Using Probabilistic Linkage to Merge Multiple Data Sources for Monitoring Population Health

Wayne Bigelow, M.S.

Trudy Karlson, Ph.D.

Patricia Beutel, B.A.

Center for Health Systems Research & Analysis

University of Wisconsin – Madison

June, 1999

Please direct inquiries to: Wayne Bigelow, CHSRA, 610 Walnut Street, Madison WI 53705

Email: wayne@chsra.wisc.edu

Background

- The National Highway Transportation Safety Agency wanted to study the health outcomes and costs associated with vehicular crashes in greater detail than had been previously possible.
- NHTSA decided to link together state specific crash data with already existing health data (hospital discharge, ambulance, emergency department) under a program called CODES.
- Given limits in data elements available to merge these different data sources, NHTSA is using a technique called probabilistic linkage to merge them. The software used to perform the linkage is AUTOMATCH.

Goal:

Merging data sources from Wisconsin DOT Crash data to Wisconsin hospital discharge data under NHTSA sponsored Crash Outcomes Data Evaluation Systems Project (CODES).

Problem:

No person level identifiers available (e.g. SSN, Name, Address) for linking crash records to hospital records.

General Problem:

- Linking records between two data sets when one or both do not have person level identifiers.
- Linking records when there is incorrect or missing data for person level identifiers.
- However, some information, such as sex, age/birthdate, date(s) of event, county and zip code may be available.
- Problem is common to a wide range of research: Outcomes, epidemiologic, quality assurance and financial.

Probabilistic Record Linkage

- Links records between 2 data sets through the calculation of linkage likelihood or probability weights, adjusting for incomplete and missing data.
- Likelihood/probability weights are estimated given all observed agreements and disagreements on all data elements used for linking records together.
- Probabilistic linkage incorporates variable levels of discriminatory power and reliability within specific linkage elements.

Linkage Weights (1)

M(i) = Reliability

- Probability that linkage element (i) agrees on a true matched pair.
- Approximately = (1 - error rate)
- Analogous to “sensitivity”
- Determined by initial manual review of data, or through previous research.

Linkage Weights (2)

U(i) = Discriminatory Power:

- Probability that linkage element (i) agrees on an unmatched pair.
- Approximately = (1 / number of values)
- Analogous to “specificity”
- In **AUTOMATCH**, ***U*** must be initially set. Software later generates ***U***s from actual frequency of different values in the two data sets.

Linkage Weights (3)

Linkage Weight for a match on a given element = $(M(i) \,/\, U(i))$

Linkage Weights for a non-match = $(1 - M(i) \,/\, 1 - U(i))$

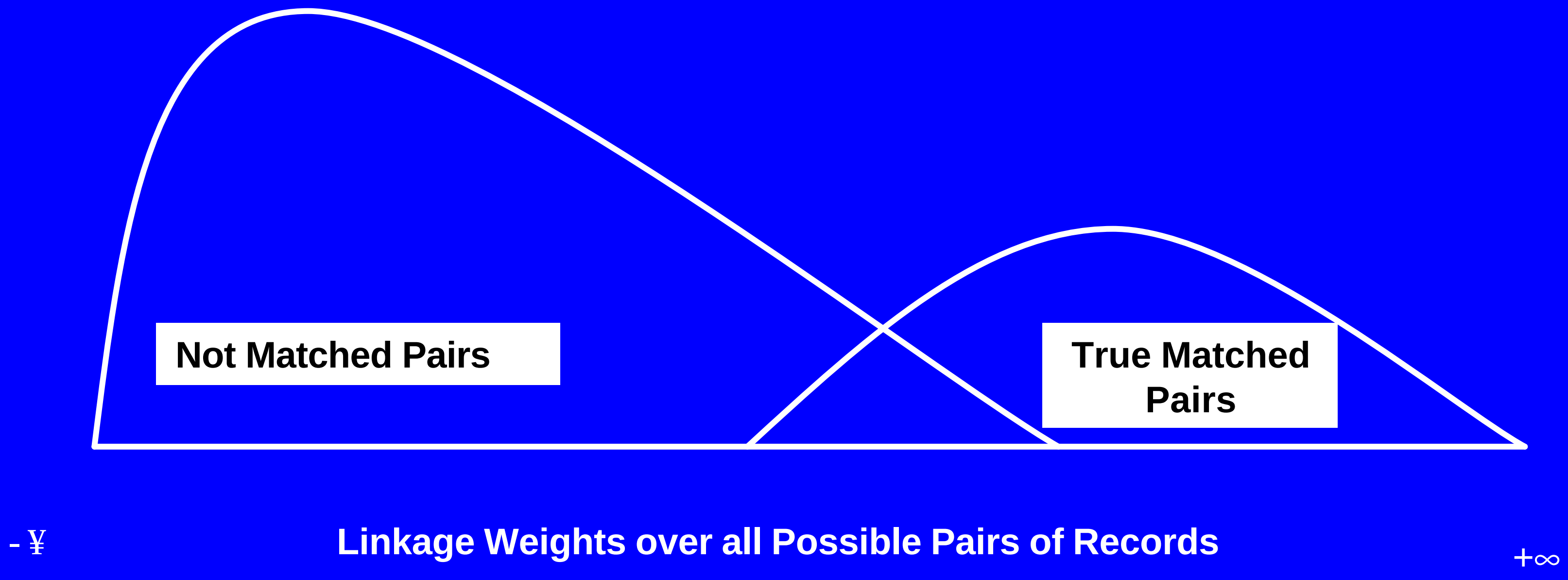
Total Linkage Weight for a Record Pair =

