Data Linkage

Tellen D. Bennett, MD, MS

Associate Professor, Pediatric Critical Care
University of Colorado School of Medicine
Children’s Hospital Colorado
Investigator, ACCORDS
Co-Director, Analytics Core, Data-Driven Decisions and Discovery (D4)

tell.bennett@ucdenver.edu

October 13, 2016
Why Link Data?

Efficiency

- ROI for funders
- More health knowledge for the public
- Rare disease, heterogeneous patients: may be necessary
Why Link Data?

To answer different questions

- To adjust for disease severity adequately
- Longitudinal or patient-reported outcomes
- Community or social determinants of health
- ...
To answer different questions...

# Data Merging - More patients, same data fields

<table>
<thead>
<tr>
<th>Last Name</th>
<th>LOS</th>
<th>Died?</th>
<th>Crani?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Willson</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Thomas</td>
<td>9</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tamburro</td>
<td>11</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Watson</td>
<td>4</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crani?</th>
<th>Died?</th>
<th>LOS</th>
<th>Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>5</td>
<td>Flori</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>8</td>
<td>Markovitz</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>3</td>
<td>Dean</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>10</td>
<td>Thompson</td>
</tr>
</tbody>
</table>
## Data Merging - More patients, same data fields

<table>
<thead>
<tr>
<th>Last Name</th>
<th>LOS</th>
<th>Died?</th>
<th>Crani?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dean</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Flori</td>
<td>5</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Markovitz</td>
<td>8</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tamburro</td>
<td>11</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Thomas</td>
<td>9</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Thompson</td>
<td>10</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Watson</td>
<td>4</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Willson</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

1. Making data 'longer'
2. Stata: 'append', R: 'rbind()'
Data Merging - ARDS Example

The Merging of 5 PALISI Studies

- Long Term FU N=183
- Prone N=102
- Fluids N=171
- Calfactant N=109
- CARDS N=152

N = 711
18 enrolled in >1 study
N = 693

\[1\] Ward S and Flori H, with permission
Data Merging - More patients, same data fields

Advantages

▶ Statistical Power
▶ Generalizability (maybe)

Disadvantages

▶ Same questions
Data Merging - Dependencies

NIH/trialist: Common Data Elements

Streamline Your Neuroscience Clinical Research using content standards that enable clinical investigators to systematically collect, analyze, and share data across the research community.

The NINDS strongly encourages researchers who receive funding from the Institute to ensure their data collection is compatible with these common data elements (CDEs). Learn more about the CDE Project.
Data Merging - Dependencies
Informaticist: Common Data Model e.g. OMOP
This idea is hot
## Data Linkage: Same patients, more data fields

<table>
<thead>
<tr>
<th>Last Name</th>
<th>LOS</th>
<th>Died?</th>
<th>Crani?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flori</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Dean</td>
<td>9</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Thompson</td>
<td>11</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Markovitz</td>
<td>4</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adm OI</th>
<th>VFD-28</th>
<th>ICU LOS</th>
<th>Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0</td>
<td>2</td>
<td>Flori</td>
</tr>
<tr>
<td>22</td>
<td>11</td>
<td>3</td>
<td>Markovitz</td>
</tr>
<tr>
<td>31</td>
<td>7</td>
<td>16</td>
<td>Dean</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>12</td>
<td>Thompson</td>
</tr>
</tbody>
</table>
Data Linkage: Same patients, more data fields

<table>
<thead>
<tr>
<th>Name</th>
<th>LOS</th>
<th>Died?</th>
<th>Crani?</th>
<th>Adm</th>
<th>OI</th>
<th>ICU</th>
<th>LOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dean</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>31</td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>Flori</td>
<td>5</td>
<td>No</td>
<td>Yes</td>
<td>16</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Markovitz</td>
<td>8</td>
<td>No</td>
<td>No</td>
<td>22</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Thompson</td>
<td>10</td>
<td>No</td>
<td>No</td>
<td>9</td>
<td></td>
<td></td>
<td>12</td>
</tr>
</tbody>
</table>

1 Making data 'wider'

2 Stata: 'merge', R: 'cbind()'
Data Linkage - Same patients, more data fields

Advantages

- Different Questions

Disadvantages

- Same statistical power
- Same generalizability
How: Interesting Challenges

- More patients: Harmonization
- More data fields: Record Linkage
- Both: Governance, Access, Authorship, ...
Focus: Patient-Level Record Linkage

- Deterministic: the data files share a unique identifier
- Probabilistic: the data files share enough variables with enough information content
Warning - Math Ahead

"Do the arithmetic or be doomed to talk nonsense."

