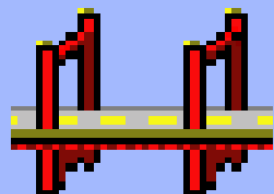


Fundamentals of Linking Public Health Datasets



Link Plus

Probabilistic Record Linkage Software

NAHDO--CDC (Assessment Initiative)

2nd Probabilistic Record Linkage Conference Call

March 30, 2007

3:00 - 4:30 P.M. EST





CDC–NPCR Link Plus Contacts

Kathleen K. Thoburn, CDC/NPCR Contractor

E-mail: kthoburn@cdc.gov

David Gu, CDC/NPCR Contractor

E-mail: dgu@cdc.gov

Tom Rawson, CDC Computer Programmer

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

- Name spelling, typo in SSN, transposition of digits in the day component of DOB
 - Would still deem a match

Probabilistic Matching

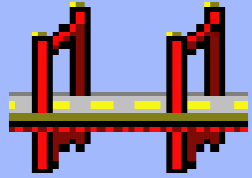
- Find the records in File 2 that **seem** to match records in File 1
- Calculate a score that indicates, for any pair of records, how **likely** it is that they both refer to the same person
- Sort the likely and possible matched pairs in order of their scores
- Define a threshold (Cut Off Value) for 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"

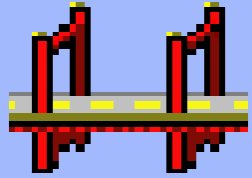
Probabilistic Matching

- **Agreement** argues **for** linkage
- **Disagreement** argues **against** linkage
- Full agreement argues more strongly for linkage than partial agreement
- Some types of partial agreements are stronger than others
 - Field-specific – Birth date versus Sex
 - Value-specific - “Jane” versus “Janiqua”



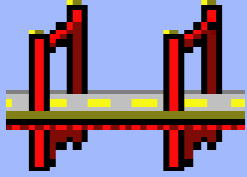
Link Plus Software

- Stand-alone probabilistic record linkage program
- Combines ease of use and statistical sophistication
- Detects duplicates within a single data file, or links two files together
- Supports fixed width files and delimited files
- Provides powerful support for manual review of uncertain matches



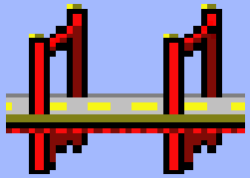
Link Plus Software

- Computes probabilistic record linkage scores based on the theoretical framework developed by Fellegi and Sunter
 - Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, pp. 1183-1210
- Handles missing values of matching variables
- Facilitates a simple and efficient blocking ("OR blocking") mechanism
 - Indexes blocking variables and compares pairs with identical values on at least one of those variables



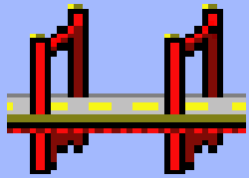
Link Plus Is Free

\$0.00



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



Link Plus Is Robust

- Program written by a mathematical statistician
- Specifications based on research into the published literature
- Tested by researchers experienced in record-linkage
- Results are clear and accessible to novice users

Link Plus Linkage Overview

Prior to Linkage

- Review and clean data files
- Set up data files
 - Same coding conventions for same variables
- Link Plus provides view of first 20 records of each input file
 - Verify data is read in properly

Link Plus Linkage Overview

External Linkage Steps:

1. Select Data Type for File 1
2. Locate/Identify File 1
3. Data Import for File 1
4. Select Data Type for File 2
5. Locate/Identify File 2
6. Data Import for File 2
7. Select Blocking Variables & Phonetic System
8. Select Matching Variables & Matching Methods
9. Select ID Variables
10. Define Missing Values
11. Select Direct/EM Method
12. Enter Cut-off Value
13. Specify Linkage File Name and Location
14. Perform Manual Review of Uncertain Matches
15. Export Merged File

Link Plus Linkage Configuration

Specify Data Type

Select Blocking Variables/ Phonetic System

Select Matching Variables/ Methods

Specify Missing Values

Identify/Import Data Files

Data Type: Fixed Width

File 1: C:\RegPlus\LinkPlus\data\STATEVS2005.dat

Data Type: Fixed Width

File 2: C:\RegPlus\LinkPlus\data\CCR10.dat

Select blocking variables

Data Item (File 1)	Data Item (File 2)	Phonetic System
DOB	Birth Date	
LNAME	Name--Last	Soundex
SSN	Social Security Number	

Select ID variables (File 1)

DCERT

Select ID Variables

Select matching variables and methods

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

Select ID variables (File 2)

Patient ID Number

Missing Value (File 1)

Day	99
Month	99
Year	9999
Format	YYYYMMDD

Missing Value (File 2)

Day	99
Month	99
Year	9999
Format	MMDDYYYY

Add Remove Add Remove

Direct Method **Direct Method/EM Algorithm**

Cutoff Value: **Enter Cutoff**

Results will be saved to

Specify Linkage File Name and Location

Advanced...