LOGN (Multiplicative sum of all linkage weights for a given record pair
times
the odds of a random true match between 2 data sets)

- Linkage weights measure how much data elements improve our ability to match two records in addition to the likelihood of a random true match.
- Linkage weights are negative if the data element(s) don't match, and are positive if they do match.

Linkage Weights (4)

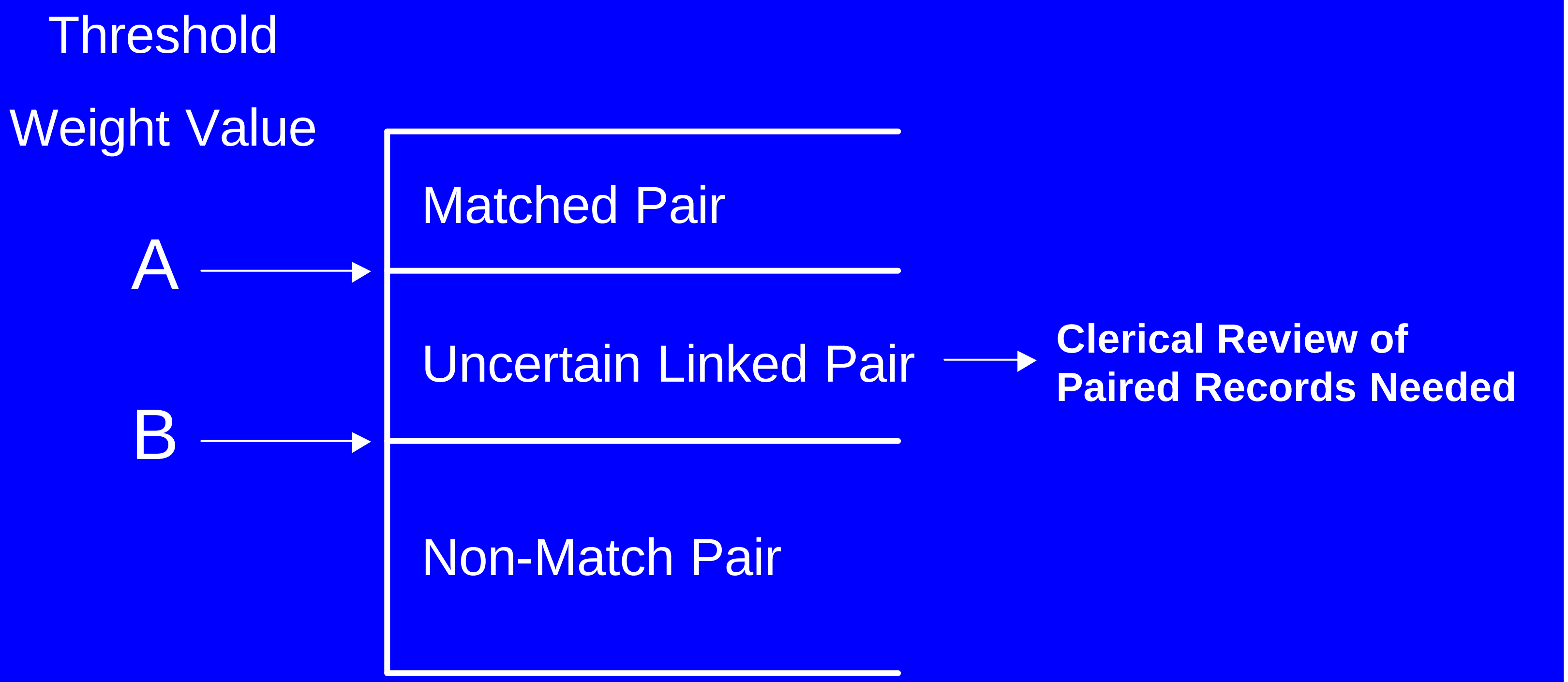
Linkage Weights are typically distributed:



