

# Overview of Link Plus

*Probabilistic  
Record Linkage Software*

# 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)

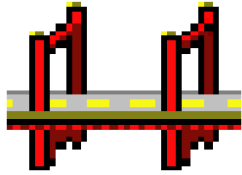
CDC/Link Plus development and training:

- Kathleen Thoburn
- David Gu

Adapted by:

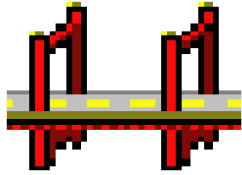
- Megan Hoopes, NW Tribal Epidemiology Center  
[ideanw@npaihb.org](mailto:ideanw@npaihb.org)





# Link Plus Software

- Stand-alone probabilistic record linkage program
- Combines ease of use and statistical sophistication
- Detects duplicates within a single database, or links 2 database files
- Supports North American Association of Central Cancer Registries files, fixed width files, delimited files, and CRS Plus database



# Link Plus Software

- Can handle missing values of matching variables
  - automatically treats null or empty values as missing data and allows user to indicate additional values to be treated as missing data
- Facilitates blocking ("OR blocking") by indexing the variables and comparing the pairs with the identical values on at least one of those variables
- Provides support for manual review of uncertain matches

Link Plus Is Free

**\$0.00**

# Link Plus Is Easy To Use

Link Plus gets you from **HERE**:

Cancer Registry data for John Smith:

Last name	First Name	Site	SSN	DOB	Sex	DateDx
SMITH	JOHN	C619	123654789	02111934	1	06152004

Vital Statistics data for John Smith:

Last name	First Name	DOB	Date of Death	COD	Death Cert #
SMITH	JOHN	02111934	03202006	123654789	01234

# Link Plus Is Easy To Use

To **HERE**:

Linked data for John Smith:

Last name	First Name	Site	SSN	DOB	Sex	DateDx	Death Date	COD	Death Cert #
SMITH	JOHN	C619	123654789	02011934	1	06152004	03202006	C100	01234

# Link Plus Is Easy To Use

Without having to go **HERE**:

$$P(\gamma \mid M) = \prod_{i=1}^K m_i^{\gamma^i} (1-m_i)^{(1-\gamma^i)}$$

and

$$P(\gamma \mid U) = \prod_{i=1}^K u_i^{\gamma^i} (1-u_i)^{(1-\gamma^i)} .$$



# Link Plus Is Easy To Use

- Designed especially for cancer registry work
  - HOWEVER, can be used with **any** data
- Mathematics largely hidden from user
- Practical default values supplied for many tasks
- Familiar Windows interface
- Includes Help and test examples

# Using Link Plus

# Obtaining/Updating Link Plus



1. Go to NPCR Home Page:  
<http://www.cdc.gov/cancer/npcr>
2. In the 'Software and Tools' Section  
- click on [Registry Plus](#)
3. Under 'Registry Plus Components'  
- click on [Link Plus - Download](#)

# Getting Started

- Make sure you know your data!
- Review and clean data files
- Frequency distributions very helpful
- Look for errant values
  - e.g. - DOB day component = 16

