Link Plus

A Probabilistic Record Linkage Tool for Cancer Registry Data Linking and Deduplicating

Joe Rogers

David Gu

Tom Rawson

DEPARTMENT OF HEALTH AND HUMAN SERVICES
CENTERS FOR DISEASE CONTROL AND PREVENTION
Atlanta, Georgia, USA

What Is Link Plus?

- An easy-to-use stand-alone Windows application developed at CDC for:
  - Detecting duplicates in the cancer registry database
  - linking a cancer registry file against external files, such as for death clearance
- Performs probabilistic record linkage to accommodate missing values, misspellings, abbreviations, typographical errors as well as other errors

Examples

Patient ID	SSN	Last Name	First Name	Middle Name	Birth Date	Sex	Race
60738	258005679	ROBERT	DAVINA	L	19191129	2	01
50932	999999999	ROBERT	DAVINA	LORENZA	19191120	2	01

Patient ID	SSN	Last Name	First Name	Middle Name	Birth Date	Sex	Race
52768	475343678	TREADAWAY	MILBRA	AMELIA	19450107	2	01
66237	475343678	TREADWAY	AMELIA		19450701	2	99

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Link Plus: Overview

- Performs probabilistic record linkage to accommodate missing values, misspellings, abbreviations, typographical errors as well as other errors
- Employs the model based on the theoretical framework developed by Fellegi and Sunter in their paper “A Theory of Record Linkage” (1969)

Link Plus: Features

- Easy to use
  - Simple interface
  - Pre-configured for one-click deduplication of standardized US cancer registry data
- Can be used with
  - CRS Plus database
  - Fixed-width format file, e.g., NAACCR records
  - Delimited format file

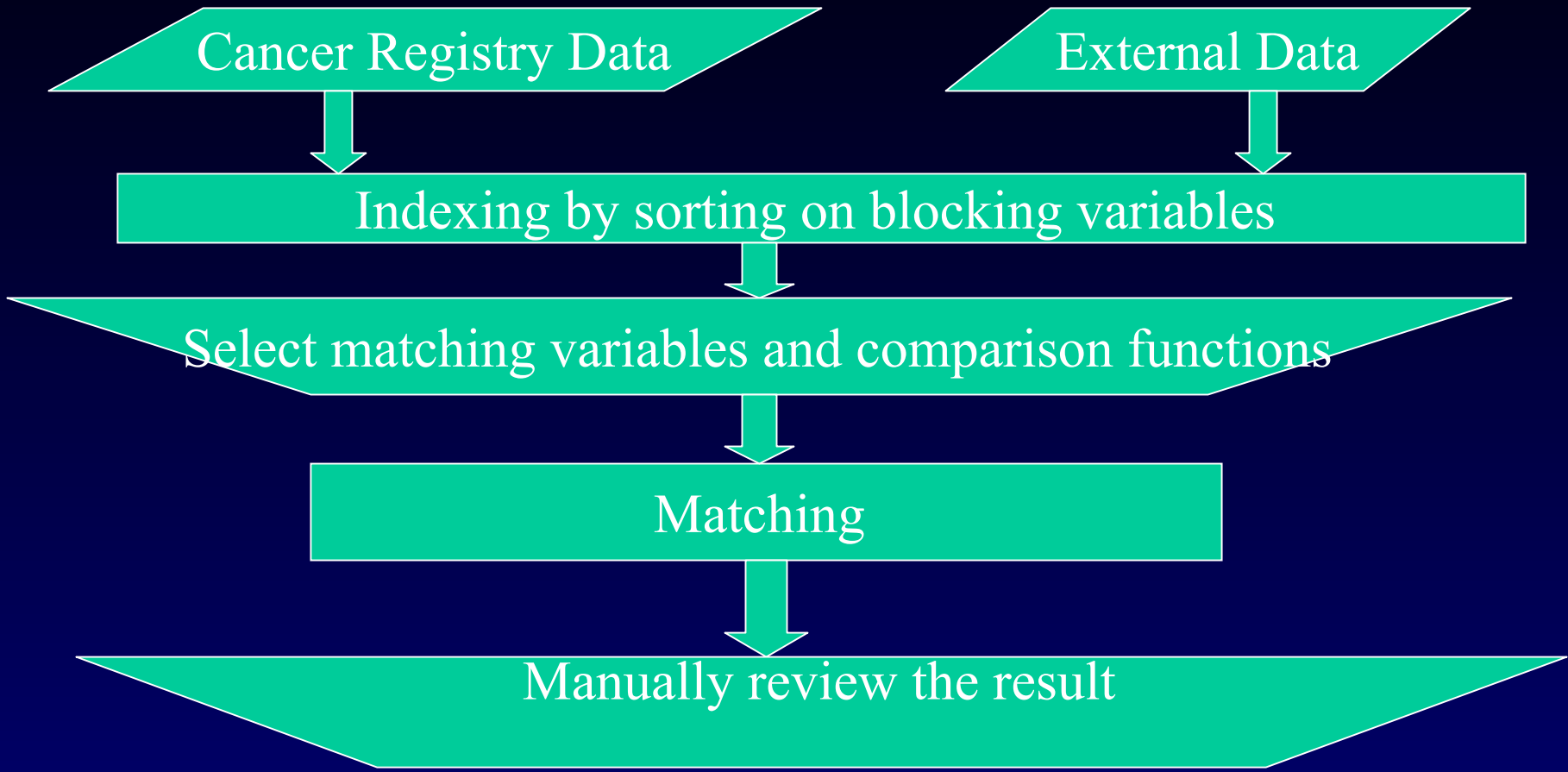
Link Plus: Features (continued)

