Comparing CIREN and NASS Cases: A Multivariate Similarity Approach

Joel Stitzel, Ph. D.,
Patrick Kilgo, M. S.,
Brian Schmotzer, M. S.,
Clay Gabler, Ph. D.,
J. Wayne Meredith, M.D.

Toyota – Wake Forest University
School of Medicine
CIREN Center
Impetus: The Question

- CI REN - excellent source for detailed injury data
- NASS weighted sample of real world cases
- Question: How do we draw conclusions about injuries and safety using CI REN data, knowing that it accurately describes real world injuries? For:
  - Biomechanics Research
  - Safety Regulation
  - Automotive Industry
NASS/CDS vs. CI REN

- **NASS/CDS**
  - National Sample
  - Crash Data
  - Entry Criteria: Tow-away Crashes
  - 4000-5000 crashes/year
  - Large contingent of MAIS 1-2 injuries
  - Reliable Sampling Weights

- **CI REN**
  - Clinical Data
  - Crash Data
  - Entry Criteria: Level I Trauma Center
  - 300-400 crashes/year
  - Minimum MAIS 3*
  - Detailed Injury Data
  - Exhaustive Outcomes

*CIREN inclusion criteria specify some exceptions to this rule.*
NASS/CDS vs. CI REN

- **Why use NASS/CDS?**
  - Allows National Estimates of Injury
  - Injury data not always complete or detailed (less clinical data)
  - Lots of data, sampling weights, well established, some injury data

- **Why use CI REN?**
  - Detailed Clinical Evaluation of Injuries
  - Cases reviewed by doctors, engineers, crash investigators, etc.
  - Not a population-based sample
  - Difficult to extrapolate to national outcome
Objectives

- **Quantify** the similarity/difference between a given CI REN case (or a subset of CI REN cases) and the population of NASS cases

- **Hypothetical questions:**
  - Given a subset of NASS cases of interest:
    - How do we identify CI REN cases that “match it”?
    - What is the best way to make a comparison?
    - What variables in NASS or CI REN are important in describing the differences?
NASS/CDS vs. CI REN

Is CI REN a subset of NASS/CDS?

What is the nature and size of the overlap?

Does CI REN contain crash scenarios not in NASS/CDS?
Main Question

- How similar are individual CI REN cases to the average NASS case?
- By extension: how similar is a group of CI REN cases of interest, to a population of NASS cases of interest?
General Approach

- Compute a “similarity score” between CI REN cases and the average NASS case.

That is, the k-dimensional “distance” between the case types.

Where k = number of variables common between the data sources.
Methods –
Basics of Mahalanobis Distance

- A multivariate measure of distance
- It puts all variables on the same scale – standardizes them
- Takes into account the correlation between the k variables
Mahalanobis Distance

- Based on **correlations** between variables.
- Different patterns can be identified and analyzed.
- Useful way of determining **similarity** of an unknown **sample set** to a known one.
- Differs from **Euclidean** distance in that it takes into account the correlations of the data set and is **scale-invariant**, i.e. not dependent on the scale of measurements.

http://en.wikipedia.org/wiki/Mahalanobis_distance
Example: Case Study of Height/Weight of Two Men

- **Male #1:**
  - 6 inches above average height
  - 100 pounds **above** average weight

- **Male #2**
  - 6 inches above average height
  - 100 pounds **below** average weight

By Euclidean Distance, these cases are equidistant from the average (centroid).

By Mahalanobis Distance, Male #2 is MUCH further from the average.
Previous Example:

Euclidean vs. Mahalanobis:

- Euclidean distance measures the straight-line distance between two points.
- Mahalanobis distance takes into account the correlation and scale of the data.

- Male #1 and Male #2 are shown on a 2D graph with Weight on the y-axis and Height on the x-axis.
- The average is marked as 'AVERAGE' and is shown as a green dot.
- Male #1 and Male #2 are represented by orange circles.
- The 1σ circle indicates the standard deviation from the average.
Mahalanobis Distance and Similarity Scores

- Mahalanobis distance is an effective multivariate measure for how far points are apart in k-space in the context of the correlations between them.

- So, we can use Mahalanobis distance as our Similarity Score.
Methods: Data Gathering – Common Variables Between Datasets

- Total Delta V
- Occupant Age
- Weight (lbs)
- Height (ft)
- MAXAIS
- ISS
- Model year
- Gender
- Maximum Intrusion
- # of Lower Extr. Injuries
- # of Upper Extr. Injuries
- # of Head Injuries
- # of Chest Injuries
Methods: Our Sample

- **NASS**
  - All NASS cases meeting following criteria:
    - MAXAIS $\geq 3$
    - 2001 to 2005
  - A subset of 1869 NASS cases (with MAXAIS $\geq 3$)

- **CIREN**
  - All CIREN cases from 2001 to the present (2819)
Methods – Dealing with Missingness

- When there are large amounts of missingness that need to be dealt with.

- Possibilities:
  - Impute with averages
  - Estimate covariance structure with multiple imputation methods
  - Delete these observations

For this study, we simply imputed column-wise averages.

