VALID EPIDEMIOLOGIC METHODS AND STUDIES BASED ON LINKED DATA:
ARE THEY COMPATIBLE?

Jack K. Leiss, PhD
Chief Epidemiologist
Constella Health Sciences
jleiss@constellagroup.com

American College of Epidemiology
Annual Scientific Sessions
Chicago, September 6-9, 2003

- What is record linkage?
- Linkage methods and epidemiologic investigation
- Solutions

National Cancer Institute

1 R43 CA88757-01

National Center for Health Statistics

UR6/CCU417428-01

National Institute of Child Health and Human Development

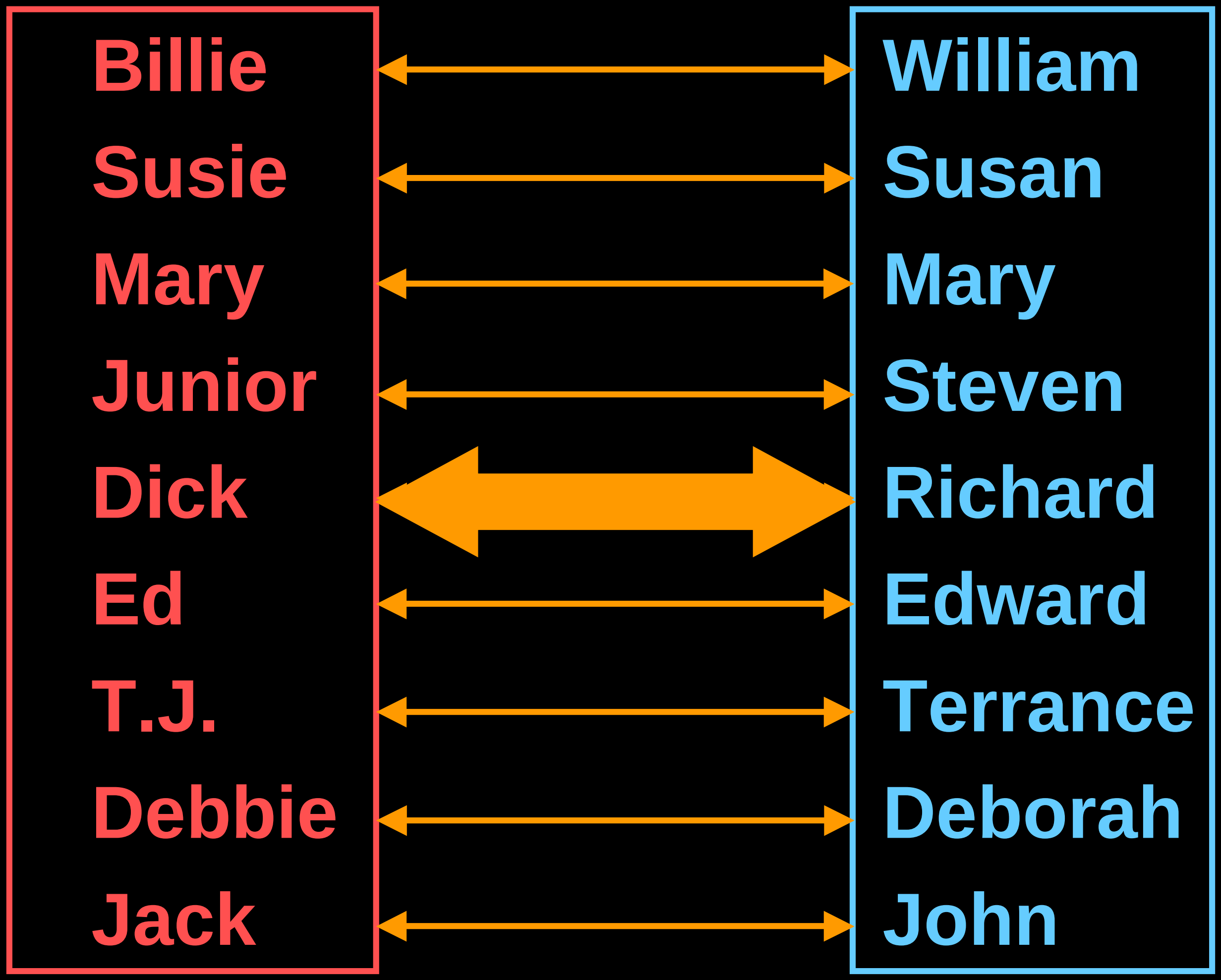
1 R43 HD35785-01A1

2 R44 HD35785-02A1

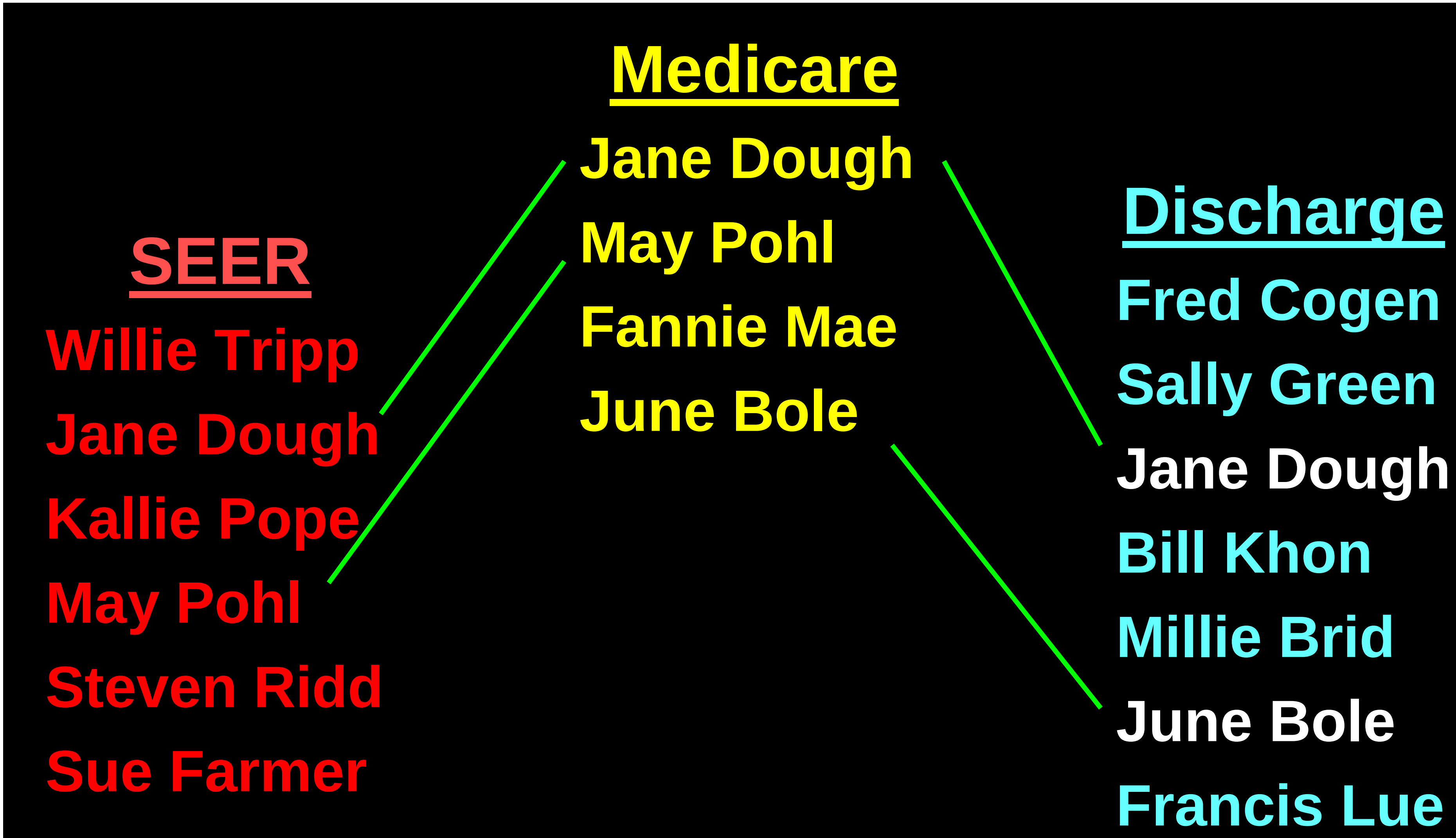
Mike McGlincy, PhD

Strategic Matching, Inc.