# Data Cleaning Tips

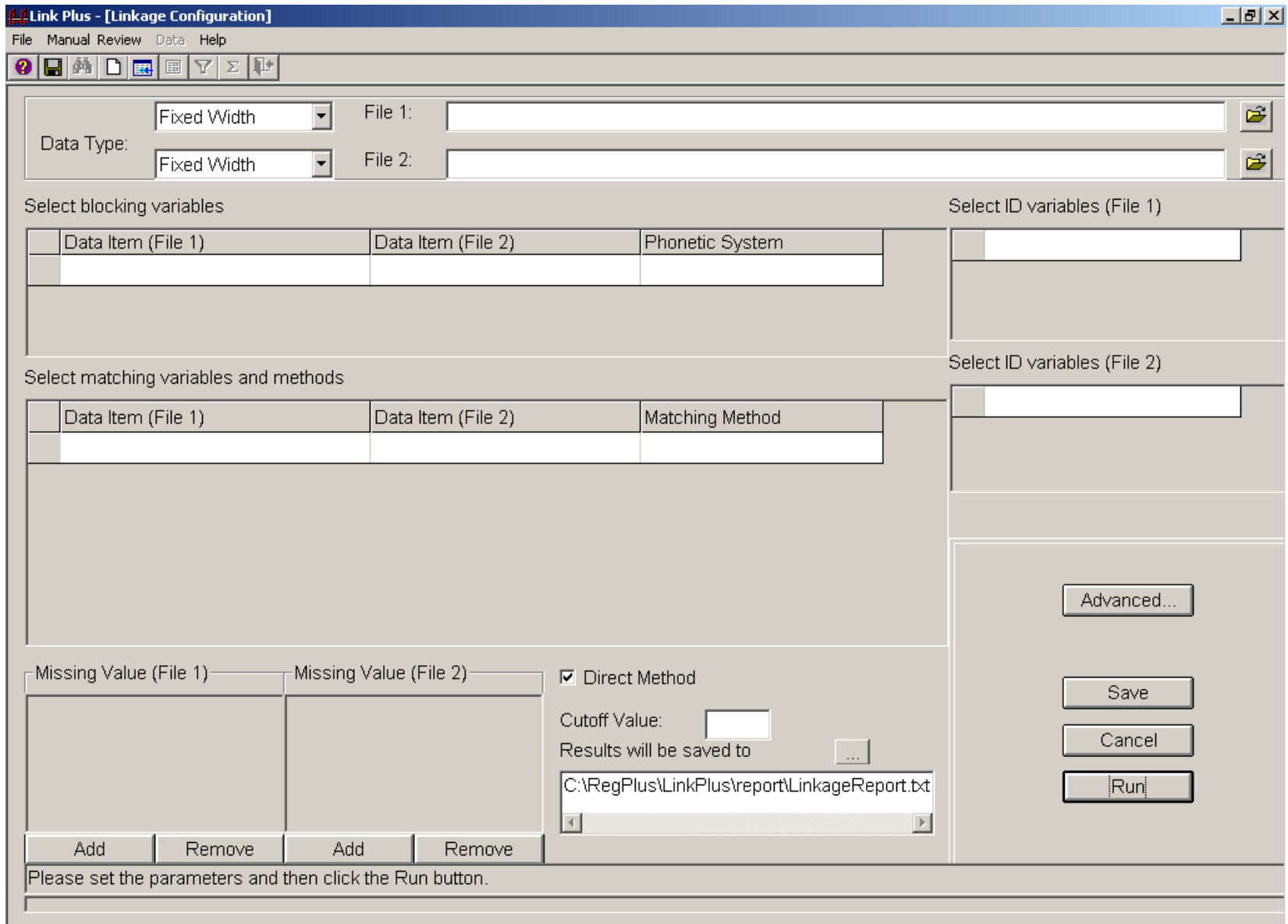
- Last Name
  - Link Plus automatically cleans punctuation and strips off suffices, numbers III
- First Name
  - May find **Dr.** Bill or **Rev** Bill or **Sister** Mary
  - Remove prefix in First Name field
- Middle Name
  - Link Plus automatically cleans numbers, weird symbols
  - Link Plus accounts for the switching of first and middle names
  - NMI-no middle initial or NMN-no middle name
- DOB
  - Review day, mo, yyyy component
  - Replace errant values with missing
- Sex
  - Make sure files use same coding convention; M, F, or Blank OR 1, 2, 9

# Linkage Overview

## Two main types of linkage:

- Linkage of 2 files
  - Probabilistically link one file to another file
- Deduplication
  - Special case of record linkage
  - Records in the same file are blocked, compared, and scored against each other
  - Result is a ranked list of record pairs
  - High-scoring pairs may be duplicates

# Linkage configuration screen



Link Plus - [Linkage Configuration]

File Manual Review Data Help

Data Type: Fixed Width File 1: File 2:

Select blocking variables

Data Item (File 1)	Data Item (File 2)	Phonetic System

Select ID variables (File 1)

Select matching variables and methods

Data Item (File 1)	Data Item (File 2)	Matching Method

Select ID variables (File 2)

Missing Value (File 1) Missing Value (File 2)

Direct Method

Cutoff Value: Results will be saved to C:\RegPlus\LinkPlus\report\LinkageReport.txt

Add Remove Add Remove

Advanced... Save Cancel Run

Please set the parameters and then click the Run button.

