>
</tr>
<tr>
<td>Defining ‘target’ variables (exposures, outcomes, covariates) and quality criteria</td>
<td>Team of researchers knowledgeable about data needs and their cohorts; workshops/meetings; reports</td>
</tr>
<tr>
<td>Determining the potential for generating target variables</td>
<td>Team within each cohort to review data collection tools, SOP, data dictionaries, and qa/qc documentation</td>
</tr>
<tr>
<td>Transforming cohort data into a harmonized dataset</td>
<td>Algorithms, team of information scientists within each cohort</td>
</tr>
</tbody>
</table>
Resources Required for Harmonization

- Time
- Provisions for updating data
- Resource management / coordination
  - Retrospective harmonization
  - Prospective harmonization
What Resources are Available?
Resources Available for Data Harmonization

- Experience and data from other consortia
  - CHARGE
  - The Cardiovascular Disease Lifetime Risk Pooling Project
  - Epidemiology of Non-fracture Fall Injuries: Cross-cohort Study of Medicare Outcomes
  - Duke BioLINCC Data Harmonization Project
  - TOPMed

- International Harmonization Initiative
International Harmonization Initiative

• Schema for selection of variables and classification of quality of harmonization

• Web-based platform for harmonization
  • https://www.maelstrom-research.org
  • http://www.obiba.org
What Type of Data Access & Control is Desirable?
Types of Data Access & Control

- Centralized / data commons
- Federated infrastructure
Centralized / Data Commons Questions?

- Where would the Coordinating Center for data harmonization be located?
- What management infrastructure would be required?
- What resources are there to support such infrastructure?
- What kind of data sharing procedures would be needed?
- What would be the roles and responsibilities of the publications committee? time and effort required?
- What are the ethical-legal considerations for each cohort / for the CC?
  - Who would be held accountable in case of a data breach / hacking?
- Would there be a common data use agreement?
- How many cohorts would be included/excluded?
- Is this feasible in the short term?
  - Sometimes - Cardiovascular Disease Lifetime Risk Pooling Project
Federated Infrastructure Questions

- Data access and control would remain with each cohort after endorsement of the concept of the project by the CCC Executive Committee.
- Data sharing procedures would be those of each cohort / in place.
- Likely to be more feasible in the short term.
Data harmonization and federated infrastructure for three Healthy Obesity Project studies

Doiron et al. Emerging Themes in Epidemiology 2013, 10:12
http://www.ete-online.com/content/10/1/12
What ought to be the next steps in promoting / supporting data harmonization in the CCC?

• Short term

• Long term
References

