Big Data to Knowledge

Unmet Opportunities and Challenges Afforded by Big Data

ECHO

Environmental influences on Child Health Outcomes

Matthew W. Gillman, MD, SM
30 November 2016
ECHO Mission

• Enhance the health of our nation’s children for generations to come
ECHO Overall Scientific Goal

• Understand effects of early environmental exposures on child health and development
ECHO Overall Scientific Goal

• Understand effects of early environmental exposures on child health and development
  – Effects: Observation & intervention
ECHO Overall Scientific Goal

• Understand **effects** of early environmental exposures on child health and development
  – Effects: Observation
    • Nationwide consortium of existing cohort studies
      – 2-year feasibility/pilot phase
      – If pass milestones, 5-year follow-on phase
ECHO—Cohorts
35 awards, 74 PIs, 84 cohorts

Cohort Prime Awardees
Cohort Sub-Awardee
ECHO—Cohorts
35 awards, 74 PIs, 84 cohorts

Current:
~33,000 mothers
~46,000 children
ECHO-wide (aka “synthetic”) cohort

- >50,000 children
- Data platform for multiple cohorts to conduct solution-oriented observational studies
- Multiple cohorts maximize
  - Sample size
  - Heterogeneity/diversity
  - Generalizability
Big Data Challenges in ECHO

- Sharing
- Stewardship
- Analysis

- Many of these within BD2K purview
Big Data Challenges in ECHO
Data Sharing

• Among investigators
• For public use
• [With individual participants]
Big Data Challenges in ECHO
Data Sharing

• Who
  – Investigator wariness
Big Data Challenges in ECHO
Data Sharing

- **Who**
  - Investigator mistrust

- **How & when**
  - **Sequence**
    - Aggregate level analyses using distributed data approach (“send programs to data”)—year 01
    - Submit existing individual-level data to Data Analysis Center—years 01-02
    - Submit newly collected data—years 02 and beyond
Big Data Challenges in ECHO
Data Sharing

• Who
• How & when
• What
  – Longitudinal
    • Many touches per individual
Big Data Challenges in ECHO Data Sharing

• **Who**

• **How & when**

• **What**
  – Longitudinal
  – Data sources
    • Primary
      – Interviews
      – Questionnaires
      – Examinations
      – Biospecimens & environmental specimens
      – Medical imaging
      – Wearable sensors
    • Secondary
      – Electronic medical records
      – Vital statistics
      – Geospatial data
Big Data Challenges in ECHO Data Sharing

• Who
• How & when
• What
  – Longitudinal
  – Data sources
  – Varying depths of phenotyping
Big Data Challenges in ECHO
Data Stewardship

• Who
  – Data Analysis Center for now
  – Longer term?

• When
  – How long?

• What
  – Metadata

• How
  – Resources
Big Data Challenges in ECHO Data Analysis

- **Who**
  - Centralized (Data Analysis Center) v. decentralized

- **What**
  - Harmonization
    - Esp. for existing data
  - Different datasets for different questions
ECHO-Wide Cohort
“Dynamic Cohort of Inception Cohorts”

Cohorts recruited at different points in the life course and in different eras.
ECHO-Wide Cohort
“Dynamic Cohort of Inception Cohorts”

Cohorts recruited at different points in the life course and in different eras with heterogeneity in retention within each and different follow-up schedules and different measures and combinations of existing and new data
Big Data Challenges in ECHO Data Analysis

- Who
- What
- How
  - Credit system?
Big Data Challenges in ECHO Data Analysis

- **Who**
- **What**
- **How**

- **Type of scientific question**
  - Prediction → “precision prevention”
  - Etiology → primordial prevention
Big Data Challenges in ECHO Data Analysis

• Etiology
  – Causal relationships between exposures and outcomes
  – Most (big data) analysis approaches don’t distinguish
    • Confounding—a nuisance, control for it
    • Mediation (pathway, mechanism)—interesting!
Extra slides
Move the Needle on Data Sharing

• Among investigators
• For public use
• With individual participants
Move the Needle on Data Sharing