Next step
Methods: NASS Weights

- Used NASS sampling weight coefficient in calculating weighted average for NASS cases
- Used in weighting our “covariance” matrix
  - It’s as if each row was repeated “k” times where “k” is the NASS weight for that row
NASS/CDS is composed of lower severity cases

(Distribution of Cases by MAIS – NASS/CDS 2005 vs. CIREN)
NASS/CDS and CI REN roughly comparable for MAIS3+ Cases

(Distribution of MAIS3+ Cases – NASS/CDS 2005 vs. CI REN)
## Weighted NASS and CI REN Means

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weighted NASS Mean</th>
<th>CI REN Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Delta V</td>
<td>37.5</td>
<td>40.9</td>
</tr>
<tr>
<td>Age</td>
<td>38.0</td>
<td>36.7</td>
</tr>
<tr>
<td>Weight (lbs)</td>
<td>167</td>
<td>160.5</td>
</tr>
<tr>
<td>Height (ft)</td>
<td>5.56</td>
<td>5.39</td>
</tr>
<tr>
<td>MAXAI S</td>
<td>3.7</td>
<td>3.39</td>
</tr>
<tr>
<td>ISS</td>
<td>23.6</td>
<td>21.6</td>
</tr>
<tr>
<td>Mortality</td>
<td>0.24</td>
<td>0.15</td>
</tr>
<tr>
<td>Model Year</td>
<td>1994</td>
<td>1997</td>
</tr>
<tr>
<td>Male</td>
<td>0.56</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Results

- Distance Distribution is roughly the same.
- CI REN has a greater proportion of its cases clustered about the mean distance than NASS does.
- “CI REN is more like NASS than NASS”
Potential outcomes of principal components analysis

- Principal Component 1
- Principal Component 2

- Population 1
- Population 2
- Population 3
Principal Components Analysis

- There are differences between NASS and CIREN.
- But more important part is overlap, they are largely the same.
- CIREN cases are driven to the right by some low MAIS.

These two components explain 38.96% of the point variability.
Could subset CIREN or NASS to minimize Mahalanobis distance depending on interest.

This is shown here in 2-d by minimizing variation in 1st 2 principal components.

These two components explain 38.98% of the point variability.
Distance vs. # Injuries

Number of Head Injuries

Number of Chest Injuries

Number of Upper Extremity Injuries

Number of Lower Extremity Injuries
Distance vs. Crash and Injury Characteristics

- Total Delta V
- Maximum AIS
- Injury Severity Score (ISS)
- Maximum Intrusion
Distance vs. Model Year and Anthropometric Variables

- **Mahalanobis Distance** vs. **Model Year**
- **Mahalanobis Distance** vs. **Age**
- **Mahalanobis Distance** vs. **Weight**
- **Mahalanobis Distance** vs. **Height**
Caveats

- Other regressions could be performed
- Shape of data suggests maybe they should be performed
Distance vs. Age

Mahalanobis Distance vs. Age
Distance vs. Max AIS
Distance vs. # Head Injuries

![Graph showing the relationship between Mahalanobis Distance and Number of Head Injuries](image-url)
## Results – Subset of CI REN cases

“Furthest” from weighted average

**NASS case**

<table>
<thead>
<tr>
<th>Delta V</th>
<th>Head</th>
<th>Chest</th>
<th>Up_ex</th>
<th>Low_ex</th>
<th>Age</th>
<th>Weight</th>
<th>Height</th>
<th>Gender</th>
<th>Max_int</th>
<th>Max_AIS</th>
<th>ISS</th>
<th>Dead</th>
<th>Model Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>64.3</td>
<td>201</td>
<td>2.2</td>
<td>1</td>
<td>148</td>
<td>5</td>
<td>43</td>
<td>1</td>
<td>1997</td>
</tr>
<tr>
<td>61</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>40.8</td>
<td>560</td>
<td>6</td>
<td>1</td>
<td>31</td>
<td>3</td>
<td>22</td>
<td>0</td>
<td>2000</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>53.5</td>
<td>220</td>
<td>5.25</td>
<td>1</td>
<td>180</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>1998</td>
</tr>
<tr>
<td>46</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>66</td>
<td>1.74</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2000</td>
</tr>
<tr>
<td>101</td>
<td>0</td>
<td>4</td>
<td>11</td>
<td>22</td>
<td>19.3</td>
<td>143</td>
<td>5.68</td>
<td>1</td>
<td>71</td>
<td>5</td>
<td>43</td>
<td>1</td>
<td>1997</td>
</tr>
<tr>
<td>36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.3</td>
<td>24</td>
<td>1.57</td>
<td>0</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2000</td>
</tr>
<tr>
<td>126</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1.4</td>
<td>26</td>
<td>2.49</td>
<td>1</td>
<td>70</td>
<td>3</td>
<td>10</td>
<td>0</td>
<td>1999</td>
</tr>
<tr>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>13</td>
<td>1.9</td>
<td>1</td>
<td>31</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2001</td>
</tr>
<tr>
<td>41</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>11</td>
<td>1.51</td>
<td>1</td>
<td>12</td>
<td>5</td>
<td>35</td>
<td>0</td>
<td>1993</td>
</tr>
<tr>
<td>45</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.1</td>
<td>11</td>
<td>1.84</td>
<td>1</td>
<td>50</td>
<td>5</td>
<td>57</td>
<td>0</td>
<td>1979</td>
</tr>
</tbody>
</table>
Conclusions

- CI REN crashes are similar to NASS crashes
  - For severe injuries (MAIS 3)
- Valuable approach to improve understanding and use of CI REN
- Could refine method to be able to assign a NASS similarity score to each CI REN crash
- Or NASS crash similarity or anthropomorphic similarity score or injury similarity score
- Good method to quality control data in CI REN and NASS
Thank you

- Acknowledgements
  - Toyota
  - Carol Flannagan and Jonathan Rupp, UMTRI
  - Analytical Committee

- Disclaimer: the published material represents the position of the authors and not necessarily that of Toyota or NHTSA

- For more information contact:
  - Dr. Joel Stitzel (jstitzel@wfubmc.edu)