# File Import

## File 1 versus File 2

- File 1 and File 2 have a one to many relationship
  - File 1 (ONE); File 2 (MANY)
- Death record is File 1; cancer registry is File 2:
  - One John Smith in death record file can link to many John Smith's in cancer registry file
- Cancer registry is File 1; death record is File 2
  - One John Smith in cancer registry file can link to many John Smith's in death record file

Data Type:	Fixed Width	File 1:	<input type="text"/>	
	Fixed Width	File 2:	<input type="text"/>	



# File Layout

- After designating File 1 and File 2, specify file layout for each file
- Link Plus reads .txt files:
  - Fixed width
  - Comma delimited
  - Tab delimited
- Link Plus provides view of first 20 records of each input file
  - Verify that data is being read in properly

# Blocking Variables

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

Select blocking variables			
	Data Item (File 1)	Data Item (File 2)	Phonetic System

# Phonetic Systems

- Phonetic coding involves coding a string based on how it is pronounced
- Link Plus offers a choice of 2 Phonetic Coding Systems:

## **Soundex**

- 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
- Reduces matching problems due to different spellings
- Simple and fast

# Phonetic Systems

## New York State Identification and Intelligence System (NYSIIS)

- Maps similar phonemes to the same letter; maintains relative vowel positioning
- String can be pronounced by the reader without decoding
- 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
- However, Soundex may bring more pairs for comparison when it used for blocking

# Matching Variables

- Up to 10 fields may be selected for matching
- Recommended variables (Matching Methods):
  - Name--Last (LastName)
  - Name--First (FirstName)
  - Name--Middle (MiddleName)
  - Sex (Exact)
  - Race (Exact)
  - Birth Date (Date)
  - Social Security Number (SSN)

Select matching variables and methods			
	Data Item (File 1)	Data Item (File 2)	Matching Method

# Matching Methods

- **Exact**
  - Case insensitive character-for-character string comparison method
  - Results are either yes or no
- **Generic String**
  - Uses edit distance function (Levenshtein distance) to compute the similarity of two long strings
    - Minimum number of operations (insertion, deletion, or substitution of a single character) needed to transform one string into the other
- **Last Name/First Name**
  - Incorporate both partial matching and value-specific matching to account for minor typographical errors, misspellings, and hyphenated names

# Matching Methods

- **SSN**
  - Specifically for Social Security Number
  - Incorporates partial matching to account for typographical errors and transposition of digits
- **Date**
  - Incorporates partial matching to account for missing month values and/or day values
- **Middle Name**
  - Accounts for occurrence of the middle initial only versus the full middle name

# Matching Methods

- **Value-Specific (Frequency-Based)**
  - Intended for advanced users
  - Sets weights for matching values based on the frequencies of values in the two files being compared
  - A match on a frequent value is associated with a low weight, while a match on a rare value is associated with a high weight
  - In a file with a high proportion of records with a white race (value of 01), a match on value 01 would be weighted lower than a match on the value 03 (American Indian)



# Missing Values

- Automatically treats null or empty values as missing data for Matching Variables
- Allows user to indicate additional values which are to be treated as missing data by the program
- Specify date format on the missing value grid when the Date Matching method is applied to a matching variable

# Missing Values

- Specify date format on the missing value grid

Select matching variables and methods

	Data Item (File 1)	Data Item (File 2)	Matching Method
*	DOB	Birth place	Date
	LNAME	Name--Last	Last Name
	FNAME	Name--First	First Name
	SSN	Social Security Number	SSN
	MI	Name--Middle	Middle Name

Missing Value (File 1)		Missing Value (File 2)	
Day	99	Day	99
Month	99	Month	99
Year	9999	Year	9999
Format	YYYYMMDD	Format	MMDDYYYY

Direct Method

Cutoff Value:

Results will be saved to

Please set the parameters and then click the Run button.

# Direct Method

Direct Method

- "Direct Method" refers to the method used to derive the M-Probabilities used in linkage
  - Probability that a matching variable agrees given that a comparison pair is a match
- Click if you prefer to use the **default** M-Probabilities or user-defined M-Probabilities
- Click on Advanced to view the default M-Probabilities or **define** your own
- Un-click if you prefer Link Plus to **compute** the M-Probabilities **based on your data** using the EM algorithm

# EM Algorithm



- Expectation-**M**aximization algorithm
- Method for maximum likelihood estimation in problems involving incomplete data
- Unclick Direct Method if you would like Link Plus to **compute** the M-Probabilities **based on your data**
  - Estimates reflect the characteristics of the data dynamically
- Very popular computational method
- Basic principle is to derive a solution in a complicated case from a corresponding solution in a simple case

# Cut Off Value

Cutoff Value:

- The score value above which comparison pairs are accepted as potential links and presented for review
- Value should always be positive
- Initial value of around 7 recommended when using the recommended Matching Variables
- Run linkage, and quickly review potential matches to identify upper and lower cut off scores
  - Upper cut off = where do “perfect” matches end?
  - Lower cut off = where do non-matches begin?

