Probabilistic Data Linkage: Basic Methods and Applications

Lawrence Cook, MStat, PhD
Department of Pediatrics
Division of Critical Care
University of Utah

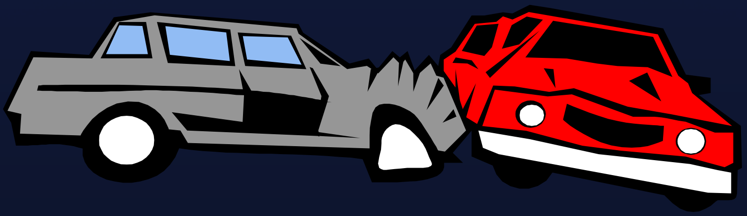
UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

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

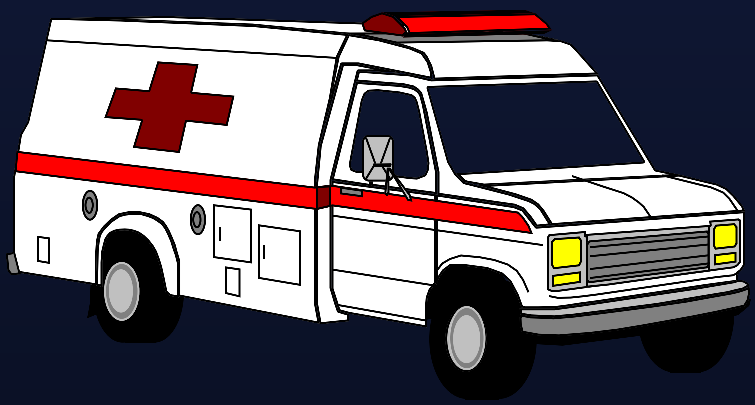


- Drivers and vehicles
  - – Speed, contributing factors, violations
- Occupant
  - – Age, gender, seating location, belt usage
- No medical information about occupants

UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

EMS Database

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

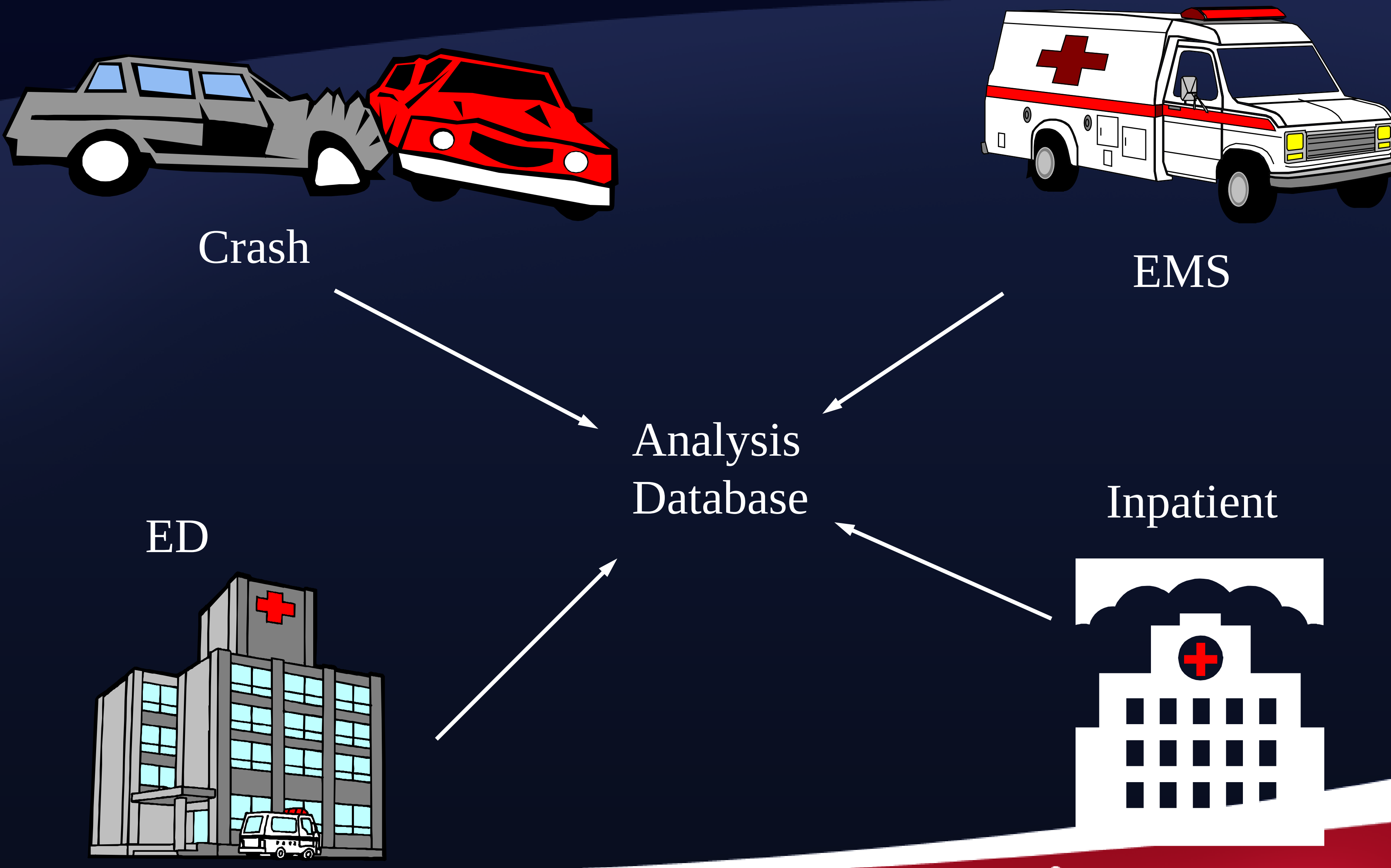


ED Database

- 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
EMS
Analysis
Database
ED
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

Mary Smith	F	05/05/45	07/15/10 11:40	Weber	US5	Seat=1	Belt=N

Hospital Record

Mary Smith Sanchez	F	05/05/44	07/15/10 11:51	Weber	Fracture	Mem Hosp

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 15/10 11:47 Weber US5 Seat=1 Belt=N

Hospital Rec

Mary Smith Sa 15/10 11:55 Weber Fracture Mem Hosp

Probability of
true match = 0.0009

Probabilistic Record Linkage

Crash Record

Mary Smith F 05 Weber US5 Seat=1 Belt=N

Probability of
true match = .0192

Hospital Record

Mary Smith Sanchez F 05 Weber Fracture Mem Hosp

Probabilistic Record Linkage

Crash Record

Mary Smith F 05/05/45 (... JS5 Seat=1 Belt=N

Hospital Record

Mary Smith Sanchez F 05/05/44 ... Fracture Mem Hosp

Probability of
true match = .0385

Probabilistic Record Linkage

Crash Record

Mary Smith F 05/05/45 07/15/ eat=1 Belt=N

Hospital Record

Mary Smith Sanchez F 05/05/44 07/15/ re Mem Hosp

Probability of a true match = 0.1429

Probabilistic Record Linkage

Crash Record

Mary Smith F 05/05/45 07/15/10 11:4 Belt=N

Hospital Record

Mary Smith Sanchez F 05/05/44 07/15/10 11:5 Hosp

Probability of a true match = 0.9836

UNIVERSITY OF UTAH SCHOOL OF MEDICINE

Probabilistic Record Linkage

Crash Record

Mary Smith) 11:47 Weber US5 Seat=1 Belt=N

Probability of a true match = 0.9817

Hospital Record

Mary Smith Sanch) 11:55 Weber Fracture Mem Hosp

Probabilistic Record Linkage

Crash Record

Mary Smith F Weber US5 Seat=1 Belt=N

Probability of a true match = 0.9999

Hospital Record

Mary Smith Sanchez F Weber Fracture Mem Hosp

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

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

University of Utah School of Medicine

Risk of Hospitalization or Death to the Teenage Driver

	Teens Odds Ratio	Adults Odds Ratio
Any passenger vs. alone	1.7 (1.4,2.2)	1.3 (1.2,1.4)
1 passenger vs. alone	1.6 (1.3,2.1)	1.3 (1.1,1.4)
$\geq$ 2 passenger vs. $\leq$ 1	1.6 (1.2,2.1)	1.2 (1.1,1.4)
$\geq$ 3 passenger vs. $\leq$ 2	1.7 (1.2,2.4)	1.1 (1.0,1.3)
$\geq$ 4 passenger vs. $\leq$ 3	1.9 (1.2,3.2)	1.3 (1.1,1.7)
$\geq$ 5 passenger vs. $\leq$ 4	2.5 (1.1,5.6)	1.8 (1.3,2.6)

UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

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
- ½ 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

UNIVERSITY OF UTAH SCHOOL OF MEDICINE

Repeat Patients to the Emergency Department

Unduplication of three-years of emergency department data

UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

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 ($\geq$ 5 visits) in year remained serial users the next year

UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

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

Utah KABCO

Pre 2006

- K – Fatal
- A – Broken bones & bleeding
- B – Bruises & abrasions
- C – Possible injury
- O – No injury