What is record linkage?



Billie	William
Susie	Susan
Mary	Mary
Junior	Steven
Dick	Richard
Ed	Edward
T.J.	Terrance
Debbie	Deborah
Jack	John



SEER

Willie Tripp

Jane Dough

Kallie Pope

May Pohl

Steven Ridd

Sue Farmer

Medicare

Jane Dough

May Pohl

Fannie Mae

June Bole

Discharge

Fred Cogen

Sally Green

Jane Dough

Bill Khon

Millie Brid

June Bole

Francis Lue

Subject	Data		
Jane Dough	SEER	Medicare	Discharge
May Pohl	SEER	Medicare	…
June Bole	SEER	…	Discharge

OR = 3.1 (2.4 – 4.7)

SEER

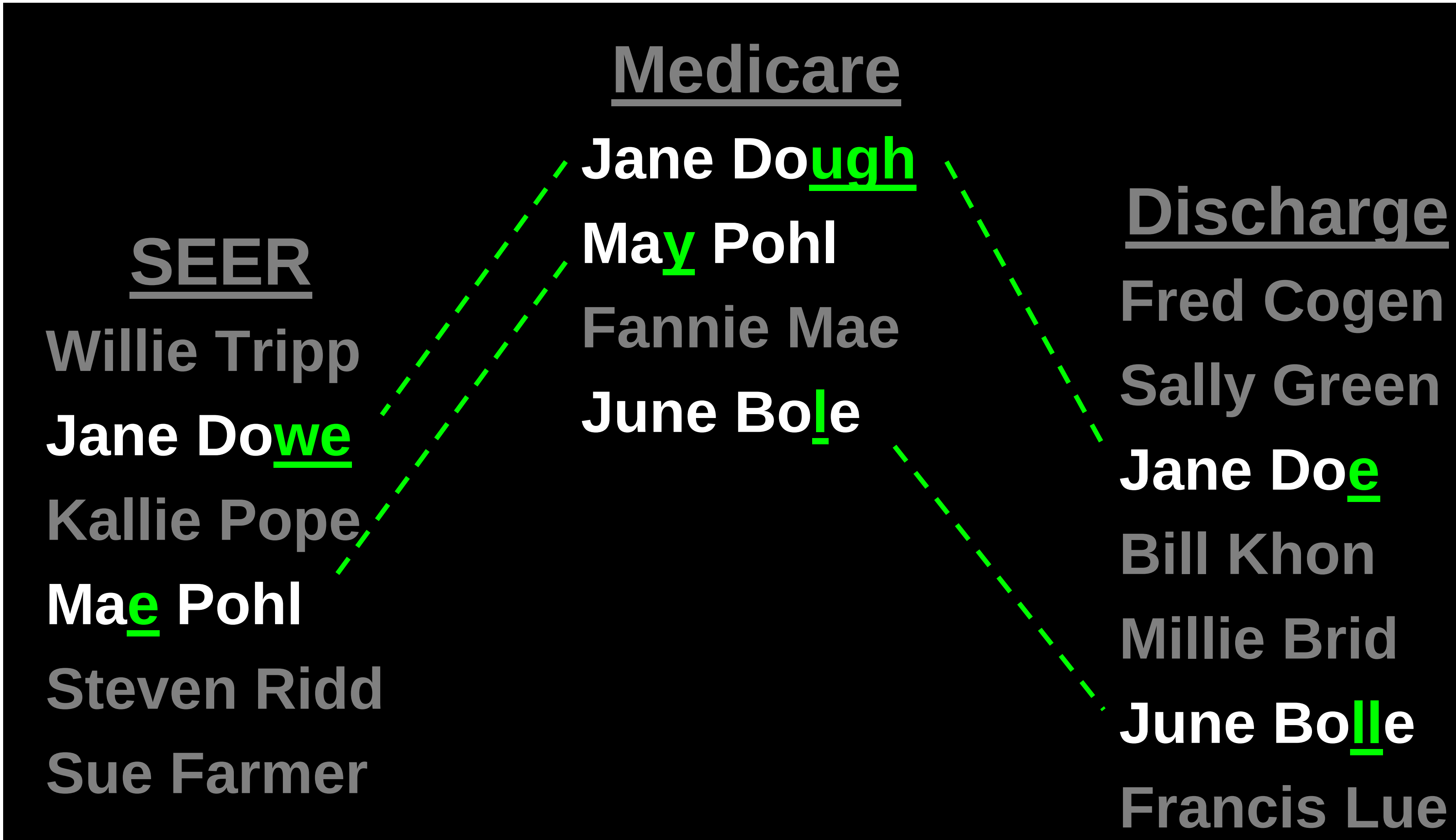
Willie Tripp
Jane Dowe
Kallie Pope
Mae Pohl
Steven Ridd
Sue Farmer

