It’s Not All Roses: When Data Integration Can Be Misleading and Criteria for When Should It Be Pursued

Chris Gennings, Ph.D.
Department of Environmental Medicine and Public Health
Icahn School of Medicine at Mount Sinai
New York
chris.gennings@mssm.edu

February 2018
Analysis of Integrated (BIG) Data – Not all Roses

• Statistical concepts are still important (e.g., type 1 & 2 errors, experimental design, randomization, reproducibility, bias)

• Exploratory versus confirmatory research
  • Hypothesis driven or “fishing expedition”

• What p value to use; meaningful effect sizes

Cynical view: Are we “chasing” latest technologies?
  • Do they answer important research questions OR do we ask questions based on technologies?

Bigger isn’t always better – but it IS generally more complex...
Data Integration

Integrating disparate study types: assumption of relevance
• Linking environmental exposures and health outcomes
  • External exposures with locality-specific health records
• Linking human data and experimental study results for risk assessment
  • EXAMPLE:

Integrating data across epidemiology studies: evaluation of compatibility
• For generalizability across studies: exposures, subject characteristics
• To increase sample sizes (BIG DATA) and MAYBE power
• EXAMPLE: OP pesticides & fetal growth; 4 birth cohorts; “limited by differences in study populations” ( - Harley et al, 2016; EHP)
Meta Analyses and Meta Regression

• A meta analysis is an analysis of analyses; generally includes a systematic review; results in a (weighted) average of estimates.
  • Dates back to 17th century studies of astronomy; Karl Pearson published collated data from several studies of typhoid inoculation in 1904 ( - Wikipedia)
  • File-drawer problem

• A meta regression adjusts meta analyses with moderator variables using regression techniques.

PROBLEM:

• Different exposure ranges across studies
• Nonlinearity in the “low dose” region

Solution: Consideration of pooled data (e.g., Lanphear et al, 2005; EHP)
Pooled Data with Heterogeneous Concentration Ranges and Nonlinearity

**Figure:** Linear models for each cohort study in the pooled analysis, adjusted for maternal IQ, HOME score, maternal education, and birth weight with intra-study ranges depicted by 5th and 95th percentiles of the concurrent blood lead level at the time of the IQ testing (Lanphear et al 2005; EHP)

**Figure:** Log-linear model for concurrent blood lead concentration along with linear models for concurrent blood lead levels among children with peak blood lead levels above and below 10 µg/dL (Lanphear et al 2005; EHP)
Considerations for Pooled Data
- Stingone et al, 2017

Focus: integration of epidemiology data with advantages to shared data centers

- Data harmonization: e.g., education, SES
- Different associations across studies between covariates and outcomes (**CASE 1: main effects; CASE 2: interactions**)
- Temporality – important in pediatric environmental health
- Multi domain outcomes (e.g., neurodevelopment) with multiple scales (**CASE 1**)
- Multiple labs (different labs per study; **CASE 3**)
- Using selected inter-study subjects for case control matching (**CASE 4**)
Children’s Health Exposure Analysis Resource (CHEAR) Goals

Goal 1
Advance understanding of the impact of environmental exposures on children’s health and development

Goal 2
Provide infrastructure for adding or expanding exposure analysis to studies involving research in children’s health
CHEAR Laboratory Quality Assurance/Quality Control

- Method Blanks
- Labeled Standards

Within Laboratory (Between Run/Batch)
- QC Samples/Standards
- Pooled Material

Between Laboratory (Within CHEAR)
- Round Robin
- Common QC materials
- NIST Samples

Between Laboratory (External to CHEAR)
- NYDOH PT Program for Trace Elements
- G-EQUAS (organics)
CASE 1: Pooled Data without increased power

3 studies of concurrent Pb and cognitive development
• PROGRESS: birth cohort in Mexico City
• ALSPAC: birth cohort in Avon, England from 1990s
• Uruguay Study: 1st grade evaluation; Montevideo, Uruguay

Harmonization issues: Uruguay study required correction by school and test indicators; heterogeneous associations with covariates

<table>
<thead>
<tr>
<th>Study</th>
<th>PROGRESS</th>
<th>ALSPAC</th>
<th>Uruguay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home eval</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Harmonization issues led to exposure evaluation from residuals of covariate only models
CASE 1: Pooled Data without increased power

Table: Intra- and inter-study analyses (Beta (SE)) of log(Pb) and adjusted Cognitive Score