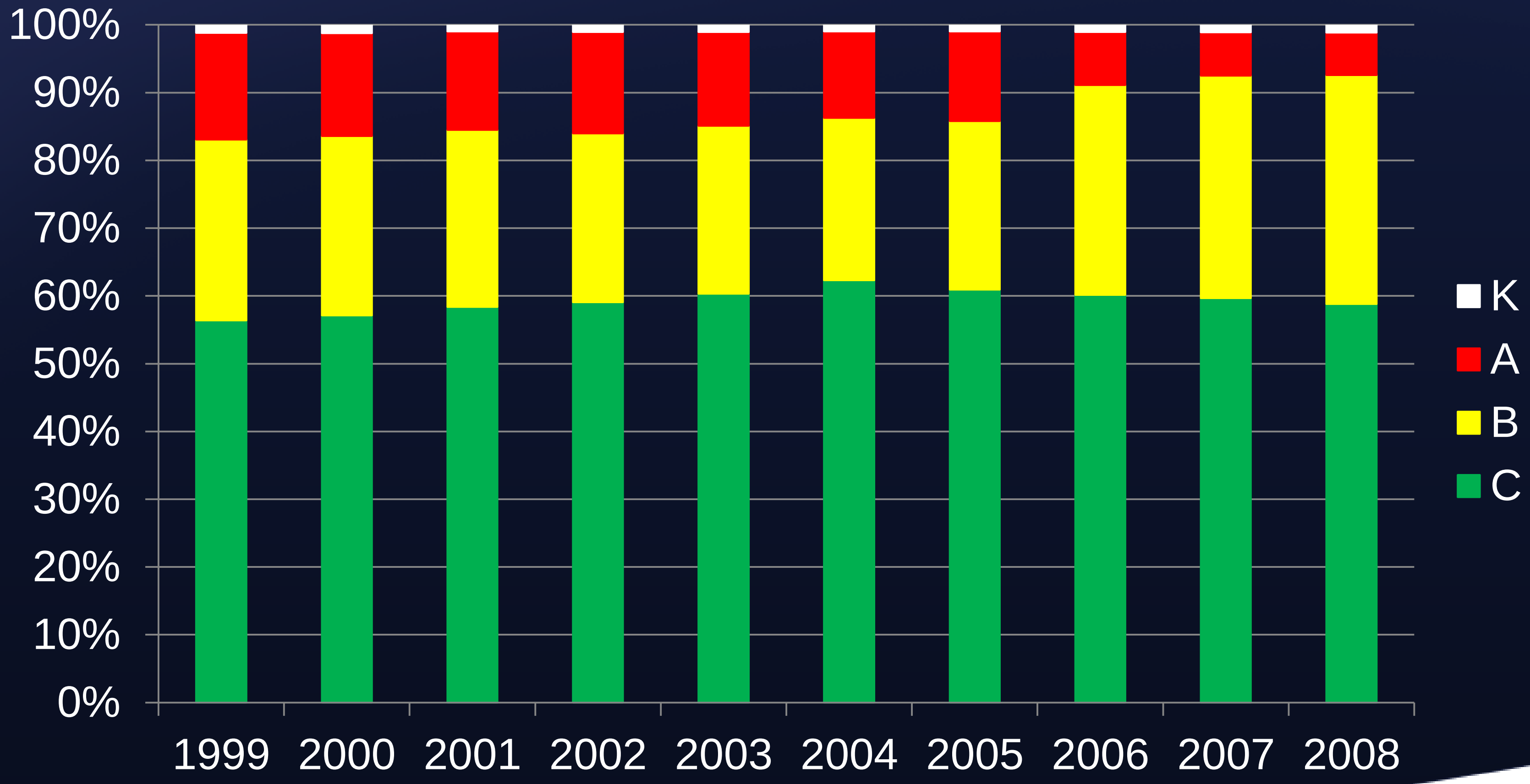
Post 2006

- K – Fatal
- A – Incapacitating injury
- B – Non-incapacitating injury
- C – Possible injury
- O – No injury

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?