– “It’s just for genetics”
– “I’ve got 10 million variables in raw form and another 10,000 derived variables, and I’ve spent years cleaning them. No one else will understand how to use them, especially longitudinally.”
– “I don’t want my data out there before my team—esp. my junior investigators—and I have a chance to analyze them.”
– “NIH says I have to do it, so I will—but just the minimum necessary.”
Move the Needle on Data Sharing

• Need for nuanced approach
  – Adheres to the principles
    • We win when we all win
    • Big data are better than small
    • Publicly funded data are, in the end, public
  – Takes into account investigators’ fears
  – Plays by the rules

• Lessons learned from IC consortia
### Sample Size and Racial/Ethnic Distribution

<table>
<thead>
<tr>
<th></th>
<th>Asian</th>
<th>Black or Af-Am</th>
<th>AI/AN</th>
<th>Cauc.</th>
<th>Multi</th>
<th>Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>(5%)</td>
<td>(17%)</td>
<td>(3%)</td>
<td>(66%)</td>
<td>(10%)</td>
<td>(26%)</td>
</tr>
<tr>
<td><strong>~46000 kids, currently</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ECHO-Wide Cohort

What is it?

• Multi-level longitudinal data platform
  – Broad reach
    • For all cohorts
    • Core data, including Common Data Elements
  – Deeper for more detailed phenotypes
    • Fewer (but still multiple) cohorts
    • Additional data at certain points in the life course depending on, e.g., focused exposures or outcomes
      – Microbiome, metagenomics, epigenomics, metabolomics
ECHO Goals

Long, medium, short term

• Improve the health of children and adolescents
  – by conducting observational and intervention research to inform high-impact programs, policies, and practices

• Institute best practices for how to conduct team science in the 21st century
ECHO Goals
Long, medium, short term

• **Early Wins** by 1 year
  – Observation (Cohorts)
    • 1+ aggregate analyses on existing multiple-cohort data
      – No data sharing needed
      – Distributed data analysis approach
    • New data collection protocol
      – Cohorts submit to central IRB
    • Review/methods papers
    • Individual cohort analyses
  – Intervention (IDeA States Pediatric Clinical Trials Network)
    • Infrastructure and training in place to begin 1+ trials
Who We Are
ECHO’s 7 Components

[Diagram showing the 7 components of ECHO: Coordinating Center, Data Analysis Center, Cohort Sites, CHEAR, IDEA States Network, PRO Core, Genetics Core]
### Current Characteristics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mothers enrolled, N</td>
<td>~33000</td>
</tr>
<tr>
<td>Children enrolled, N</td>
<td>~46000</td>
</tr>
<tr>
<td>Age of children, y</td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td>0 – 36</td>
</tr>
<tr>
<td>Minimum age, median</td>
<td>1.5</td>
</tr>
<tr>
<td>Maximum age, median</td>
<td>7.0</td>
</tr>
</tbody>
</table>
### Future

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mothers enrolled, N</td>
<td>~33000</td>
<td></td>
</tr>
<tr>
<td>Children enrolled, N</td>
<td>~46000</td>
<td>Expanding enrollment in some cohorts</td>
</tr>
</tbody>
</table>
ECHO Overall Scientific Goal

- Understand effects of *early* environmental exposures on child health and development
  - Effects: Observation
  - Early: conception to age 5 y
ECHO Overall Scientific Goal

- Understand effects of early environmental exposures on child health and development
  - Effects: Observation
  - Early: conception to age 5 y
  - Environmental exposures: Society to biology
    - Physical and chemical
      - Air pollution
      - Chemicals in our neighborhoods
    - Societal factors—stress, maltreatment, etc.
    - Social factors—networks, SES, family dynamics, etc.
    - Behavior—sleep, diet, etc.
    - Biology—epigenetics, microbiota, etc.
ECHO Overall Scientific Goal

• Understand effects of early environmental exposures on child health and development
  – Effects: Observation
  – Early: conception to age 5 y
  – Environmental exposures: Society to biology
  – Child health and development
    • High-impact conditions
    • 4 original focus areas
      – Pre/peri/post-natal outcomes
      – Upper and lower airway
      – Obesity/dysmetabolism
      – Neurodevelopment
    • + Child Health