

# Probabilistic Record Matching and Deduplication Using Open Source Software

Immunization Registry Conference  
Atlanta, GA  
October 19, 2004

Magaly Angeloni  
Rhode Island Department of Health  
[www.health.ri.gov](http://www.health.ri.gov)

Mike Berry  
HLN Consulting, LLC  
[www.hln.com](http://www.hln.com)

# Agenda

- Discussion of the Problem
- Deterministic/Probabilistic Matching
- Matching Architecture
- Software
- KIDSNET Background
- Records on Hold
- KIDSNET Matching Results

# Discussion of the Problem

- What is Data Linkage? Record Linkage? Matching? Deduplication?
- The bringing together of records of information concerning an individual using demographic information.
- Used to support public health surveillance, epidemiological research, health services research, and program planning.

*(Source: CDC NCCDPHP)*

# Deterministic Matching

- Rule-based
- Exact matches; sets of rules
- Apply rules based on common errors, nicknames, abbreviations, etc.
- Experience with the data helps
- Simple, efficient, fast
- Successful when: high quality data or unique identifiers
- Less successful when: incomplete or inaccurate data; spelling or transcription errors; nicknames; name changes; etc.
- Sometimes components of probabilistic matching are utilized in deterministic schemes

# Probabilistic Matching

- Estimate the probability that two records are the same person vs. not the same person based on a degree of match on selected fields.
- Define a probability level above which all pairs are assumed to be the same person. Define a probability level below which all pairs are assumed **not** to be the same person.
- Send pairs that fall between these two levels to “human review.”

# Probabilistic Matching (cont.)

- Pre-processing/Standardization
- Blocking
- String Comparison
- Frequency Analysis
- Probability Scoring, Assignment
- Human Review

Not a match

Human Review

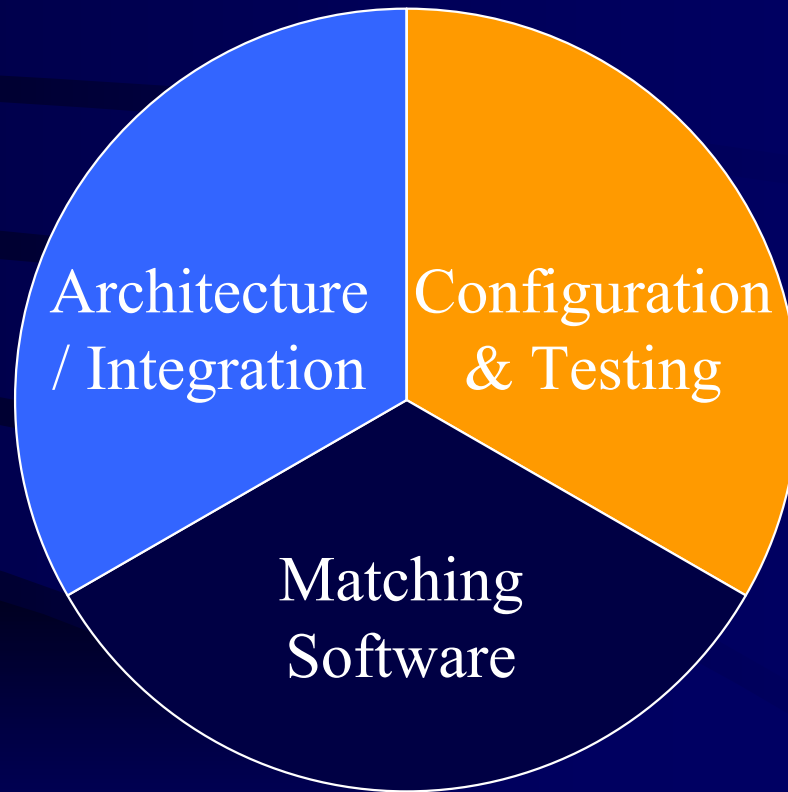
Definite Match

# Probabilistic Matching (cont.)

Fellegi-Sunter method for computing probabilities:

- $m$  – probability that a field is a match given that a pair is a match
- $u$  – probability that a field is a match given that a pair is NOT a match

# Implementing Matching



# Software - Commercial

- Ascential
- AutoMatch
- ChoiceMaker
- LinkSoft
- Madison
- MatchWare
- Search Software America (SSA)
- Many others...

