Probabilistic Data Linkage: Basic Methods and Applications

Lawrence Cook, MStat, PhD
Department of Pediatrics
Division of Critical Care
University of Utah
Crash Outcome Data Evaluation System (CODES)

- Initiated in 1992 by the US National Highway Traffic Safety Administration (NHTSA)
- Are safety belts and motorcycle helmets effective at preventing injuries resulting from motor vehicle crashes?
Crash Database

- **Crash**
  - Date, time, crash type
- **Drivers and vehicles**
  - Speed, contributing factors, violations
- **Occupant**
  - Age, gender, seating location, belt usage
- **No medical information about occupants**
EMS Database

- Patient
- Time
- Scene
- Procedures
- Treatments
- Medications
- No information once dropped off at hospital
ED Database

- Patient
- Time
- ICD-9, Procedures, and E Codes
- ED Charges
- No information once admitted to hospital
- No information prior to arrival at ED
Inpatient Database

- Patient
- Time
- ICD-9, Procedures, and E Codes, ISS
- Hospital Charges
- No information prior to admission to hospital
Crash → Analysis Database → ED
EMS → Analysis Database → Inpatient
Benefits of Safety Belts

- Odds of being admitted or dying
  - 4.3 – 6.5 times higher if not belted
- Odds of emergency department or worse
  - 2.8 – 3.5 times higher if not belted
- Odds of any injury
  - 1.9 – 4.1 times higher if not belted
- Hospital charges for unbelted
  - 55% increase among hospitalized persons
  - 400% increase among all persons
Probabilistic Linkage

- Probabilistic linkage is a method that uses properties of variables common to databases to determine the probability that two records refer to the same person and/or event.
Let’s Play 20 Questions

I’m thinking of a person
Record Linkage with Imperfect Data

### Crash Record

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>DOB</th>
<th>Date</th>
<th>Time</th>
<th>Location</th>
<th>Seat</th>
<th>Belt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary Smith</td>
<td>F</td>
<td>05/05/45</td>
<td>07/15/10</td>
<td>11:40</td>
<td>Weber</td>
<td>US5</td>
<td>N</td>
</tr>
</tbody>
</table>

### Hospital Record

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>DOB</th>
<th>Date</th>
<th>Time</th>
<th>Location</th>
<th>Diagnosis</th>
<th>Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary Smith Sanchez</td>
<td>F</td>
<td>05/05/44</td>
<td>07/15/10</td>
<td>11:51</td>
<td>Weber</td>
<td>Fracture</td>
<td>Mem Hosp</td>
</tr>
</tbody>
</table>
Probabilistic Linkage Theory

Reliability ($m$)

Probability that a common variable agrees on a matched pair.
Approximately 1 - error rate.

Discriminating Power ($u$)

Probability that a common variable agrees on an unmatched pair.
Approximately the probability of agreeing by chance.
Probabilistic Record Linkage

Crash Record
Mary Smith                  F  05/05/45  07/15/10 11:47  Weber  US5  Seat=1  Belt=N

Hospital Record
Mary Smith Sanchez   F  05/05/44  07/15/10 11:55  Weber  Fracture  Mem Hosp

Probability of true match = 0.0009
# Probabilistic Record Linkage

## Crash Record

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>DOB</th>
<th>Date Input</th>
<th>Time Input</th>
<th>Location</th>
<th>Seat</th>
<th>Belt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary Smith</td>
<td>F</td>
<td>05/05/45</td>
<td>07/15/96</td>
<td>11:47</td>
<td>Weber</td>
<td>US5</td>
<td>Seat=1 Belt=N</td>
</tr>
</tbody>
</table>

## Hospital Record

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>DOB</th>
<th>Date Input</th>
<th>Time Input</th>
<th>Diagnosis</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary Smith Sanchez</td>
<td>F</td>
<td>05/05/44</td>
<td>07/15/96</td>
<td>11:55</td>
<td>Fracture</td>
<td>Mem Hosp</td>
</tr>
</tbody>
</table>

**Probability of true match = .0192**
Probabilistic Record Linkage

Crash Record

Mary Smith  F 05/05/45 07/15/96 11:47  Weber  US5 Seat=1 Belt=N

Hospital Record

Mary Smith Sanchez  F 05/05/44 07/15/96 11:55  Weber  Fracture Mem Hosp

Probability of true match = .0385
### Probabilistic Record Linkage

#### Crash Record

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>DOB</th>
<th>Date of Incident</th>
<th>Seat</th>
<th>Belt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary Smith</td>
<td>F</td>
<td>05/05/45</td>
<td>07/15/96</td>
<td>1</td>
<td>N</td>
</tr>
</tbody>
</table>