Utah KABCO Data



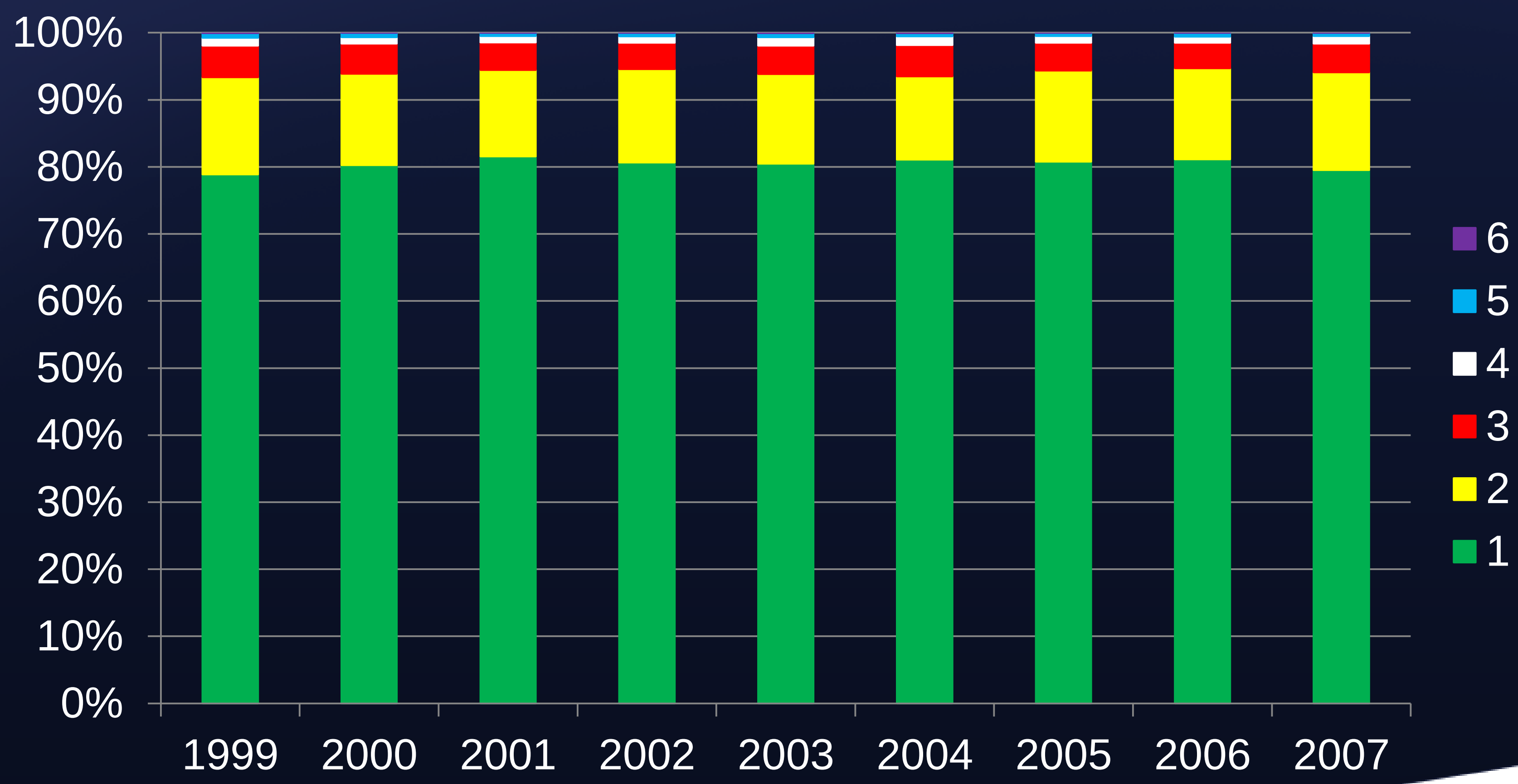
Can Hospital Files be Used to Measure Serious Injury Rates?

- 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

Utah MAIS Data



Summary

- Does wording on crash report matter?
  - KABCO distribution appears to change
  - MAIS remained more consistent
- Extend study to multiple states

Multi-State Analysis

UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

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

UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

Crash Severity of Injury

State A

- K – Fatal
- A – Incapacitated
- B – Visible Injury
- C – Momentary unconsciousness/ Complaint of pain
- O – No injury