- Customizable
  - Blocking criteria configurable
  - Matching variables easily selected
  - Several approximate string comparators available
  - Pre-set to use SSN, name, birth date, race and sex, but user- defined matching variables allowed

Link Plus: Features (continued)

- Fast
  - Linkage engine written in C++.
  - Effective sorting and indexing algorithms
  - Efficient scoring method by computing linkage score in two stages
    - 1: discard unlinked pairs from coarse computing-cheap match
    - 2: more expensive and accurate functions compute final linkage score
- Free of charge

Link Plus: Linking Process

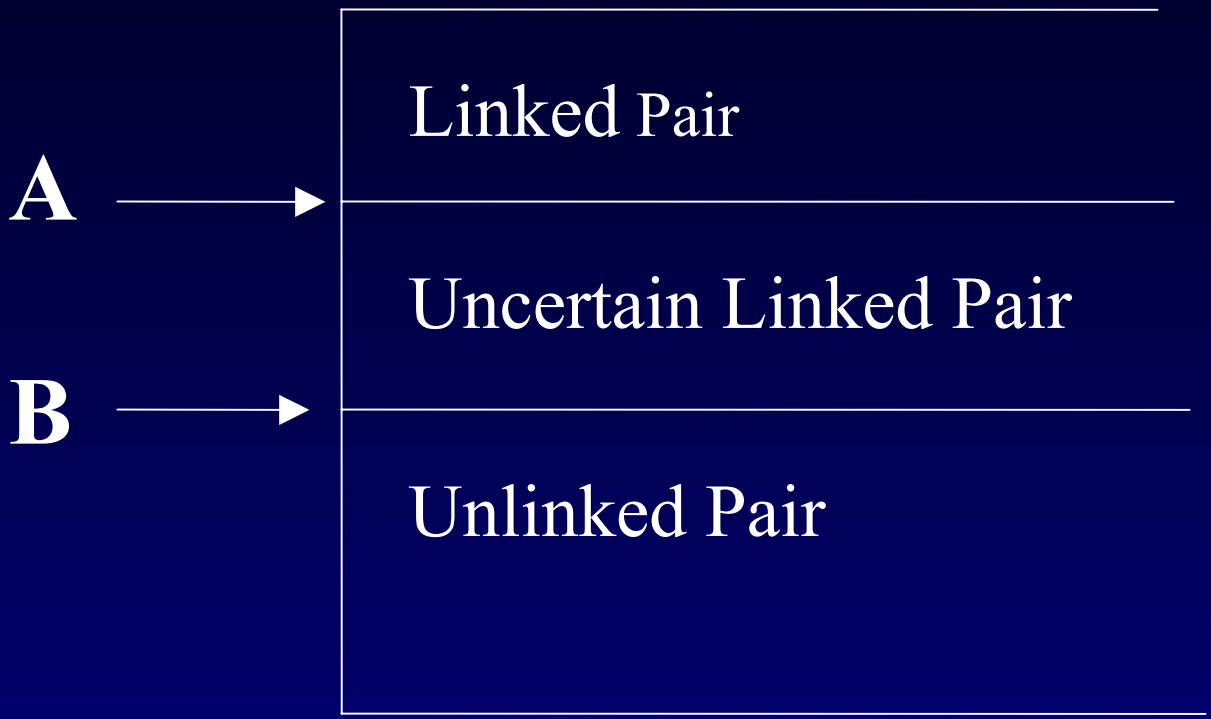


Link Plus: Blocking

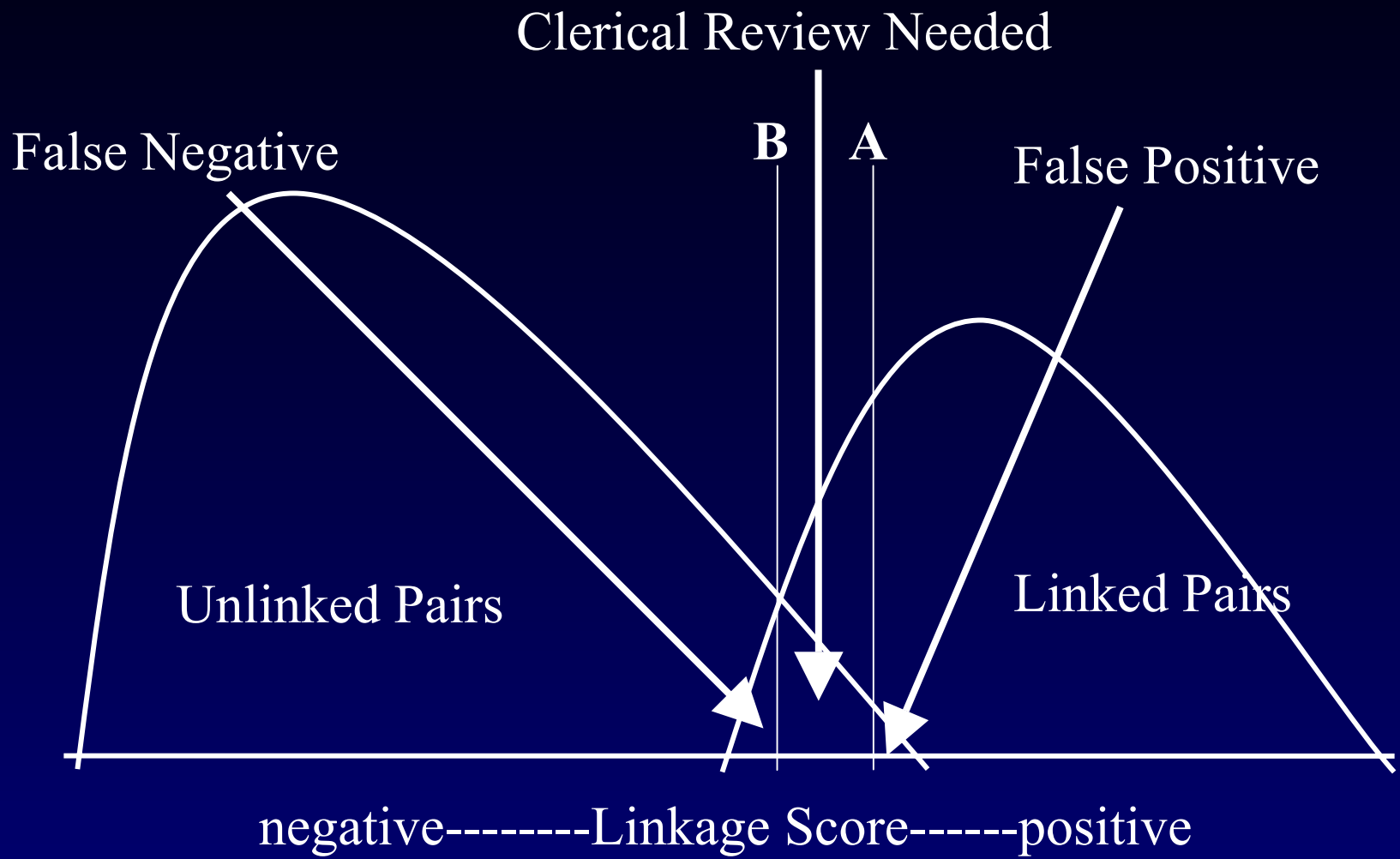
- For files with millions of records, total of all possible record pairs is too large
- Program portions files into blocks through sorting and indexing. Comparisons made between records in same block.
- Decrease in comparisons comes with cost of missing potential matches between records in different blocks
- Blocking is a trade-off between computing cost and missing potential matches