#### Hospital Record

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>DOB</th>
<th>Date of Incident</th>
<th>Hospital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary Smith Sanchez</td>
<td>F</td>
<td>05/05/44</td>
<td>07/15/96</td>
<td>Mem Hosp</td>
</tr>
</tbody>
</table>

**Probability of a true match = 0.1429**
Probabilistic Record Linkage

Crash Record

Mary Smith                  F  05/05/45  07/15/10 11:47  Weber  US5  Seat=1  Belt=N

Hospital Record

Mary Smith Sanchez   F  05/05/44  07/15/10 11:55  Weber  Fracture  Hosp

Probability of a true match = 0.9836
Probabilistic Record Linkage

Crash Record
Mary Smith                  F  05/05/45  07/15/10 11:47  Weber  US5  Seat=1  Belt=N

Hospital Record
Mary Smith Sanchez   F  05/05/44  07/15/10 11:55  Weber  Fracture  Mem Hosp

Probability of a true match = 0.9817
Probabilistic Record Linkage

Crash Record

Mary Smith  F  Weber US5 Seat=1 Belt=N

Hospital Record

Mary Smith Sanchez  F  Weber Fracture Mem Hosp

Probability of a true match = 0.9999
Probabilistic Record Linkage

This pair of records has both agreements and disagreements. Our calculations say that the odds are $p = 0.9999$ that the records refer to the same individual and crash event.
Research Studies
Impact of Passengers on Crash Outcomes of Teenage Drivers?

Motor Vehicle Crash Hospital Discharge Vital Records
## Risk of Hospitalization or Death to the Teenage Driver

<table>
<thead>
<tr>
<th></th>
<th>Teens Odds Ratio</th>
<th>Adults Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any passenger vs. alone</td>
<td>1.7 (1.4,2.2)</td>
<td>1.3 (1.2,1.4)</td>
</tr>
<tr>
<td>1 passenger vs. alone</td>
<td>1.6 (1.3,2.1)</td>
<td>1.3 (1.1,1.4)</td>
</tr>
<tr>
<td>≥ 2 passenger vs. ≤ 1</td>
<td>1.6 (1.2,2.1)</td>
<td>1.2 (1.1,1.4)</td>
</tr>
<tr>
<td>≥ 3 passenger vs. ≤ 2</td>
<td>1.7 (1.2,2.4)</td>
<td>1.1 (1.0,1.3)</td>
</tr>
<tr>
<td>≥ 4 passenger vs. ≤ 3</td>
<td>1.9 (1.2,3.2)</td>
<td>1.3 (1.1,1.7)</td>
</tr>
<tr>
<td>≥ 5 passenger vs. ≤ 4</td>
<td>2.5 (1.1,5.6)</td>
<td>1.8 (1.3,2.6)</td>
</tr>
</tbody>
</table>
What types and how many injuries will occur in shop class over a one year period?

Student Injury Reports
Emergency Department
Hospital Discharge
Shop Class Injuries

One-year ED
- 167 in class injuries
- 45 seen at ED
- \( \frac{1}{2} \) were saw related
- Open wounds, 64%
- Fractures, 9%
- 2 amputations
- $16,571 ED charges

Five-years Inpatient
- 1,008
- 7 admitted
- 6 table saw related
- 3 amputations
- 2 open wound with tendon damage
- $26,767 hospital charges
Repeat Patients to the Emergency Department

Unduplication of three-years of emergency department data
Findings

- 1.37 million visits by 780,000 patients
- Repeat and frequent users account for 1/3 of patients by 2/3 of visits
- Patients attending five or more EDs were more likely to not have insurance
- 1/3 of serial users (≥ 5 visits) in year remained serial users the next year
Defining Serious Injuries for Motor Vehicle Crashes
Crash View of Injuries

- **KABCO**
  - K or killed within 30 days of the crash date
  - A or incapacitating injury
  - B or non-incapacitating injury
  - C or possible injury
  - O or no injury

- Assigned by investigating officer at the crash scene
Serious Injury Rates

- Serious = K or A injuries
- Can serious injury rates be measured similarly across states or over time?
- Case study – Utah
  - Complete redesign of crash report in 2006
  - New definitions for KABCO
<table>
<thead>
<tr>
<th>Utah KABCO</th>
<th>Pre 2006</th>
<th>Post 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>K</strong> – Fatal</td>
<td><strong>K</strong> – Fatal</td>
<td></td>
</tr>
<tr>
<td><strong>A</strong> – Broken bones &amp; bleeding</td>
<td><strong>A</strong> – Incapacitating injury</td>
<td></td>
</tr>
<tr>
<td><strong>B</strong> – Bruises &amp; abrasions</td>
<td><strong>B</strong> – Non-incapacitating injury</td>
<td></td>
</tr>
<tr>
<td><strong>C</strong> – Possible injury</td>
<td><strong>C</strong> – Possible injury</td>
<td></td>
</tr>
<tr>
<td><strong>O</strong> – No injury</td>
<td><strong>O</strong> – No injury</td>
<td></td>
</tr>
</tbody>
</table>
Methods

• Remove all non-injured occupants
• Compare distribution of K, A, B, C injuries before and after crash report change
• Will there be a difference?
Can Hospital Files be Used to Measure Serious Injury Rates?