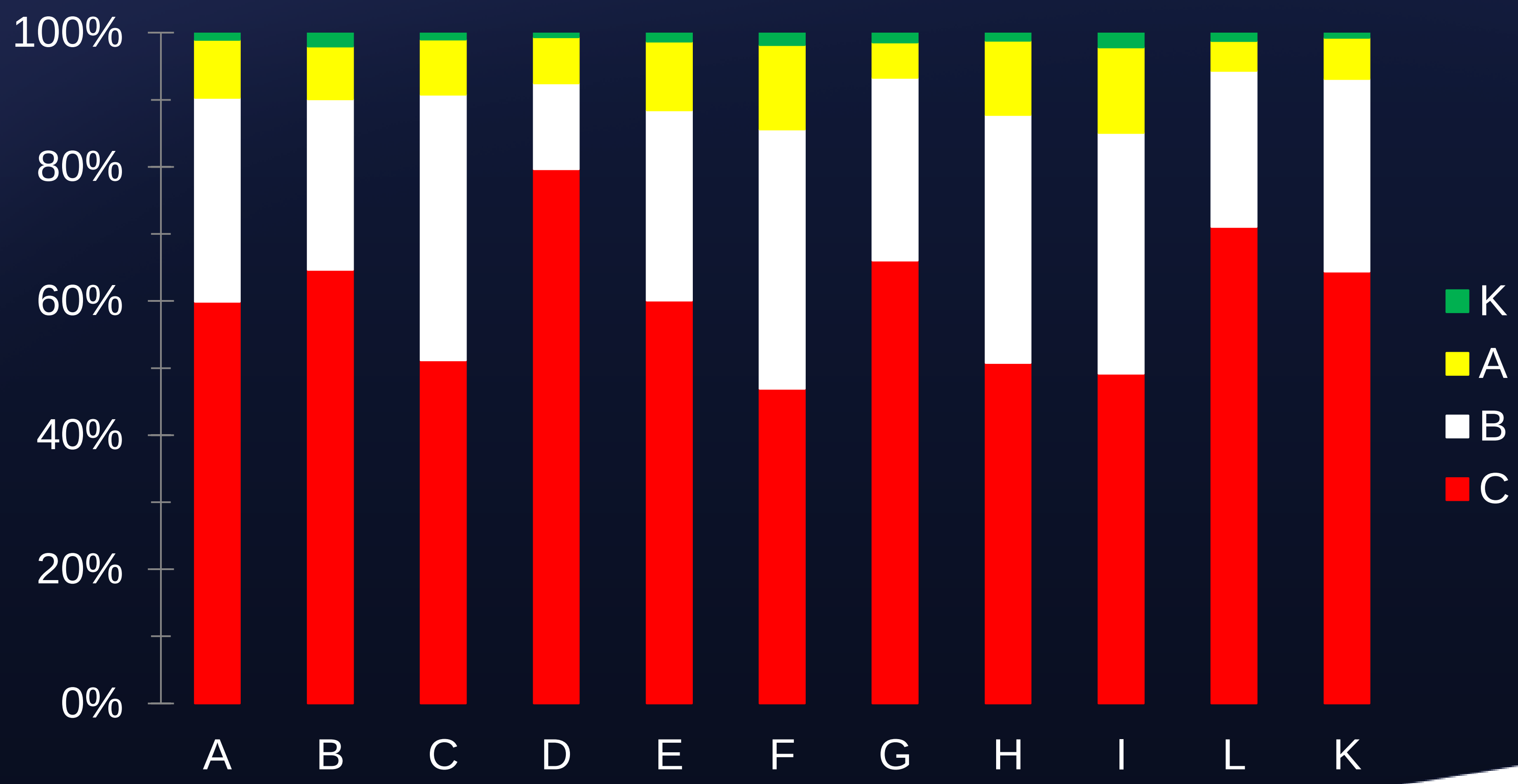
State B

- K – Fatal
- A – Life Threatening
- B – Serious
- C – Complaint of Pain
- O – No injury