Medicare

Jane Dough
May Pohl
Fannie Mae
June Bole

Discharge

Fred Cogen
Sally Green
Jane Doe
Bill Khon
Millie Brid
June Bolle
Francis Lue



SEER

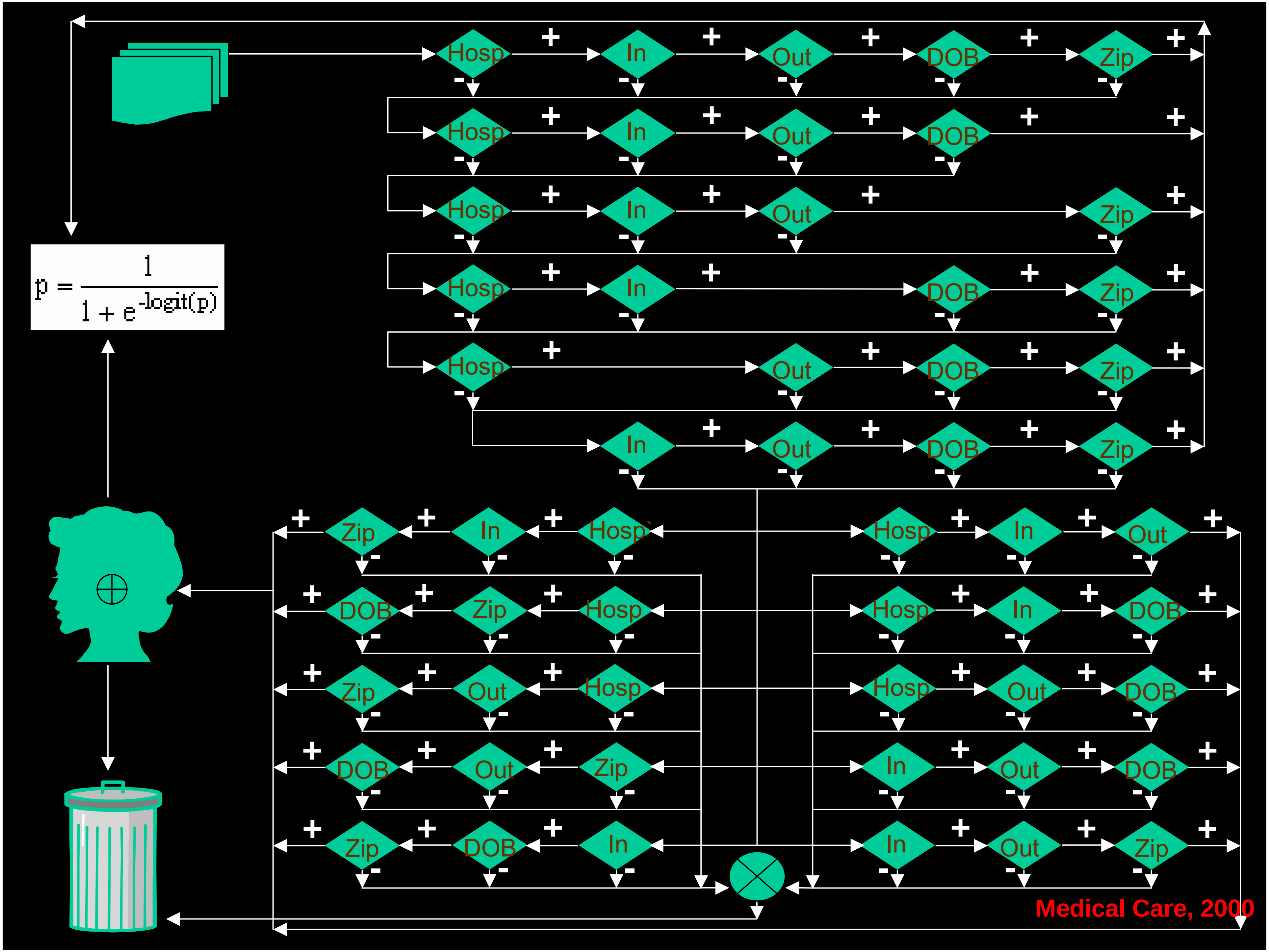
Willie Tripp
Jane Dowe
Kallie Pope
Mae Pohl
Steven Ridd
Sue Farmer