Selecting Matched Records Using Probabilistic Linkage (1)

- Match occurs when the total record pair linkage weight is greater than threshold value A.
- Non-Match occurs when the total record pair linkage weight is less than threshold value B.
- Uncertain Linkage occurs when the total record pair linkage weight is between Value A and Value B. Further clerical review is needed.

Selecting Matched Records Using Probabilistic Linkage (2)



Threshold

Weight Value

A

B

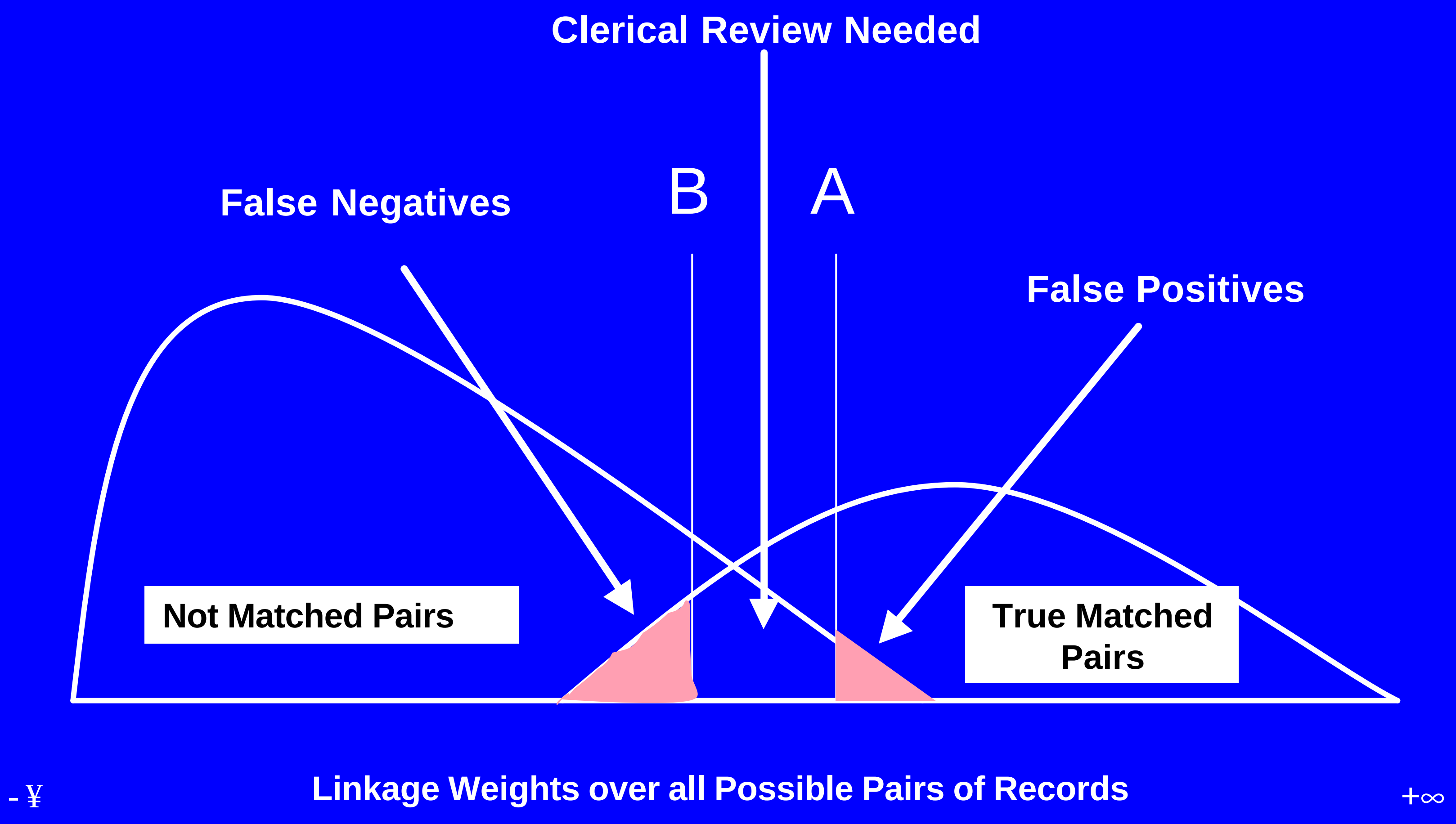
Matched Pair

Uncertain Linked Pair

Non-Match Pair

Clerical Review of Paired Records Needed

Selecting Matched Records Using Probabilistic Linkage (3)