Finding Linked Records (1)

Cut-off Value



Finding Linked Records (2)



Link Plus: Sample Report (1)

report - Notepad

File Edit Format View Help

Matching Parameters

Matching Field	m-prob	u-prob	agree	disagree
Race1	0.98851	0.61456	0.43263	-3.19772
Sex	0.99610	0.50000	0.62738	-4.41868
BirthDate	0.95000	0.00005	8.89882	-2.72678
LastName	0.97339	0.00093	6.32906	-3.30026
FirstName	0.95166	0.00351	5.09885	-2.75428
MiddleName	0.96542	0.00439	4.90960	-3.05854
SocSec	0.95465	0.00000	14.62908	-2.81563

Cutoff Value=14.5

Match Report
Sorted by Score
Number of Matched Pairs = 506

PatientID	Race1	Sex	BirthDate	LastName	FirstName	MiddleName	SocSec	Score
100379	01	1	19630808	JUBINVILLE	THAD	LEIGH	408745011	
77090	01	1	19630808	JUBINVILLe	thad	LEIGH	408745011	48.5
100211	01	2	19790515	ENFIELD	FRIEDA	E	405763776	
44548	01	2	19790515	ENFIELD	FRIEDA	E	405763776	47.6
100125	02	2	19301019	SNEAD	SHERRI	CASSITY	403331729	
29206	02	2	19301019	SNEAD	SHERRI	CASSITY	403331729	45.7
100083	01	2	19710326	BURGOON	LYNDA	X	403034460	
21594	01	2	19710326	BURGOON	LYNDA	X	403034460	45.7
41054	01	2	19320622	BRUBECK	RHONDA	TRACE	405580860	
100189	01	2	19320622	BRUBECK	RHONDA	TRACE	405580860	45.7
31339	01	2	19320717	TROTTER	JENNA	JEANNINE	404316233	
100140	01	2	19320717	TROTTER	JENNA	JEANNINE	404316233	45.6
100371	01	1	19360521	SCHAEFFER	SALVATORE	AMY	408087161	
74411	01	1	19360521	SCHAEFFER	SALVATORE	AMY	408087161	45.6