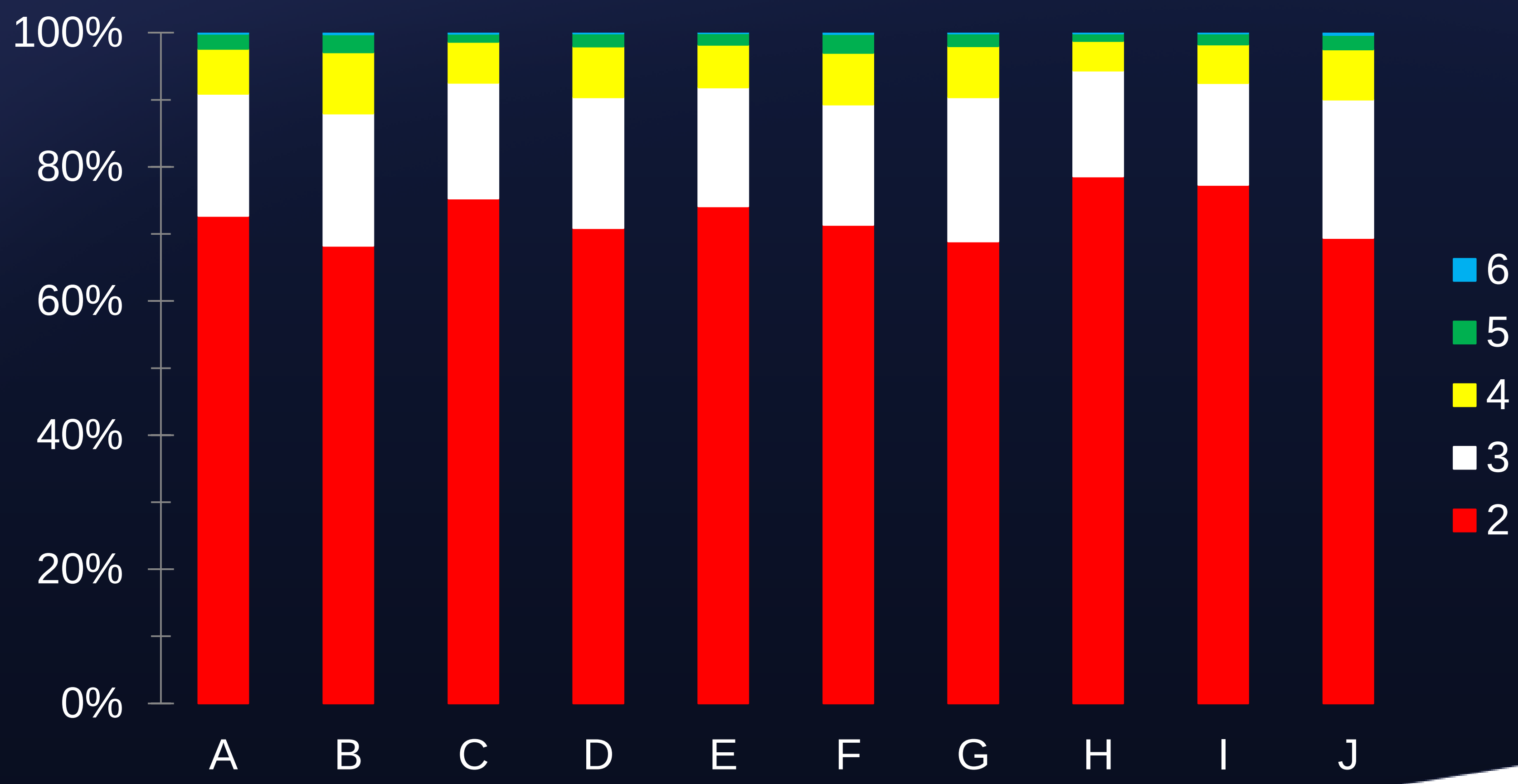
Methods

- 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



UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

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?

UNIVERSITY OF UTAH SCHOOL OF MEDICINE

Name Dilema

- Name 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

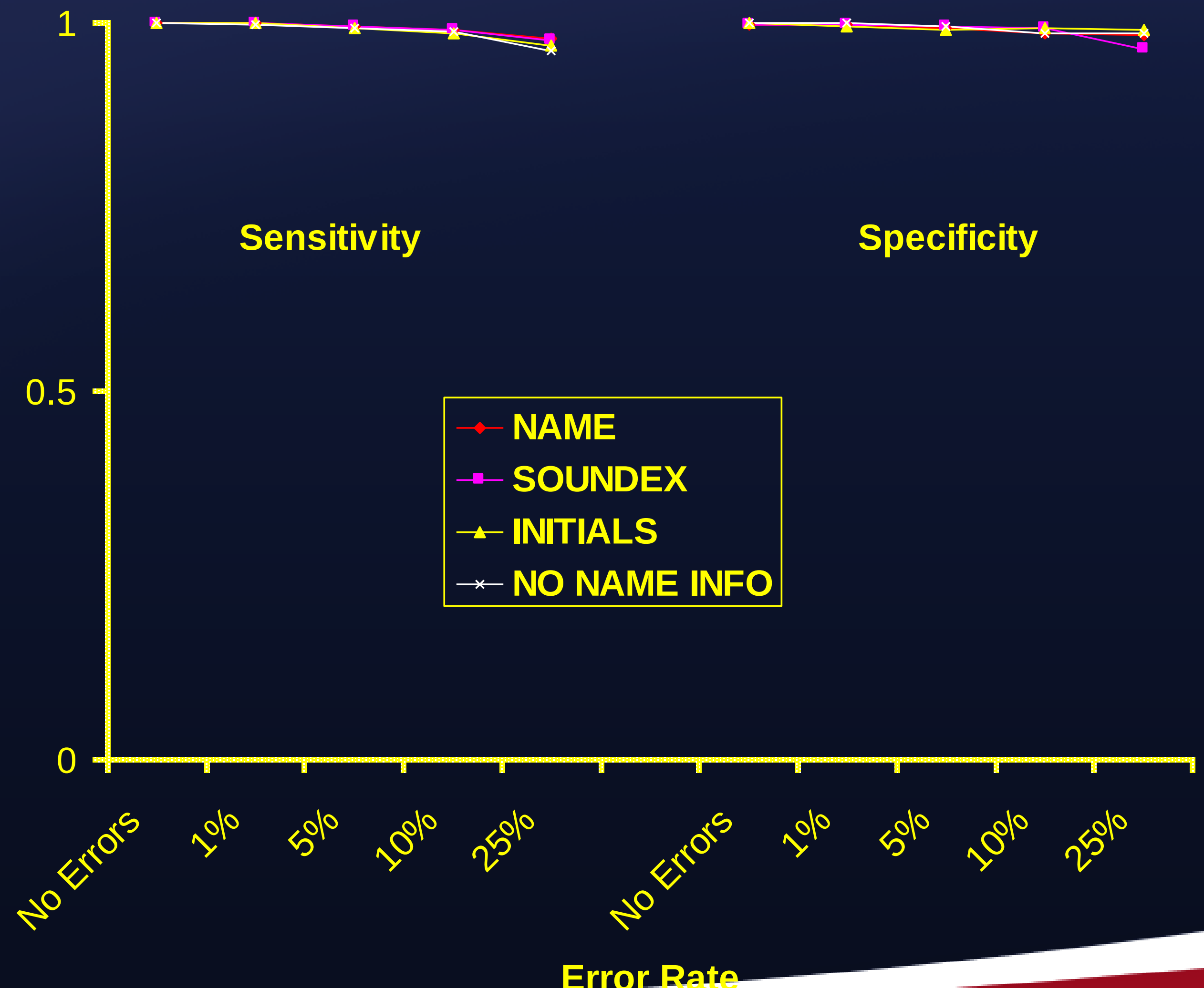
- Sensitivity - Ability to recognize true matches