Example:

- Solid matches somewhere between 20-30
- Non-matches 10-20
- Solid = 27 and higher, and Non-matches = 17 and lower
- Gray Area = Greater than 17 and less than 27

# Manual Review of Uncertain Matches

Link Plus - [Linkage Report=C:\RegPlus\LinkPlus\Report\LinkageReport.txt]

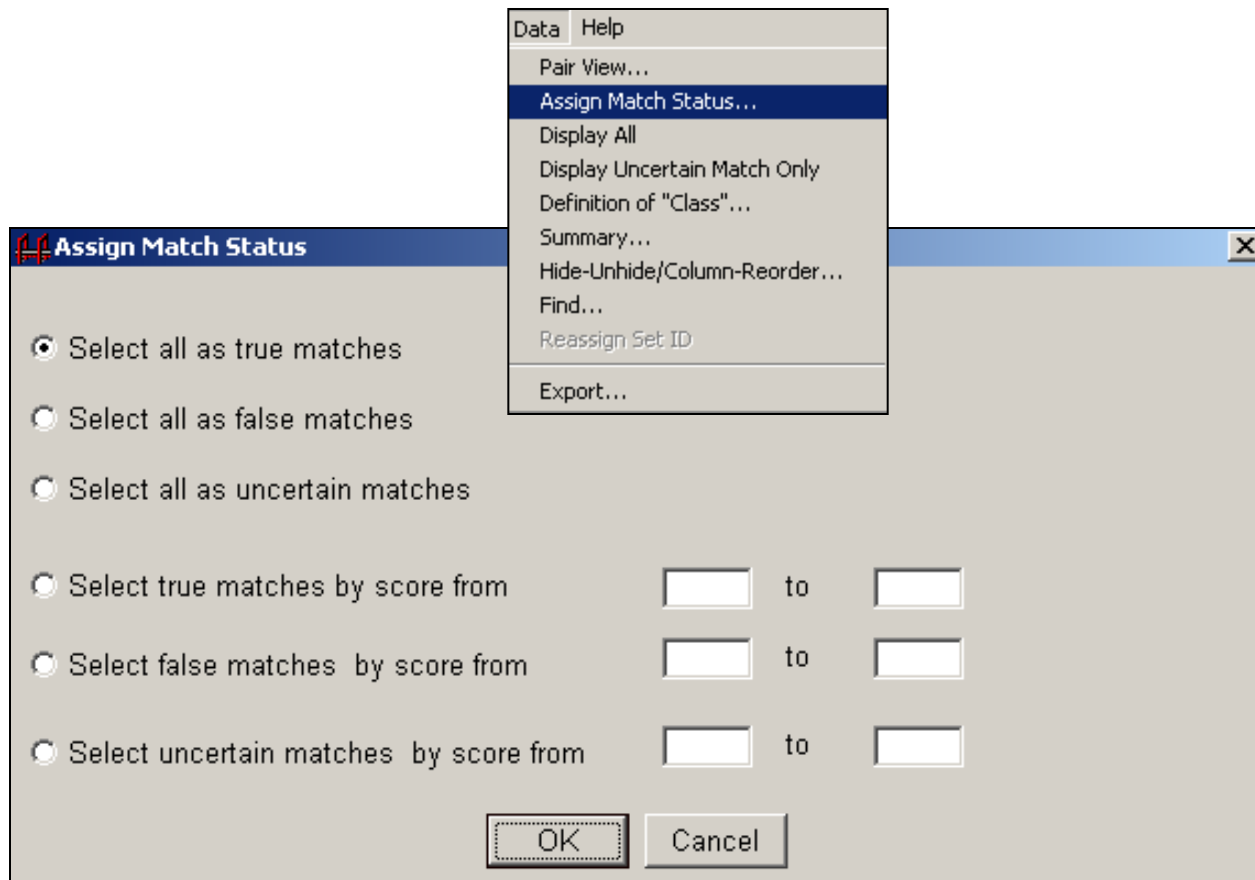
Manual Review Data Help

= true matches   
 = false matches   
 = uncertain matches   
  = unmatched values   
  = missing values