<table>
<thead>
<tr>
<th>PROGRESS (N=193)</th>
<th>ALSPAC (N=193)</th>
<th>Uruguay (N=288)</th>
<th>Combined</th>
<th>Combined w/ Random Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7 (3.7)</td>
<td>-6.1 (3.3)</td>
<td>0.08 (3.4)</td>
<td>-1.64 (2.0)</td>
<td>-1.3 (1.8)</td>
</tr>
<tr>
<td>p=0.651</td>
<td>p=0.064</td>
<td>p=0.981</td>
<td>p=0.418</td>
<td>p=0.452</td>
</tr>
</tbody>
</table>

Increased sample size did not provide more power to detect an association. The wider concentration range in ALSPAC made the signal stronger there.
CASE 2: Study by covariate interactions

Combining 2 CHEAR Pilot Studies: Prenatal Metals and Birth Weight

**Pilot 34** (Christiani): 192 pregnant (<16 wks) women from two provinces in Bangladesh

**Pilot 36** (Ferguson): 390 pregnant women recruited from Brigham and Women’s Hospital in Boston (2008-2011)

**Outcome variable**: Birth weight

**Covariates to be HARMONIZED**: Parity, sex, gestational age at birth, maternal BMI (1st Trim), maternal age, education level; but, different dilution factors

**Common metals**: As, Cd, Mn, Mo, Pb (ug/L)
Comparison of Concentration Distributions:

Violin plots (Bangladesh = Red; Boston = Blue)

Log As
Log Cd
Log Mn
Log Mo
Log Pb
Characterizing Study Differences: Logistic regression

**Modeling \( P(\text{Study=Bang}) \)**

- Beta > 0: Higher in Bang
- Beta < 0: Lower in Bang

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Beta Estimate</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Weight</td>
<td>Pos</td>
<td>0.016</td>
</tr>
<tr>
<td>Parity (0, &gt;0)</td>
<td>Pos</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Female</td>
<td>Pos</td>
<td>0.082</td>
</tr>
<tr>
<td>Gestational Age at Birth</td>
<td>Pos</td>
<td>0.002</td>
</tr>
<tr>
<td>Maternal BMI (1st trim)</td>
<td>Neg</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Maternal Age</td>
<td>Neg</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Urinary Specific Gravity</td>
<td>Neg</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Birth Weight * Gest Age</td>
<td>Neg</td>
<td>0.006</td>
</tr>
</tbody>
</table>

**Figure**: Different associations between gestational age at birth and birth weight
CASE 3: Combining Data Across Labs and Studies

Testing for Process Control Across Labs, Studies and Batches

- Common QC pools (urine, blood) run in every CHEAR study and all batches
  - 2 pools generated per matrix
  - Batch sizes of ~100 use 6 QC pools per batch from both pools
  - Batch sizes of ~50 use 3 QC pools per batch from both pools

- Plans to evaluate this process (e.g., # of QCs per batch) after implementation

- Multivariate Control Charts
Multivariate Control Charts

**OBJECTIVE**: evaluate whether the process is “in control”
- Generalizable across labs & studies

**Case study**:
- 17 metals evaluated from whole blood in 10 batches
- 6 QC pool samples run per batch
  - 3 of QC pool=0 & 3 of QC pool=1
- T Square statistic is multivariate version of a t statistic
  - 17-dimensional mean vector and covariance matrix
- We approximated T Square using first 5 PCs (accounted for >90% of variability)
- Alpha=0.01
**CASE 4**: Selection of Inter-study subjects for Case Control Matching

- **Cases** from national disease registries: e.g., NASH Clinical Research Network
- **Matched Controls**: Stratification Score Matching using available covariates
  - e.g., age, cohort year maternal age, sex, parental education, child’s race, maternal birth place, homeowner status
- **Per variable % reduction in bias after matching range:**
  - Internal controls: 0 - 12%
  - External controls: 61 - 100%
- **RESULT**: Average exposure differences from matched pairs using, say, 1000 random case orders for matching
Final Thoughts

RECOMMENDATIONS: Do NOT integrate “blindly”

• Thoughtful integration of study data and resulting analyses
• Investigate potential heterogeneity across studies
• Develop a metrics/figures to demonstrate heterogeneity and evaluation strategy for adjusting for it

• Advantages of team science (exposure experts, subject-matter experts; quantitative experts) - do not work alone!
• Conduct confirmatory studies.
Thank you!
And Thanks to:

Pb Biomarker Consortium:
  Katarzyna Kordas
  Mara Tellez Rojo
  Caroline Taylor
  Bob Wright

CHEAR Network; particularly, Data Center Statistics Team:
  Elena Colocino         Stefanie Busgang
  Paul Curtin           Matthew Mazzella
  John Doucette
  Maya Deyssenroth
  Megan Niedzwiecki

CHEAR Pilot Study PI’s:
  David Christiani
  Kelly Ferguson