  % of true matches identified

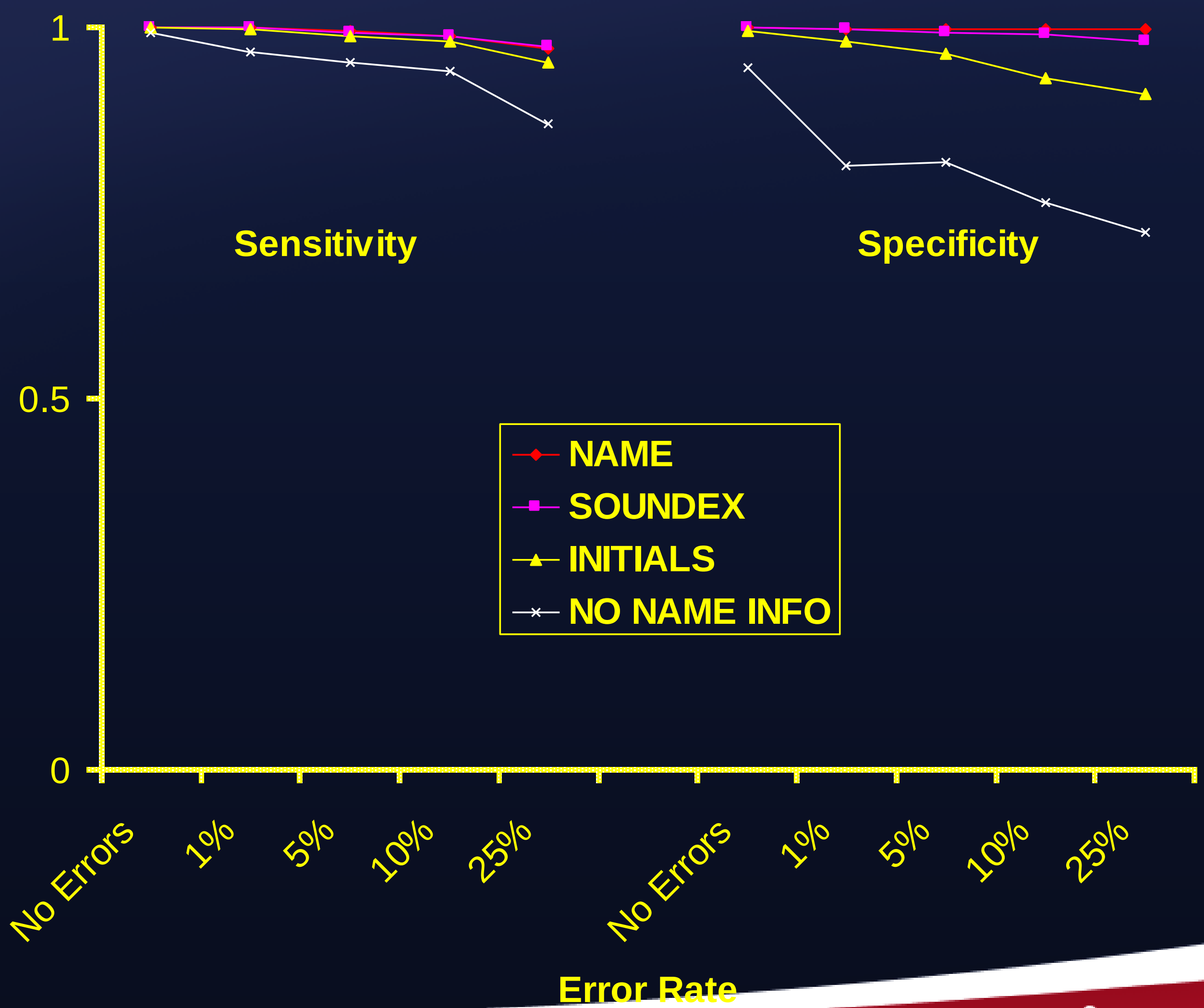
- Specificity - Ability to recognize incorrect matches