Save

Cancel

Run

Save linkage configuration

Run linkage!

The configuration file C:\RegPlus\LinkPlus\Configuration\screenshot.cfg has been saved

Link Plus Manual Review

Link Plus - [View=C:\RegPlus\LinkPlus\Report\trainingview.view]

Manual Review Data Tools Help

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

Score	Class	Link ID	File	Record #	LNAME;Name--La	FNAME;Name--Fi	DOB;Birth Dat	SSN;Social Security Num	MI;Name--Middl	SEX;Sex	RACE;Race 1	DT	
<input checked="" type="checkbox"/>	4	60	1	41	JONES	GINA	01071937	806126055	C	2	1	00	
<input checked="" type="checkbox"/>	18.7	4	60	2	20002	JONES	GINA	01071937	806126055	CHRISTINE	2	1	
<input checked="" type="checkbox"/>	4	61	1	27	FOSTER	LINDA	12011928	836926285	B	2	1	00	
<input checked="" type="checkbox"/>	18.6	4	61	2	13002	FOSTER	LINDA	12011928	836926285	BARBARA	2	1	
<input checked="" type="checkbox"/>	4	62	1	40	LONG	NORMAN	11051933	801825875		1	1	00	
<input checked="" type="checkbox"/>	18.4	4	62	2	19502	LONG	NORMAN	11051933	801825875		1	1	
<input checked="" type="checkbox"/>	4	63	1	21	RICKARD	AUDREY	01141921	801624953		2	1	00	
<input checked="" type="checkbox"/>	18.4	4	63	2	10002	RICKARD	AUDREY	01141921	801624953		2	1	
<input checked="" type="checkbox"/>	4	64	1	91	AVER	DAVID	09151930	806228612	D	1	1	00	
<input checked="" type="checkbox"/>	18.2	4	64	2	45002	AVER	DAVID	09151930	806228612	DALLAS	1	1	
<input checked="" type="checkbox"/>	4	65	1	89	KEEL	DAVID	08221922	822228984	J	1	1	00	
<input checked="" type="checkbox"/>	18.2	4	65	2	44002	KEEL	DAVID	08221922	822228984	JIM	1	1	
<input checked="" type="checkbox"/>	4	66	1	72	EDWARDS	ALLISON	04051931	911630798		2	1	00	
<input checked="" type="checkbox"/>	18.2	4	66	2	35502	EDWARDS	ALLISON	04051931	911630798		2	1	
<input checked="" type="checkbox"/>	4	67	1	66	CHAPPELL	WILLIAM	12041924	812127511	A	1	1	00	
<input checked="" type="checkbox"/>	18.2	4	67	2	32502	CHAPPELL	WILLIAM	12041924	812127511	ABRAHAM	1	1	
<input checked="" type="checkbox"/>	15	68	1	4	COGGINS	BARBRA	01271940	870224354	T	2	1	00	
<input checked="" type="checkbox"/>	18.1	15	68	2	1502	COGGINS	BARBARA	01271942	807224354	T	2	1	
<input checked="" type="checkbox"/>	14	69	1	9	PRICE	CLARE	08161926	812214719		2	1	00	
<input checked="" type="checkbox"/>	18.0	14	69	2	4002	PRICE	CLARE	07161926	812124719	SESTER	2	1	
<input type="checkbox"/>	4	70	1	78	SPICER	CATHERINE	05301927	802827835		2	1	00	
<input type="checkbox"/>	17.9	4	70	2	38502	SPICER	CATHERINE	05301927	802827835		2	1	
<input type="checkbox"/>	4	71	1	64	MAYNARD	JOHN	01221955	884129570	D	1	1	00	
<input type="checkbox"/>	17.9	4	71	2	31502	MAYNARD	JOHN	01221955	884129570	DOMINIC	1	1	
<input type="checkbox"/>	4	72	1	39	SENA	JOHN	03101913	807425994	G	1	1	00	
<input type="checkbox"/>	17.9	4	72	2	19002	SENA	JOHN	03101913	807425994	GEORGE	1	1	
<input type="checkbox"/>	4	73	1	32	NORRIS	MICHAEL	05091957	800025427	P	1	1	00	
<input type="checkbox"/>	17.9	4	73	2	15502	NORRIS	MICHAEL	05091957	800025427	PELLOT	1	1	

Link Plus Version 2 June 2007



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