Clerical Review Needed

False Negatives

B

A

False Positives

Not Matched Pairs

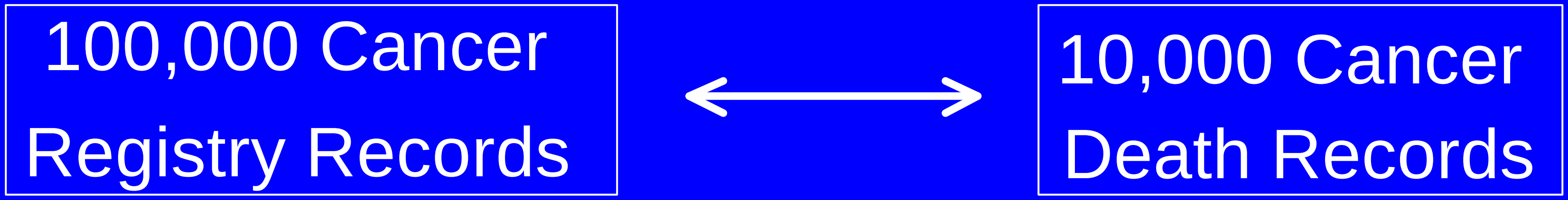
True Matched Pairs

- ¥

Linkage Weights over all Possible Pairs of Records

+∞

Example (1)



100,000 Cancer Registry Records

10,000 Cancer Death Records

- Assume all 10,000 death records will have a corresponding registry record.
- The odds for a match at random for any 2 records is 1:99,999

Example (2)

Elements used to link Cancer registry records and death records

	M(i)	U(i)
Sex	.999	.50
Date of Birth	.999	.001
Last Name	.999	.01
Type of Cancer	.90	.05
Zip Code of Residence	.99	.05

Example (3)