Medicare

Jane Dough
May Pohl
Fannie Mae
June Bole

Discharge

Fred Cogen
Sally Green
Jane Doe
Bill Khon
Millie Brid
June Bolle
Francis Lue



$$p = \frac{1}{1 + e^{-\mathrm{logit}(p)}}$$

Medical Care, 2000

Once

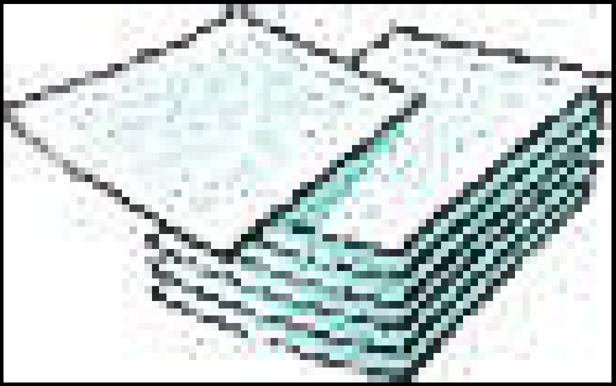
Upon

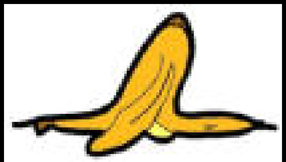
A

Time

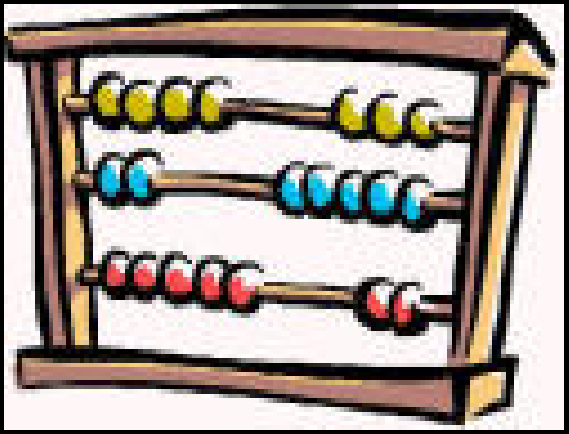
→

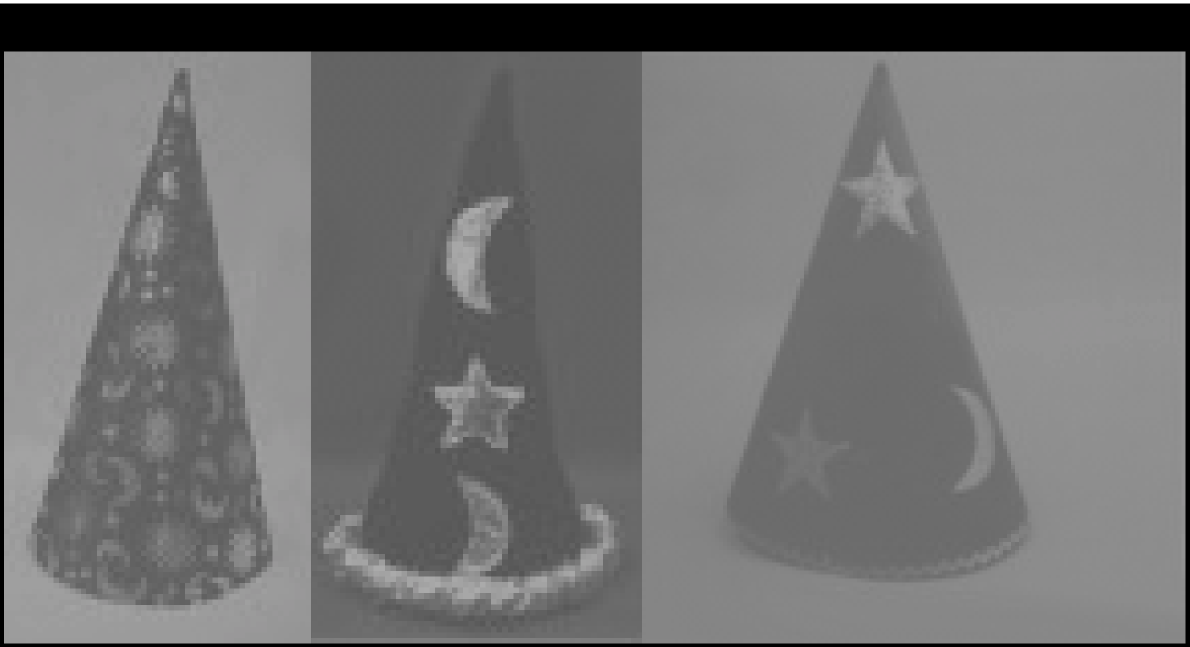
www.yesicankids.gov





www.pretendware.com
www.costumezone.com
www.partydomain.co.uk





Linkage Errors: Unmet Epi Goals

- Minimize bias
- Account for bias
- Maximize precision
- Adjust precision

Linkage Methods vs. Epi Methods

- Goals of linkage vs. goals of study
- Data management vs. data collection
- Logic vs. probability

Goals of Linkage vs. Goals of Study