  1 – false positive rate

DOB, Gender, County, Time, Incident Date

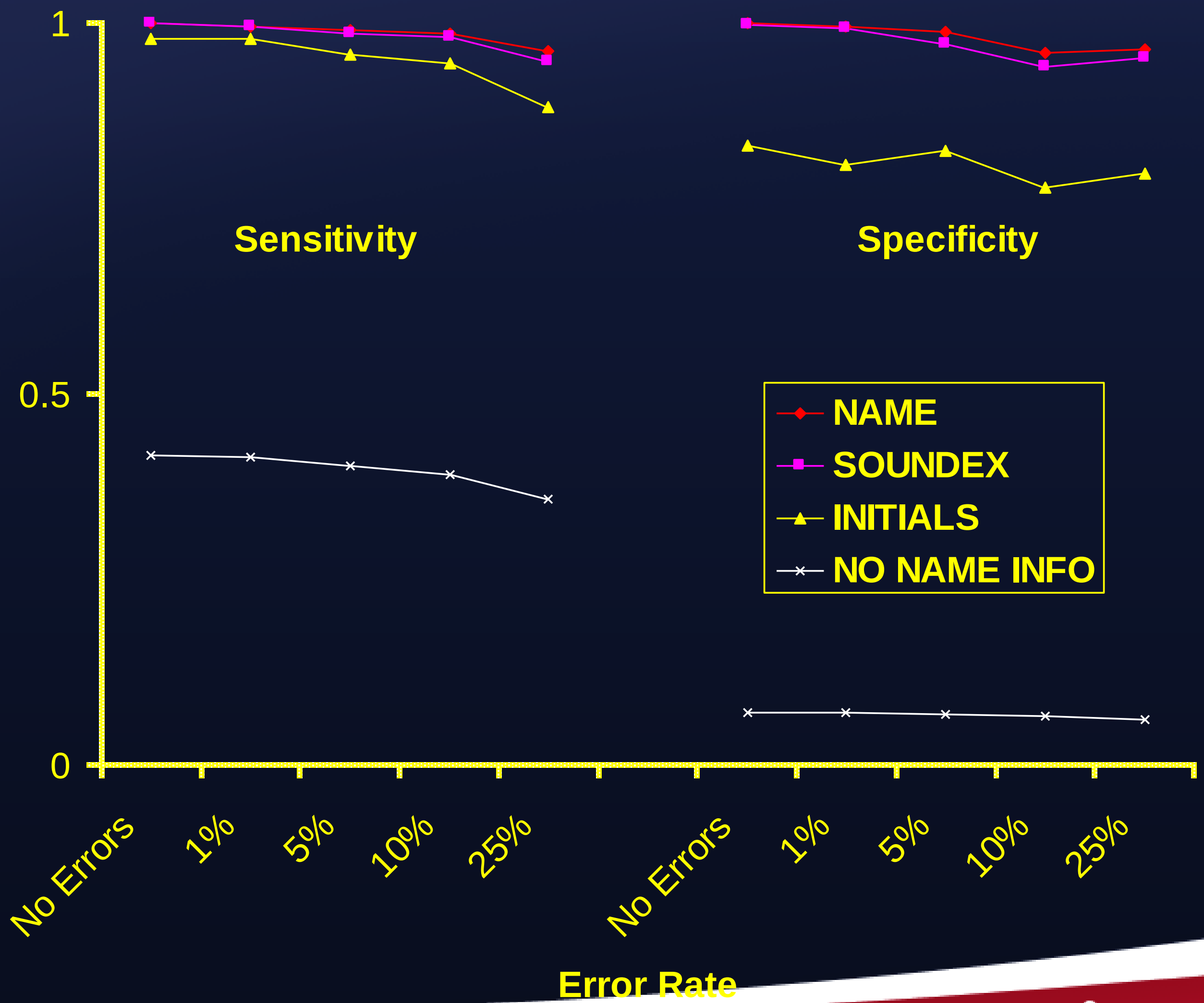


~~DOB~~, Age, Gender, ~~County~~, Time, Incident Date



UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

~~DOB~~, Age, Gender, ~~County~~, ~~Time~~, Incident Date



UNIVERSITY OF UTAH
SCHOOL OF MEDICINE

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
  - Cook LJ, Olson LM, Dean JM. (2001). Probabilistic record linkage: relationships between file sizes, identifiers and match weights. *Methods Inf Med*, *40*(3), 196-203.

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

UNIVERSITY OF UTAH SCHOOL OF MEDICINE

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

UNIVERSITY OF UTAH
SCHOOL OF MEDICINE