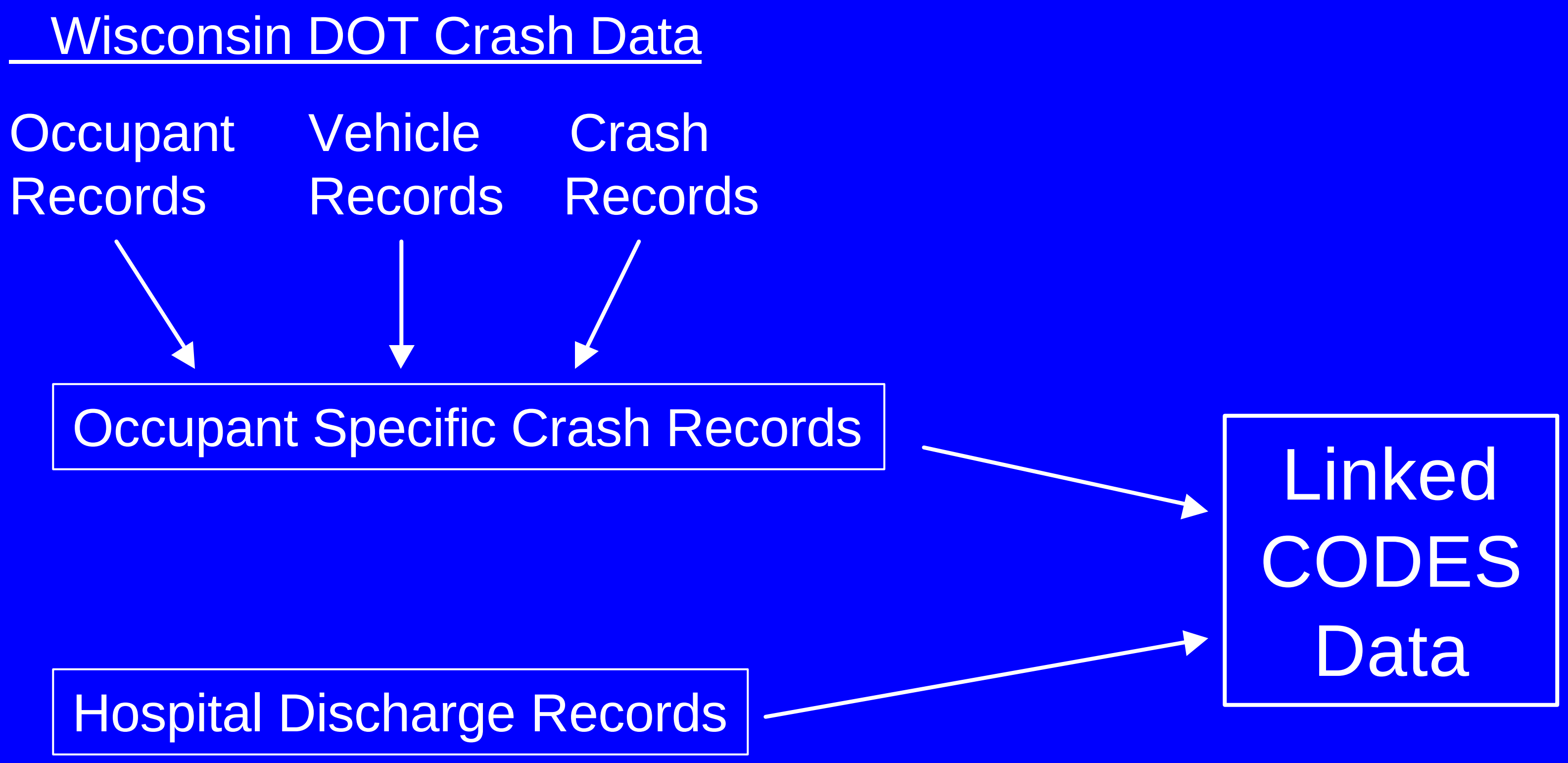
Odds of Random Match = 1/99,999 = .00001

❖ Match on Sex:	.999 / .50	≈	2
❖ Match on Date of Birth:	.999 / .001	≈	999
❖ Match on Last Name:	.999 / .01	≈	100
❖ Match on Cancer Type:	.90 / .10	≈	9
❖ Match on Zip Code:	.99 / .02	≈	50

- ❖ Multiplicative Sum * Random Odds ≈ 899.1

- ❖ LOG$_N$ (899.1) = 9.8123

Merging 1996 Wisconsin Crash and Hospital Discharge Data



Wisconsin DOT Crash Data

Occupant Records

Vehicle Records

Crash Records

Occupant Specific Crash Records

Hospital Discharge Records

Linked CODES Data

Process:

- AUTOMATCH software used.
- Creation of blocks of records which match on at least one variable for further linkage (for computational efficiency).
- Estimate ***M*** and ***U*** for each linkage data element.
- Determine threshold weight values for linked, not linked and uncertainly linked pairs of records.
- Generate total weight for potential record linkage pairs.

Elements Used to Link Wisconsin CODES Data

- Injury diagnoses used for initial selection of hospital discharge records for possible linkage.