- Examine an injury severity measure based on hospital information
- Consider non-linked occupants as uninjured
- Maximum Abbreviated Injury Scale (MAIS)
Severe Injury – Medical Record

- MAIS
  - 1 – Minor
  - 2 – Moderate
  - 3 – Serious
  - 4 – Severe
  - 5 – Critical
  - 6 – Not survivable

- Derived from ICD-9 codes using ICDMap90
Summary

• Does wording on crash report matter?
  – KABCO distribution appears to change
  – MAIS remained more consistent

• Extend study to multiple states
Multi-State Analysis
Comparing Serious Injury Rates Across US States

- States determine the reporting criteria for motor vehicle crashes
  - Monetary
  - Injury

- States also control
  - Design and format of crash report
  - Definitions of fields on crash report
# Crash Severity of Injury

<table>
<thead>
<tr>
<th>State A</th>
<th>State B</th>
</tr>
</thead>
<tbody>
<tr>
<td>• K – Fatal</td>
<td>• K – Fatal</td>
</tr>
<tr>
<td>• A – Incapacitated</td>
<td>• A – Life Threatening</td>
</tr>
<tr>
<td>• B – Visible Injury</td>
<td>• B – Serious</td>
</tr>
<tr>
<td>• C – Momentary unconsciousness/Complaint of</td>
<td>• C – Complaint of Pain</td>
</tr>
<tr>
<td>pain</td>
<td></td>
</tr>
<tr>
<td>• O – No injury</td>
<td>• O – No injury</td>
</tr>
</tbody>
</table>
Methods

- Collected data from 11 states from crash years 2005 to 2008
- Remove all non-injured occupants
- Compare distribution of K, A, B, C injuries
KABCO by State
MAIS by State

[Bar chart showing the percentage distribution of MAIS by state for A to J, with different colors representing different severity levels: 6, 5, 4, 3, 2.]
Summary

• A lot of variation between severity of injury coding on state crash reports
• Using MAIS helps to smooth the injury distribution
• More research needed
More Linkage Studies

- Crash to birth certificates
- Crash to bankruptcy
- Poison control to hospital and death
- EMS to hospital, trauma, and death
- Endotracheal intubation outcomes
What Do You Need For Probabilistic Linkage
Data Files

- Data use agreements
- Institutional Review Board (IRB) Approvals
- Memoranda of understanding
- Variables common to both files
Linkage Variables

- Many levels
- Observations spread throughout levels
- Reasonable accuracy
- Mix of person and event information
- Variable definitions same on each file
- Missing values represented by NULL
Common Linkage Variables

First and Last Names

Soundex of Names (Sounds like)
  - Lawrence Cook = L652 C200
  - Laurence Cooke = L652 C200

Date of Birth and Age

Incident Date

Time of Incident

Location: County, City, Zip, Latitude/Longitude
Are Names Necessary for Probabilistic Linkage?
Name Dilema

- Names are powerful identifiers
- Confidentiality concerns
- Names may not be collected in database
- Simulation study to determine effect of name information on linkage projects
  - We know the answers
Linkage Performance Measures

• Sensitivity - Ability to recognize true matches
  % of true matches identified

• Specificity - Ability to recognize incorrect matches
  1 – false positive rate
DOB, Gender, County, Time, Incident Date

Sensitivity

Specificity

Error Rate

NAME

SOUNDEX

INITIALS

NO NAME INFO

Sensitivity Specificity

No Errors 1% 5% 10% 25%

No Errors 1% 5% 10% 25%
Summary

• Is name information necessary?
  – If many non-name identifiers are available then name information may not be needed
  – If few non-name identifiers are available then name information becomes crucial

• Linkage feasibility test
Other Linkage Considerations

• Confidentiality concerns
  – IRBs & data sharing/use agreements
  – Separate tables of identifiers

• Databases
  – Missingness and accuracy of matching fields
  – Timeliness

• Analysis
Probabilistic Linkage Software

- LinkSolv
- Link Plus (CDC)
- Link King
- RecordLinkage (R)
- FRIL
- FEBRL
- Write your own
  - *Handbook of Record Linkage Methods for Health and Statistical Studies*, Howard Newcombe
Software Checklist

- Size of databases
- Add custom variable types and comparisons
- Unduplication / self match
- Link more than two files
- Training and documentation
Questions?

Larry Cook
larry.cook@hsc.utah.edu
801-585-9760
295 Chipeta Way
PO Box 581289
Salt Lake City, UT 84158-0289