\(^1\)credited to John McCarthy
Information Content ~ Discriminating Power

An adaption of Shannon’s general formula for uncertainty/entropy

\[ H = \sum_{i=1}^{n} p_i \cdot \left( \log_2 \frac{1}{p_i} \right) \]
Information Content

- Units = bits (from $\log_2 x$)

- $\log_2 2 = 1$ bit (evenly split biological gender)

$$H = \sum_{i=1}^{n} p_i \cdot \left( \log_2 \frac{1}{p_i} \right)$$
Information Content

For a given variable, information content \( \sim \)

- \# of values = \( n \)
- \( \text{Pr(each value)} = p_i \)
- % missing

\[
H = \sum_{i=1}^{n} p_i \cdot \left( \log_2 \frac{1}{p_i} \right)
\]
Probabilistic Linkage

- Bayesian process
- Reasonable analogy: true link status is "missing" data to be imputed as a function of known data

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]

Bayes’ theorem

\(^1\)Image: Matt Buck
## Goal: 2 x 2 table

<table>
<thead>
<tr>
<th>Linkage</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>True+</td>
</tr>
<tr>
<td>No</td>
<td>False-</td>
</tr>
</tbody>
</table>

Yes

No
### Example

<table>
<thead>
<tr>
<th>Last Name</th>
<th>LOS</th>
<th>Died?</th>
<th>Crani?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elway</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Kempe</td>
<td>9</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Smith</td>
<td>11</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Fairclough</td>
<td>5</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crani?</th>
<th>Died?</th>
<th>LOS</th>
<th>Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>5</td>
<td>Fairclough</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>8</td>
<td>Kempe</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>3</td>
<td>Elway</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>10</td>
<td>Smith</td>
</tr>
</tbody>
</table>
**Example**

<table>
<thead>
<tr>
<th>Last Name</th>
<th>LOS</th>
<th>Died?</th>
<th>Crani?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elway</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crani?</th>
<th>Died?</th>
<th>LOS</th>
<th>Last Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>3</td>
<td>Elway</td>
</tr>
</tbody>
</table>
Example - no PHI

<table>
<thead>
<tr>
<th>LOS</th>
<th>Died?</th>
<th>Crani?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crani?</th>
<th>Died?</th>
<th>LOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>8</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>10</td>
</tr>
</tbody>
</table>
Traumatic Brain Injury

Pediatric TBI in the United States, annually:

- 2,200 deaths\(^1\)\(^2\)
- 35,000 hospitalizations
- 474,000 ED visits
- est. $2-3 billion in hospital costs

Clinical Problem

Elevated intracranial pressure (ICP) may develop after TBI

- cellular swelling
- blood-brain barrier disruption
- intracranial hemorrhage

and cause secondary injury

---

ICP monitoring
Research Question

Do ICP monitors improve outcome in children with severe TBI?
Research Challenge

- No single database contains the necessary variables to answer the question
Linkage Aim

- Specific Aim 1: **Link** two databases with complementary clinical information
SA 1: Two Complementary Databases

**Pediatric Health Information Systems (PHIS) Database**
- Children’s Hospital Association (CHA)
- Business alliance of 44 U.S. Children’s hospitals

**National Trauma Data Bank (NTDB)**
- American College of Surgeons (ACS)
- Standardized collection of trauma registries from 700+ hospitals
Aim 1 Results: Validation Linkage, 2007-2010

For patients with severe TBI eligible for ICP monitor study:

- Sensitivity: 88%
- Specificity: 99.99%
- Positive Predictive Value: 98%

\(^1\)Bennett T et al. Methods Inf Med 2015.
Aim 1 Results: Overall Linkage, 2007-2010

- Age <18 years, trauma diagnosis
- hospital match (30 hospitals)

**PHIS file** 156,357 records

**NTDB file** 104,049 records

---

1 Bennett T et al. Methods Inf Med 2015.
Specific Aim 1: Why does this Work?
Specific Aim 1: Why does this Work?

- Children with severe TBI are sick
  - Linkage Rate $\propto$ Injury Severity
    - Mechanism: information content
Specific Aim 1: Why does this Work?

- Children with severe TBI are sick
  - Linkage Rate $\propto$ Injury Severity
  - Mechanism: information content

![Graph showing sensitivity and specificity for different categories of TBI and trauma. The graph plots sensitivity on the y-axis against 1 - specificity (false positive rate) on the x-axis. The lines represent Severe TBI, All TBI, All Trauma, and Severe Trauma, respectively.](image)
Specific Aim 1: Why does this Work?

- Children’s hospitals are not common
  - Iteratively, a hospital identifier can be estimated well
Specific Aim 1: Why does this Work?

- Children’s hospitals are not common
  - Iteratively, a hospital identifier can be estimated well
- Trauma ICD-9-CM codes contain extra information
  - injury type
  - mechanism
International Linkage Community

2016 International Population Data Linkage Conference

TWEET US #IPDLN2016

Linking Data - Improving Lives

PRE-CONFERENCE WORKSHOPS
AUGUST 22 - 23

MAIN CONFERENCE
AUGUST 24 - 26

THE GREAT HALL, SWANSEA UNIVERSITY BAY CAMPUS, WALES, UNITED KINGDOM
Local Linkage Community

Informatics Core of

Data-Driven
Decisions and
Discovery (D4)

Michael Kahn

Toan Ong
Conclusions

- Data Linkage is a powerful tool
Conclusions

- Data Linkage is a powerful tool
- Interest and Anschutz Campus capability are growing
Disclosures

Financial Support

- NICHD K23HD074620
- PCCTSDP, NICHD K12HD047349
- Children’s Hospital Colorado Research Institute
- NIH/NCRR Colorado CTSI UL1 TR000154