

# Record Linkage Concepts

# Acknowledgements

Slides adapted from training materials developed by CDC–NPCR Faculty:

- Melissa Jim, CDC/IHS  
[melissa.jim@ihs.gov](mailto:melissa.jim@ihs.gov)
- David Espey, CDC/IHS  
[david.espey@ihs.gov](mailto:david.espey@ihs.gov)

CDC/Link Plus development and training:

- Kathleen Thoburn
- David Gu

Adapted by:

- Megan Hoopes, NW Tribal Epidemiology Center  
[ideanw@npaihb.org](mailto:ideanw@npaihb.org)



# Overview of Record Linkage

- “Record Linkage” aka “Matching” aka “Merge”
- Combining information from a variety of data sources for the same individual
- Merge information from a record in one data source (file 1) with information from another data source (file 2)
  - Example: merging cancer information from cancer registry file with death information from vital statistics file

# 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

# Duplicate Detection

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

# Deterministic Matching

- Computerized comparison where EVERYTHING needs to match EXACTLY:


<b>Last Name</b>	<b>First Name</b>	<b>Site</b>	<b>SSN</b>	<b>DOB</b>	<b>Sex</b>	<b>DateDx</b>
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	123456789	02011934	1	06152004
SMYTH	JOHN	C619	123456786	02081934	1	06102004

- 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

## 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	12345678 <b>9</b>	02 <b>01</b> 1934	1	06152004
SMITH	JOHN	C619	12345678 <b>6</b>	02 <b>10</b> 1934	1	06152004

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



# Probabilistic Matching

- Translating intuition into formal decision rules
- 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

- Find the records in File 2 that seem to match records in File 1
- Calculate a 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 values) for automatically accepting and rejecting a potential link
  - Discard unlikely matched pairs (scores below 2<sup>nd</sup> Cut Off)
  - Gray area: range of scores between the two cut off values considered 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
- **Disagreement** argues **against** linkage
- Full agreement argues more strongly for linkage than partial agreement
- Some types of partial agreements are stronger than others
  - Rare surname versus residence county code

# Probabilistic Matching

- Agreement on an uncommon value argues more strongly for linkage than a common value
  - Espey versus Smith
- Agreement on a more specific variable argues more strongly for linkage than agreement on a less specific one
  - SSN versus Sex
- Agreement on more variables/disagreement on few argues for linkage

# Probabilistic Matching

- Once comparisons are made, a **weight** is calculated for each field comparison
- A total weight (or “score”) is derived by summing these separate field comparisons across all fields being compared
- Probabilistic weights are
  - Field-specific – Birth date versus Sex
  - Value-specific - “Jane” versus “Janiqua”

# Linkage basics

- **Blocking variables**
- **Matching variables**
- **Advantages of Link Plus**
- **Using Link Plus**

# 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



# Sorting socks analogy

**Blocking variable:  
Pattern**



**7 of 13 socks fall  
outside pattern block →  
Non-matches**



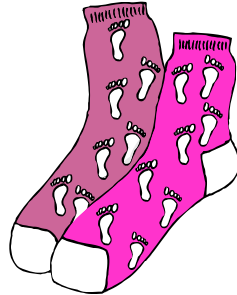
**6 of 13 socks within  
pattern block →  
compare matching  
variables**

# Compare matching variables color & size within blocked pairs



**High Score**

**Possible matches**



**Gray Area**



**Low Score**

# Probabilistic linkage concepts (1)

	Description	Common usage*
<b>Blocking</b>	An initial step to reduce the number of record comparisons and increase efficiency of linkage. At least one blocking variable must match exactly (or phonetically) between the two records being compared; subsequent comparisons are made after blocking.	Blocking variables: <ul style="list-style-type: none"><li>• Last name</li><li>• First name</li><li>• Social security number</li><li>• Date of birth</li></ul>
<b>Matching</b>	<p>After blocking, matching variables are compared to generate a match score for each record pair.</p> <p>Match scores for each variable are:</p> <ul style="list-style-type: none"><li>• Field-specific (matching DOB is scored higher than matching sex)</li><li>• Value-specific (last name of 'Hoopes' is scored higher than 'Smith' due to frequency of occurrence)</li></ul>	<p>Matching variables:</p> <ul style="list-style-type: none"><li>• Last name</li><li>• First name</li><li>• Social Security Number</li><li>• Date of Birth</li><li>• Sex</li><li>• Address</li></ul> <p>The user may designate matching algorithms &amp; M-probabilities for each variable.</p>

\* May vary based on data items and quality of data in available in matching data sets

## Probabilistic linkage concepts (2)

	Description	Common usage*
Match score	The total probability weight assigned to each record pair; equal to the sum of scores generated by comparing each match field. Based on software-calculated M probability (sensitivity) and U probability (specificity).	The range of match scores is examined to determine upper and lower cut-off values. High match scores are likely true matches and scores below cut-off value are automatically designated false matches. Record pairs between cut-off values are clerically reviewed.
Clerical review	Case-by-case review of uncertain matches that fall between the upper and lower cut-off values. Additional variables can be added to record layout to assist in the designation of match status. This process can be completed independently by two or more reviewers to increase reliability.	Additional variables may include: <ul style="list-style-type: none"><li>• Street address</li><li>• City, state, zip code</li><li>• Suffix</li><li>• Race/ethnicity</li><li>• Maiden name</li></ul>

\* May vary based on data items and quality of data in available in matching data sets