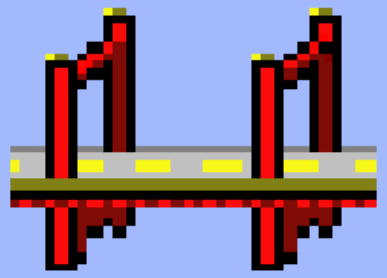
NPCR NATIONAL PROGRAM OF CANCER REGISTRIES

REGISTRY PLUS



# Link Plus Workshop

## *Record Linkage Concepts*

Application Overview & User Training
NAACCR 2009 Pre-conference Workshop
June 14, 2009
8:00 a.m. – 12:00 p.m.

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

# CDC–NPCR Link Plus Faculty

**Rich Pinder,** Analyst
Los Angeles Cancer Surveillance Program

**Kathleen Thoburn,** CTR, CDC/NPCR Contractor

**David Gu,** CDC/NPCR Contractor

# Training Outline

- Brief Overview of Record Linkage
- Central Cancer Registry Record Linkage
- Deterministic Matching
- Probabilistic Matching
- Link Plus Software Overview
- Link Plus Linkage Overview
- Linkage Exercises
- Open Discussion

# Record Linkage Concepts

# Overview of Record Linkage

- Combine or merge together information describing the same individual from a variety of data sources
- Merge information from individual's record in 1st data source (file 1) with information from individual's record in 2nd second data source (file 2)
  - Cancer information from cancer registry file, death information from vital statistics file
- "Merge" aka "Record Linkage"

# Overview of Record Linkage

- Can be accomplished manually, by visually comparing records from two separate sources
- Approach becomes time consuming, tedious, inefficient, and unpractical as the number of records in file 1 and file 2 increases
- Technological advances in computer systems and programming techniques
  - Economically feasible to perform computerized record linkage between large files
  - Efficient and relatively accurate

# Central Cancer Registry Record Linkage

- Case Finding
- Linking New Reports → Consolidation
- Duplicate Detection
- Follow Up
- Special Studies

# Case Finding

- Matching reports from
  - Pathology labs
  - Medical Records Disease Index
  - Treatment centers
- No Match: tumor has not yet been reported
  - Request report of cancer from facility of diagnosis
- Positive Match: tumor is already reported
  - New diagnostic/treatment information can be added to existing tumor record

# Linking New Reports

- Multiple notifications of the same cancer due to multiple reporting sources
- Efficient record linkage procedures on same individual very important
  - Consolidation…Is this a
    - new person?
    - new tumor for an existing person?
    - new report for an existing person/tumor?
- Failure in record linkage process results in missed cases and/or duplicate registrations
  - Leads to inaccurate counts and rates

# Duplicate Detection

- Fundamental requirement for accuracy and validity of counts in any disease registry
- National Program of Cancer Registries/ North American Association of Central Cancer Registries standard
  - Maintain <= 0.1% (<=1 per 1,000) duplicates

# Follow Up

- Death Clearance – State vital statistic file
- Hospital discharge data – Statewide file
- Department of Motor Vehicles – Drivers' licenses and renewals
- Social Security Death Master – SSA maintained file of death benefit claims
- Medicare/Medicaid – Files of state enrollees
- Voter Registration/Voter History - Statewide file of last 6 elections
- National Change of Address (U.S. Postal Service) - File of individuals reporting change of address in last 3 years

# Special Studies

- Research questions often require linking external data against the registry
- Allows hypothesis testing not available using other methods
- Efficient record linkage software is essential

# Deterministic Matching

- Computerized comparison where EVERYTHING needs to match EXACTLY:

| Last Name | First Name | Site | SSN | DOB | Sex | DateDx |
|---|---|---|---|---|---|---|
| SMITH | JOHN | C619 | 123654789 | 02011934 | 1 | 06152004 |
| SMITH | JOHN | C619 | 123456789 | 02011934 | 1 | 06152004 |

# Deterministic Matching

- Often slight variations exist in the data between the two files for the same variables:

| Last Name | First Name | Site | SSN | DOB | Sex | DateDx |
|---|---|---|---|---|---|---|
| SMITH | JOHN | C619 | 123654789 | 02011934 | 1 | 06152004 |
| SM**Y**TH | JOHN | C619 | 12345678**6** | 02**08**1934 | 1 | 06**10**2004 |

- Or variables are missing from one of the files:

| Last Name | First Name | Site | SSN | DOB | Sex | DateDx |
|---|---|---|---|---|---|---|
| SMITH | JOHN | C619 | 123654789 | 02011934 | 1 | 06152004 |
| SMITH | JOHN | C619 | | 02011934 | 1 | 06152004 |