“The goal ... was to obtain only cases [that were] an accurate match.”

Medical Care, 1999

Valid, precise estimates

Goals of Linkage vs. Goals of Study

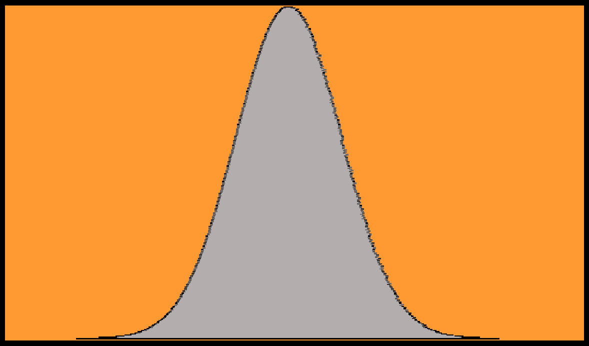
- Individuals vs. Populations
- Accuracy vs. validity

Individuals vs. Populations

"… the matched database represents valid matches

and is representative of the larger population."

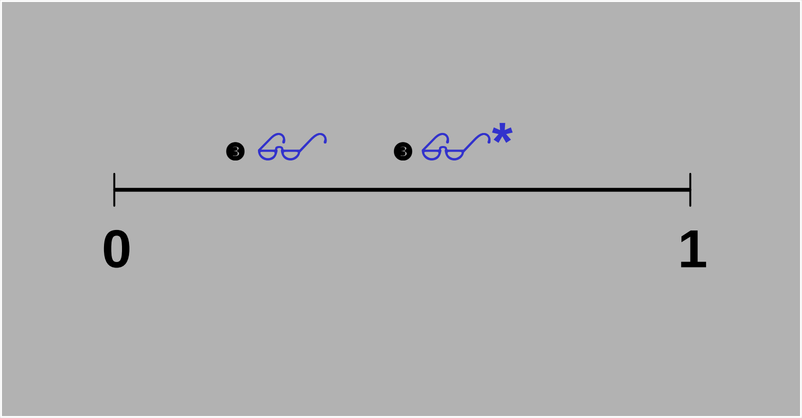
Medical Care, 2000



Accuracy vs. Validity

"We reviewed the … links and excluded links lacking face validity."

Medical Care, 2000



Goals of Linkage vs. Goals of Study

- Individuals vs. Populations
- Accuracy vs. validity

Data Management vs. Data Collection

“Patients who matched on [specified] variables … were considered valid matches.”