Score	Class	Link ID	File	Record #	lastname;last nar	first name;first nar	dob;dob	ssn;ssn
<input type="checkbox"/> 14.3	1	122	2	110	WINTHROP	JOHN	03081912	789999999
	1	123	1	146	WINSLOW	JOHN	12091922	775000154
<input type="checkbox"/> 14.3	1	123	2	146	WINSLOW	JOHN	12091922	775000154
	1	124	1	77	WINSLOW	ELIZABETH	08161928	790006570
<input type="checkbox"/> 14.3	1	124	2	77	WINSLOW	ELIZABETH	08161928	790006570
	1	125	1	45	CHRISTMANN	ELIZABETH	10181970	780990000
<input type="checkbox"/> 14.3	1	125	2	45	CHRISTMANN	ELIZABETH	10181970	780990000
	3	126	1	19	LEACHH	LAWRENCE	06301921	790002277
<input type="checkbox"/> 14.3	3	126	2	19	LEACH	LAWRENCE	06301921	790002277
	2	127	1	13	EATON	ERNIST	11251934	770001234
<input type="checkbox"/> 14.1	2	127	2	13	EATON	ERNEST	11251934	770001234
	3	128	1	21	COOK	FRANCIS	10291914	782109285
<input type="checkbox"/> 14.1	3	128	2	21	COOKE	FRANCIS	10291914	782109285
	3	129	1	3	BAGIN	HENRY	10271929	750008679
<input type="checkbox"/> 14.0	3	129	2	3	BAGIN-Danes	HENRY	10271929	750008679
	3	130	1	134	BASETT	WILLIAM	04271906	763003422
<input type="checkbox"/> 13.5	3	130	2	134	BASSETT	WILLIAM	04271906	763003422
	3	131	1	24	READ	WILLIAM	01131926	782121845
<input type="checkbox"/> 13.1	3	131	2	24	READE	WILLIAM	01131926	782121845
	4	132	1	92	WOOD	ANNE	07991928	773001234
<input type="checkbox"/> 12.9	4	132	2	92	WOOD	ANNE	07091928	773001234
	4	133	1	66	MAULDER	PHEBE	99171918	790001001
<input type="checkbox"/> 12.6	4	133	2	66	MAULDER	PHEBE	06171918	790001001
	4	134	1	93	FRENCH	ELIZABETH	06991938	778007600
<input type="checkbox"/> 11.9	4	134	2	93	FRENCH	ELIZABETH	06281938	778007600
	4	135	1	73	JACKSON	JOHN	99991954	768500000
<input type="checkbox"/> 11.5	4	135	2	73	JACKSON	JOHN	06171954	768500000
	4	136	1	63	HUBBARD	WILLIAM	99991931	755051021

# Manual Review of Uncertain Matches

- Once Cut Off scores have been identified



# Manual Review of Uncertain Matches

- Manual review screen provides special sorting that always keeps the comparison pairs together
- Fields that are not on the linkage report can be included for manual review
- Users can switch between the two viewing modes: the Datasheet View and the Pair view
- Match status may be assigned manually, or may be assigned automatically by match score



# Manual Review of Uncertain Matches

- Includes the option to restrict view to only uncertain matches
- Individual columns may be sorted, hidden and unhidden from view, and re-sized
- Order of the columns on the review screen may be modified
- Review sessions may be saved and re-opened at a later time
- Allows two reviews to be compared so that the difference can be resolved into a final review file

# Tips for Manual Review

- Focus initially on SSN and DOB
  - Names have a lot of issues (spelling and spacing)
- Once matches on SSN go away, pay attention to DOB, name, and sex
  - First and middle name switches are common
- Use race & address variables if available
- First time you start doubting if a pair is a match
  - Score will be upper cut off score
  - Anything above is considered a match
- When start to see junk
  - Score will be lower cut off score
  - Anything below is considered a non-match
- Keep an eye out for
  - Husbands and wives matching (SSN's match/sex different))
  - Brothers, sisters, and twins (LN match, SSN off by 1)
- Two people should review so that results can be combined and resolved
- These are suggestions - need to know your own data

# Helpful Link Plus Tips

- With real data Link Plus may take awhile to read data files
  - Be patient - Linkage times vary
    - IHS Linkage (2.4 million records)


With VS Data:	With Cancer Registry Data:
• Oregon: 2 hrs/400,000 recs	• Oregon: 36 min/94,172 recs
• Michigan: 6.5 hrs/1,207,000 recs	• California: 13 hrs/1,935,255 recs
- Wait (but not days) – something went wrong
  - Can run out of virtual memory
- Takes a lot of CPU
  - Shutdown and restart computer right before linkage to clear up as much space as possible prior to linkage
  - Turn off screen saver/Close all other programs