- These variations would prevent a match from being identified

# Deterministic Matching

- Describes an algorithm in which the correct next step is PRE-defined (match/no match)
- Good for production environments
- Easily incorporated into existing data systems

**HOWEVER,**

- Will miss significant numbers of true matches
- Will require enormous amount of manual review of results for missed matches

# Deterministic Matching Manual Review

- When we manually review, we use intuition to help us identify positive matches for records containing slight variations in, or missing information for, data between the two files for the same variables

| Last name | First Name | Site | SSN | DOB | Sex | DateDx |
|---|---|---|---|---|---|---|
| SMITH | JOHN | C619 | 123654789 | 02011934 | 1 | 06152004 |
| SMITH | JOHN | C619 | 123456786 | 02101934 | 1 | 06152004 |

- Typo in SSN, transposition of digits in the day component of DOB, but would still deem a match

# Probabilistic Matching

- What do Humans know?
- How can we translate intuition into formal decision rules to be used by a computer?
- Use the concept of **PROBABILITY** and perform **PROBABILISTIC** matching
- Recommended over traditional deterministic (exact matching) methods when:
  - coding errors, reporting variations, missing data or duplicate records
- Estimate probability/likelihood that two records are for the same person versus not

# Probabilistic Matching

**Definition of Probability:**

- Measure of **how likely** it is that some event will occur
  - "What is the probability of rain tonight?"
- The **likelihood** that a given event will occur
  - "There is little probability of rain tonight."

# Probabilistic Matching

- Find the records in File 2 that seem to match records in File 1
- Calculate a linkage score that indicates, for any pair of records, how **likely** it is that they both refer to the same person
- Sort the likely and possible matched pairs in order of their scores
- Define a threshold (Cut Off value) for automatically accepting and rejecting a potential link
  - Discard unlikely matched pairs (scores below Cut Off)
  - Gray area: range of scores considered as uncertain matches
- Manually review uncertain matches

# Probabilistic Matching

- The total score for a linkage between any two records is the sum of the scores generated from matching individual fields
- The score assigned to a matching of individual fields is:
  - Based on the probability that a matching variable agrees given that a comparison pair is a match
    - **M Probability** - similar to "sensitivity"
  - Reduced by the probability that a matching variable agrees given that a comparison pair is **not** a match
    - **U Probability** - similar to "specificity"

# Probabilistic Matching

- **Agreement** argues **for** linkage (higher score)
- **Disagreement** argues **against** linkage (lower score)
- Full agreement argues more strongly for linkage than partial agreement
- Some types of partial agreements are stronger than others; probabilistic scores are
  - Field-specific – Birth date versus Sex
  - Value-specific - "Jane" versus "Janiqua"

# Phonetic Systems

- Phonetic coding involves coding a string based on how it is pronounced

**Soundex (developed in 1918)**

- Code for a name consisting of a letter followed by three numbers: the letter is the first letter of the name, and the numbers encode the remaining consonants
- Zeroes are added at the end if necessary to produce a four-character code. Additional letters are disregarded.
  - Washington is coded W-252 (W, 2 for the S, 5 for the N, 2 for the G, remaining letters disregarded
- Reduces matching problems due to different spellings
- Simple and fast

# Phonetic Systems

## New York State Identification and Intelligence System (NYSIIS; 1970 +)

- Maps similar phonemes to the same letter; maintains relative vowel positioning
- String can be pronounced by the reader without decoding
  - Deborah Walker = DABARA WALCAR
- Improvement to the Soundex algorithm
  - More distinctive; people are more likely to have the same Soundex than the same NYSIIS
  - Reported accuracy increase of 2.7% over Soundex
  - Studies suggest NYSIIS performs better than Soundex when Spanish names are used
- Soundex may bring more pairs for comparison when used for blocking

# Concept of Blocking

- With so many comparisons, large files can make impossible resource demands
- Blocking is an initial probabilistic linkage step that reduces the number of record comparisons between files
- Sort and match the two files by one or more identifying ("blocking") variables
- Comparisons subsequently made only **within** blocks
  - Discard very unlikely record-pairings from the start