# Software – Non-commercial

- AJAX (INRIA/NYU)
- Census Bureau (Winkler, Jaro)
- Febr1 (ANU / NSW DOH)
- GRLS (Statistics Canada)
- Link Plus (CDC)
- OX-Link (Oxford)
- TAILOR (Purdue/Drexel)

# Febrl – Freely Extensible Biomedical Record Linkage

## Pros

- Open Source / Free
- Sophisticated data standardization
- Fellegi-Sunter implementation
- Fast
- Many string comparators
- New features to come

## Cons

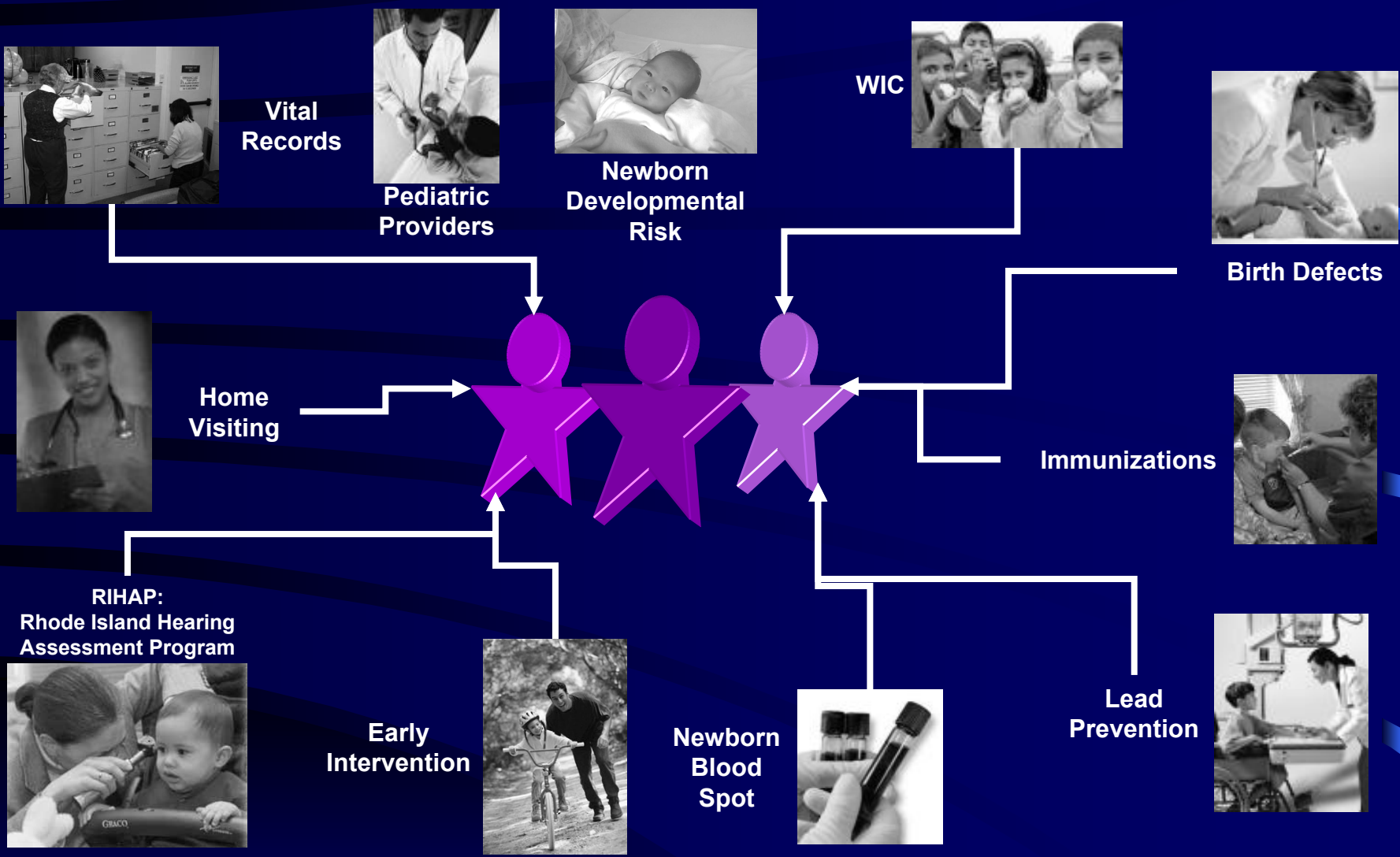
- Not widely used (yet)
- No user interface
- Limited support for developing parameters
- Limited support for testing
- Limited integration support
- Not perfect

# Implementing Febrl

- Requirements solitication
- Integration (batch vs. real-time)
- Parameter development
  - Standardization
  - Blocking
  - Probabilities
  - Frequencies
- Testing
- Adjustments

# Testing and the CDC Toolkit

- Cost
- Effort to Integrate
- Effort to Configure and Test
- Effort to Operate
- Sensitivity
- Specificity
- Speed



# KIDSNET Web Application



## Practice Reports - Patient List

Generate Report

Select Practice (For management users only)

ALLEN BERRY HEALTH CENTER-PCHC

[Search](#)

[Recently Viewed](#)

[User Profile](#)

[Practice Reports](#)

[Forms](#)

[Help](#)

[Logoff](#)

Date of Birth: From:  /  /  / (mm/dd/yyyy)

To:  /  /  / (mm/dd/yyyy)

Or...