Cancer, 2001

- Minimize bias
- Estimate bias
- Maximize precision
- Estimate precision

Data Management vs. Data Collection

"Patients who matched on [specified] variables … were considered valid matches."

Cancer, 2001

"… partial matches … were reviewed independently … and then discussed to reach consensus about whether a correct match had occurred."

Medical Care, 1999

Logic vs. Probability

The set of true links can be known.

"… the … goal … was to retain only those cases [with] a **very high likelihood** of an accurate match."

Medical Care, 1999

```
if HOSP=1 & IN=1 & OUT=1 & DOB=1 & ZIP=1 then LINK=1;
   else if HOSP=1 & IN=1 & OUT=1 & DOB=1 then LINK=1;
   else if HOSP=1 & IN=1 & OUT=1 & ZIP=1 then LINK=1;
   else if HOSP=1 & IN=1 & DOB=1 & ZIP=1 then LINK=1;
   else if HOSP=1 & OUT=1 & DOB=1 & ZIP=1 then LINK=1;
   else if IN=1 & OUT=1 & DOB=1 & ZIP=1 then LINK=1;
if LINK=0 then do;
   if HOSP=1 & IN=1 & OUT=1 then LINK=1;
    if HOSP=1 & IN=1 & ZIP=1 then LINK=1;
    if HOSP=1 & IN=1 & DOB=1 & ZIP=1 then LINK=1;
    if HOSP=1 & OUT=1 & ZIP=1 then LINK=1;
    if HOSP=1 & OUT=1 & DOB=1 then LINK=1;
    if HOSP=1 & DOB=1 & ZIP=1 then LINK=1;
    if IN=1 & OUT=1 & DOB=1 then LINK=1;
    if IN=1 & OUT=1 & ZIP=1 then LINK=1;
    if IN=1 & DOB=1 & ZIP=1 then LINK=1;
    if OUT=1 & DOB=1 & ZIP=1 then LINK=1;
end;
```

Medical Care, 2000

Logic vs. Probability

The set of true links can be known.

If it doesn’t look good, then LINK=0;

Linkage Methods vs. Epi Methods

- Goals of linkage vs. goals of study
- Data management vs. data collection
- Logic vs. probability

Solution: Probabilistic Record Linkage

- Statistical theory
- Probability distributions
- Unbiased method
- Adjust for bias
- Calculate precision
  - Estimate error rates
  - Differential error rates

Probabilistic Record Linkage in Theory

$m \equiv \Pr(\gamma \mid (a, b) \text{ in } M) = \Pr(\gamma \mid M)$

$u \equiv \Pr(\gamma \mid (a, b) \text{ in } U) = \Pr(\gamma \mid U)$

$\text{Match Weight} = \log(m/u)$

Fellegi & Sunter, JASA, 1969.

Probabilistic Record Linkage in Practice

- Jaro, Stat Med, 1995
- ~~Fellegi & Sunter~~ Heuristic
- Weights, not probabilities
- “Set of true links”
- Discard lower end of distribution
- Clerical review
- Dependent fields
- Missing values

Solutions: Investigators

- LinkSolv (mcglincym@strategicmatching.com)
- Return to Fellegi & Sunter
- Linkage as probabilistic process
- Match probabilities
- Multiple imputation of linkages
- Dependent fields

Solutions: Publishers and Reviewers

- Full description of linkage method
- Authors report linkage error rates
- Challenge internal validation
- Authors address impact of linkage errors on results: bias, precision
- Quantification, adjustment

Solutions: Funding Agencies

- Measures
  - Linkage error
  - Differential linkage error
- Techniques
  - Adjust for nondifferential linkage error
  - Account for differential linkage error
  - Include added variance in precision calculations

Are they compatible?

✓ Linkage: Truly probabilistic

✓ Results:

- ✓ Quantify
- ✓ Adjust
- ✓ Interpret

www.pretendware.com
www.costumezone.com
www.partydomain.co.uk

Solutions: Funding Agencies

- Overall 95%
- Younger African-American women 80%
- Older white women 60%
- Specific histological types 50%

“The quality of linkages was examined by calculating the percentage of linked individuals who also shared the same date of birth (92.5%).”