# Blocking Variables

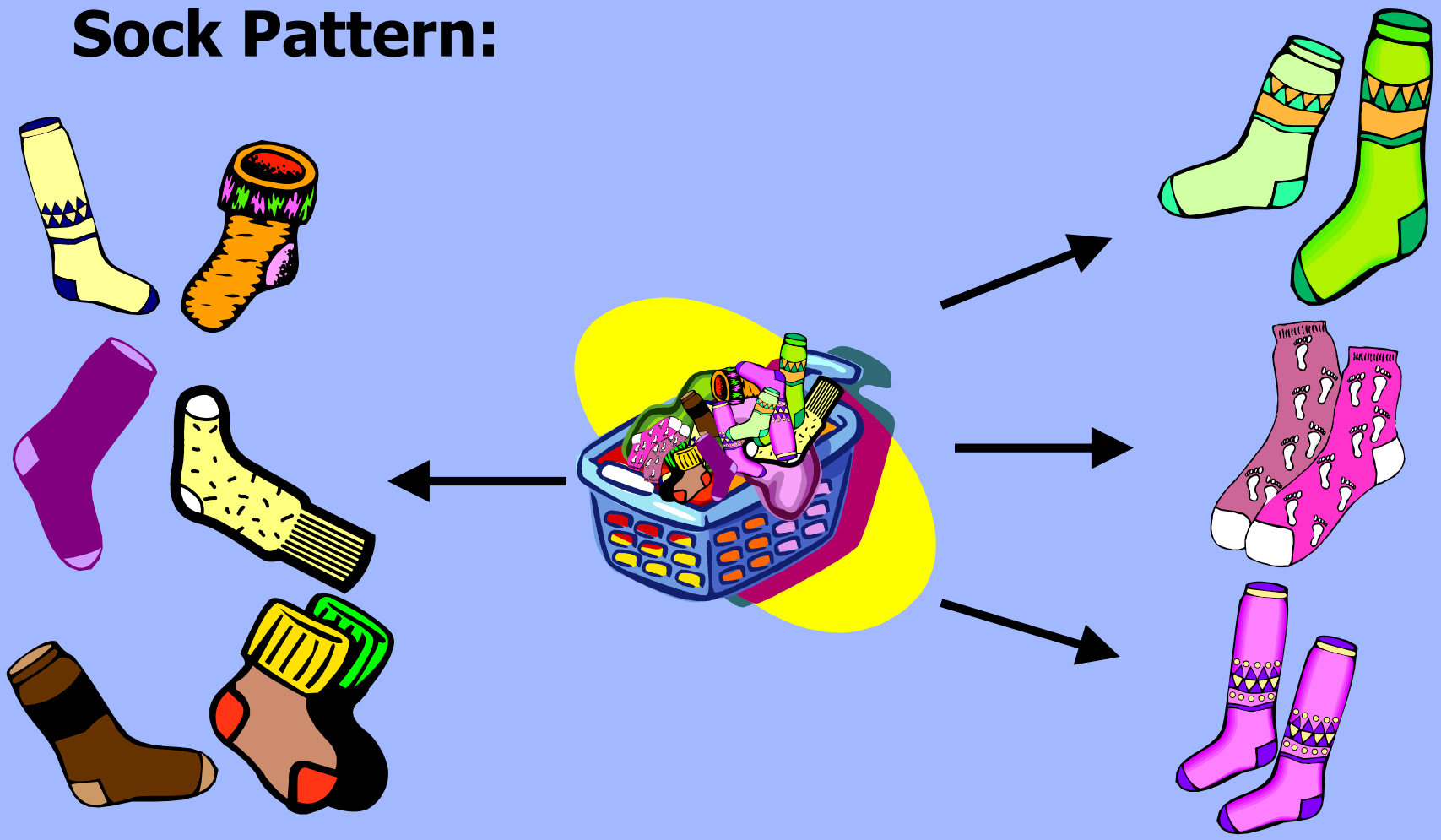
- Exact matches
- Blocks of data to compare variables within
- Common blocking variables are:
  - Last Name
  - Social Security Number
  - Date of Birth

# Matching Variables

- Probabilistic matching algorithms
- Comparing variables within blocks
- Common matching variables:
  - Name--Last
  - Name--First
  - Name--Middle
  - Sex
  - Race
  - Birth Date
  - Social Security Number

# Blocking Variables

## Sock Pattern:



7 of 13 socks fall outside pattern block

6 of 13 socks within pattern block

# Matching Within Blocks

**Blocking: Pattern**
**Matching: Color & Size**

**High Score**

**Gray Area**

**Low Score**

# Summary

- Record linkage is a vital activity for cancer registries
- Record linkage is becoming easier
- Efficiency is a key feature
  - Faster, more efficient linkage process allows more linkages for less $$ and staff time
    - More accurate counts
    - More research
    - Increased utilization of registry data