Data Elements used to link records:

Sex

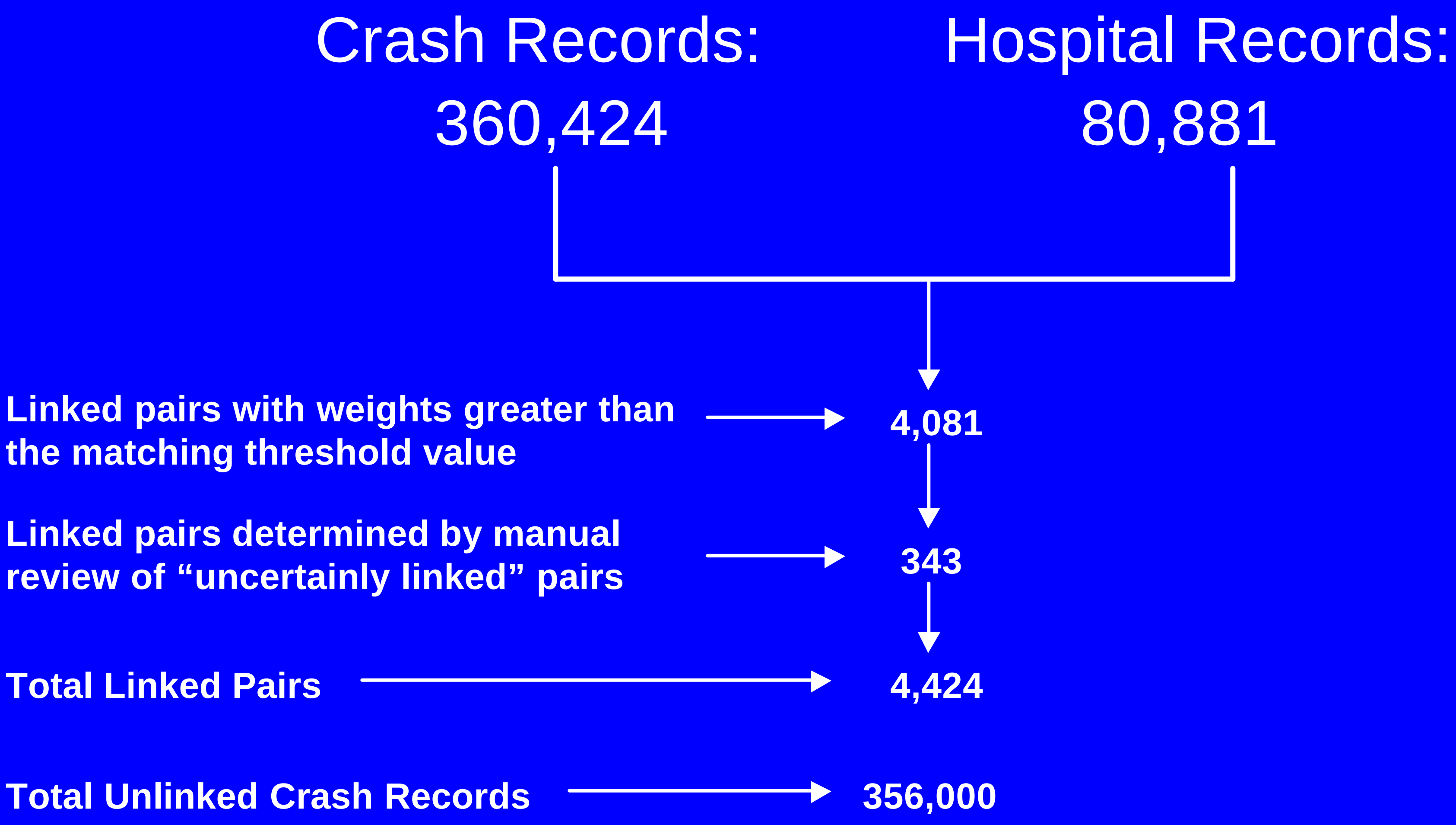
Age / Date of Birth

County of Accident/Hospitalization

Zip Code of Residence

Date of Crash/Date of Hospitalization

Results of 1996 Crash and Hospital Data Linkage



Crash Records: 360,424

Hospital Records: 80,881

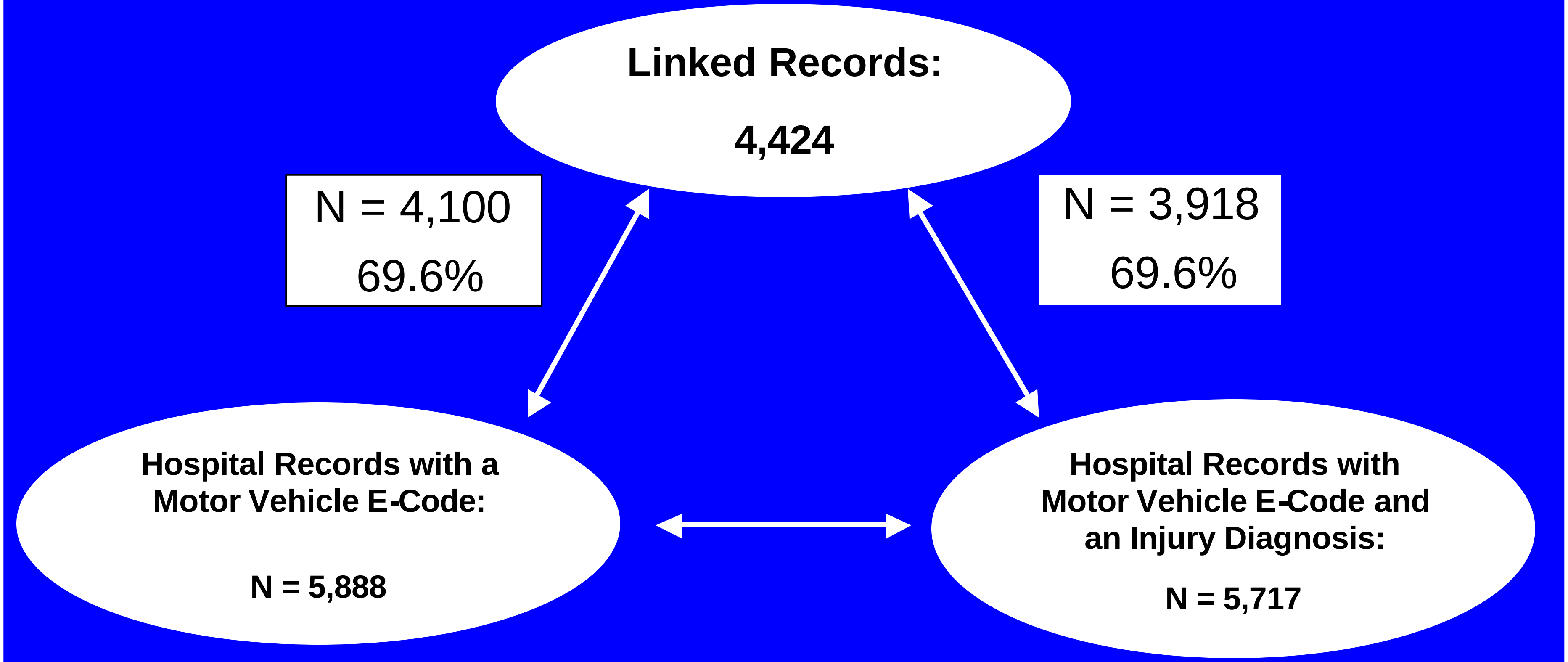
Linked pairs with weights greater than the matching threshold value → 4,081

Linked pairs determined by manual review of “uncertainly linked” pairs → 343

Total Linked Pairs → 4,424

Total Unlinked Crash Records → 356,000

Comparison of Linked Data Results to Hospital Motor Vehicle Injury Information



Linked Records:

4,424

N = 4,100

69.6%

N = 3,918

69.6%

Hospital Records with a Motor Vehicle E-Code:

N = 5,888

Hospital Records with Motor Vehicle E-Code and an Injury Diagnosis:

N = 5,717

Conclusions

- Almost 70% of Hospital Cases with an E-Code indicative of a motor vehicle injury were matched to vehicle crash information.
- After accounting for crash related hospital admissions occurring long after the crash occurred, upwards of 80% of all crash related hospitalizations were linked to DOT crash information.
- Probabilistic linkage provides a statistically sound method of linking multiple data sources in the absence of person level identifiers and missing information.
- Probabilistic linkage offers health services researchers and epidemiologists the opportunity to more cheaply and effectively perform research by utilizing existing data through record linkage.

Other Sources of Information on Record/Probabilistic Linkage

- *A Theory for Record Linkage*, I.P.Felligi and A.B.Sumter; Journal of the American Statistical Association; 1969
- *Textbook of medical record linkage*; Edited by J.A.Baldwin et al.; Oxford University Press; Oxford/New York; 1987.
- *Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa*; Florida, Matthew Jaro; Journal of the American Statistical Association; 1989
- http://www.matchware.com