GUILD: GUidance for Information about Linked Datasets
by Ruth Gilbert (University College London)

Presenter: Christine P. Chai

Department of Statistical Science
Duke University

September 20, 2016
Why is Guidance in Data Linkage Important?

- Large administrative datasets involve many teams in processing
- Information should be passed through the data linkage pathway
- Sharing information leads to more robust methods
- Example: data analysis that takes linkage error into account

https://www.newton.ac.uk/seminar/20160913140014301
Four Steps in the Data Linkage Pathway

The methods and results done in earlier steps should be communicated to the later steps

1. **Data Provision**: data descriptions, applications, agreements
2. **Data Linkage**: both at aggregate and record levels
3. **Data Analysis**: methodology and model validation
4. **Reporting**: writing up a technical report
If we do not communicate between steps, these can happen...

1. **Data Provision**: limited awareness of data quality → linkage error

2. **Data Linkage**: insufficient knowledge of linkage error effect

3. **Data Analysis**: incorrect data values due to pre-processing error

4. **Reporting**: lack information on earlier processing and analysis
GRE AWA Argument Question: “A recent study shows that people living on the continent of North America suffer 9 times more chronic fatigue and 31 times more chronic depression than do people living on the continent of Asia. Interestingly, Asians, on average, eat 20 grams of soy per day, whereas North Americans eat virtually none. It turns out that soy contains phytochemicals called Isoflavones, which have been found to possess disease-preventing properties. Thus, North Americans should consider eating soy on a regular basis as a way of preventing fatigue and depression.”

[http://www.testpreppractice.net/GRE/awa-samples/gre-argument-essay-1.html](http://www.testpreppractice.net/GRE/awa-samples/gre-argument-essay-1.html)
Bad Report Example 1 – Problems

- No information is given about the study design
- Fatigue and depression are different and should not be combined
- Too many potential confounding factors: climate, lifestyle, diet, etc.

If more information were given about the clinical dataset, we could make inferences with more concrete evidence.
Paper: Which Nudges Do People Like? A National Survey

Table 2: American Attitudes Toward Five Educational Campaigns

<table>
<thead>
<tr>
<th></th>
<th>Childhood obesity</th>
<th>Distracted driving</th>
<th>Sexual orientation discrimination</th>
<th>Movie theaters</th>
<th>Animal Welfare Society</th>
<th>Obesity (arguably manipulative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total support (in percentages)</td>
<td>82/18</td>
<td>85/15</td>
<td>75/25</td>
<td>53/47</td>
<td>52/48</td>
<td>57/43</td>
</tr>
<tr>
<td>Democrats</td>
<td>90/11</td>
<td>88/12</td>
<td>85/15</td>
<td>61/39</td>
<td>59/41</td>
<td>61/40</td>
</tr>
<tr>
<td>Independents</td>
<td>81/19</td>
<td>84/16</td>
<td>75/25</td>
<td>51/49</td>
<td>55/45</td>
<td>60/40</td>
</tr>
<tr>
<td>Republicans</td>
<td>70/30</td>
<td>80/20</td>
<td>57/43</td>
<td>41/59</td>
<td>34/66</td>
<td>47/53</td>
</tr>
</tbody>
</table>

---

Bad Report Example 2 – Problems

- Sampling frame not well-defined (just “Americans”?)
- No raw number of respondents is reported, only the percentages
- What is the survey response rate?
  Watch out for survey non-response bias

- Missing information about the data is either unintentional (lack of communication) or intentional (deliberately concealing)
Equator Network (Enhancing the QUAlity and Transparency Of health Research) - reporting guidelines and checklists

http://www.equator-network.org/

Christine P. Chai (Duke Stat)
Quoted part of the “CONSORT 2010 checklist of information to include when reporting a randomised trial”

- Trial design: Description of trial design (such as parallel, factorial) including allocation ratio
- Participants: Eligibility criteria for participants
- Interventions: The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
- Blinding: If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how
Conclusion

- **Guidance** should be communicated across steps

Data Provision
Data Linkage
Data Analysis
Reporting