Age From:  months

To:  months

Or...

All Patients

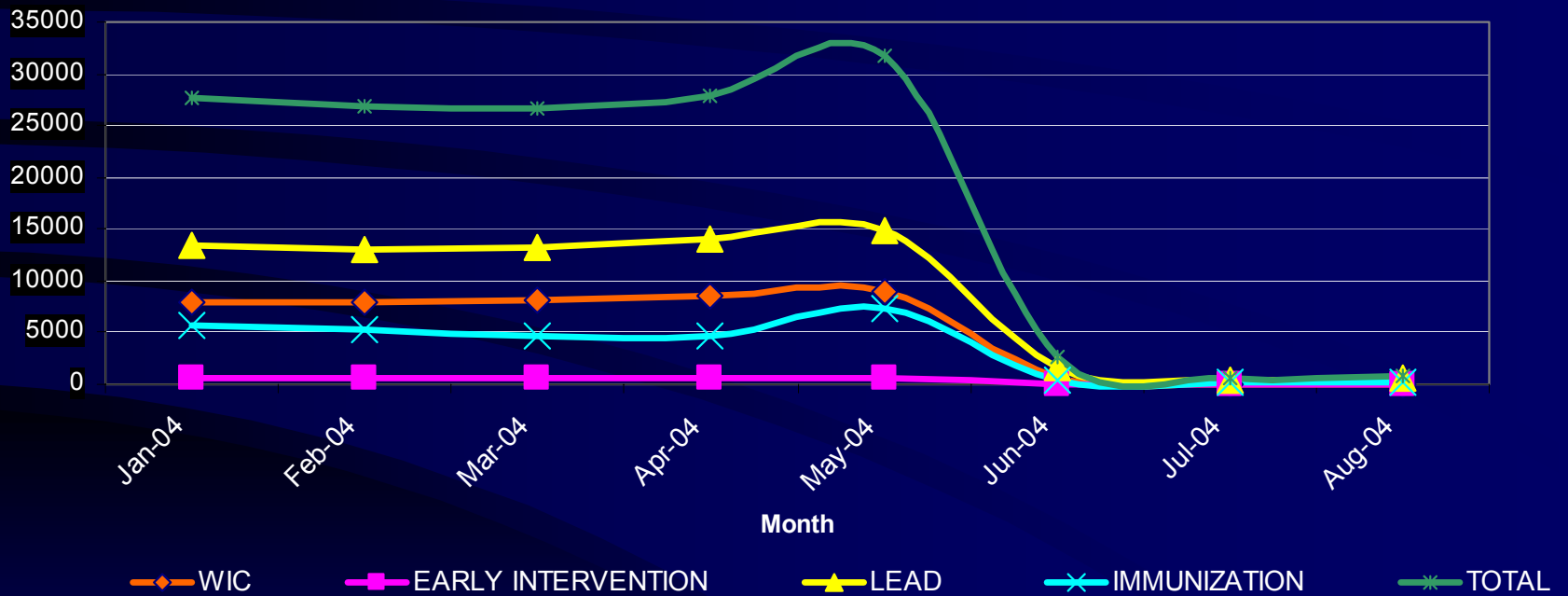
# Before...

- 48,685 records on hold
- Manual, time consuming process that took approximately 3 minutes to resolve a record
- Estimated 70 weeks or 17 months to resolve backlog
- Lack of resources
- Inaccurate reporting
- Limitations to use data
- Little incentive to enroll new providers

# After...

- >45,000 (95%) of 48,685 were removed from the "on hold" upon implementation of tools
- ~11,000 records of children were added to KN
- 78% decrease in time to resolve an error (from 3 minutes to about 40 seconds per record)
- 95% of the data that comes to KIDSNET are now imported and used for reports
- Lack of resources

# History of Records on Hold



# How we got there...

- Funding
- Leadership support
- Prioritization of work
- Long term plan
- Mechanism to hire resources
- In house resources: testing and implementation
- Tedious, continuous, exhausting work

# References

- RI Data Linkage Bibliography -  
[contact berrym@hln.com](mailto:berrym@hln.com)
- PHII Deduplication Study -  
<http://www.phii.org/reading.html>
- Freely Extensible Biomedical Record Linkage (Feb1) Website -  
<http://datamining.anu.edu.au/projects/linkage.html>
- CDC Deduplication Test Kit -  
<http://www.cdc.gov/nip/registry/dedup/dedup.htm>

# Contact Information

Magaly Angeloni

Rhode Island Department of Health

401-222-4602

[MagalyA@doh.state.ri.us](mailto:MagalyA@doh.state.ri.us)

Mike Berry

HLN Consulting, LLC

215-568-3005

[berrym@hln.com](mailto:berrym@hln.com)

# Discussion Questions