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Link Plus: Sample Report (2)

report - Notepad

File Edit Format View Help

100476	01	1	19550324	SMITH	STEVE		412282869	
95863	01	1	19550324	SMITH	STEVE	Z	999999999	20.0
100055	01	2	99999999	HAMMONS	ROSE	BOSTOCK	999999999	
15725	01	2	19351216	HAMMONS	ROSE	BOSTOCK	402381385	19.9
49148	98	2	99999999	TUGGLE	JESSICA	SHEATS	405374703	
100232	98	2	19550527	TUGGLE	JESSICA	SHEATS	999999999	19.8
30317	01	2	19710212	CARDENAS	KELLI	T	999999999	
41792	01	2	19850108	CARDENAS	KELLI	T	405846977	18.7
23074	02	2	19390106	HOWERON	SOMER	MATTSON	999999999	
100091	01	2	19390106	HOWERTON	SOMER	MATTSON	403051908	17.8
100326	01	1	99999999	JIMENEZ	JOSEPH	WILMOT	407068436	
66256	01	1	19300904	JIMENEZ	JOSEPH	WILMOT	999999999	17.7
78837	01	1	99999999	THOMPSON	EDWIN	MCCOY	408807317	
100386	01	1	19330909	THOMPSON	EDWIN	MCCOY	999999999	17.6
30975	01	2	99999999	WASHINGTON	ELIZABETH	CORTEZ	999999999	
100137	01	2	19490603	WASHINGTON	ELIZABETH	CORTEZ	404129883	17.0
44465	01	2	19500312	MARANGONI	LILIAN	M	405160724	
100209	01	2	19501203	MARANGONI	MARIE	L	405160724	17.0
79258	01	1	19840619	CAIN	MATTHEW	N	999999999	
88353	01	1	19601209	CAIN	MATTHEW	N	409394246	16.1
42872	01	2	19361123	GRAY	MELISSA	J	999999999	
36271	01	2	19620619	GRAY	MELISSA	J	404520105	14.7
70791	01	1	19850407	SMITH	JAMES	N	408792016	
63246	01	1	19850407	SMITH	JAMES	O	999999999	14.6

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Link Plus: Use for Death Clearance

- Requires customization for each state's mortality file format and characteristics
- Has been alpha tested in one state, with 45K death records for one year

Link Plus: System Requirements

- **Operation System:**
  - Windows 95, 98, NT, 2000, XP
- **Processor:** Pentium II and above
- **Memory requirement**
  - 64 M for small size data files (<100,000 records)
  - 128 M for files having over 100,000 records
  - 512 M for files having over one million records
- **Data Set:** Flat file, Access, SQL Server, ORACLE, Fox Pro, Informix, Sybase.

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Link Plus: Performance

- Computer:
  - CPU: Pentium 4 1.7 Ghz
  - memory: 512 M
  - operating system: Windows 2000 Professional
- De-duplicating process time:
  - < 1 minute (100,000 records)
  - 10~15 minutes (500,000 records)
  - 40~60 minutes (1 million records)

Link Plus: Release Schedule

- Release to beta testers June 2003
- Public release later in summer
- Distribution via CDC web site or ftp site

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Description of the data file used for demo

- Size:
  - 100,500 records
- Number of duplicate record pairs:
  - 500
- Methods:
  - random sample using the 1990 Census frequency data of last name, first name, and race

Continued

- Generate missing values randomly:
  - SSN: 10% missing values
  - DOB: 2.5% missing all, 2.5% missing both month and day
  - middle name: 5% missing values
  - race: 0.5% missing values.
- Generate errors on the whole data randomly and then manually introduce various types of errors on duplicate pairs using realistic data.

Detail Report (1)

Patient ID	Race	Sex	DOB	Last name	First name	Middle name	SSN	Score
100288	04	1	19810713	HOLLINGSWORTH	LELAND	LUMLEY	407401447	
60670	04	1	19810713	HOLLINGSWORTH	LUMLEY		407401447	30.3
100095	01	2	19570701	GIVENS	BONICERAYE		403248771	
23547	01	2	19570701	GIVENS	BONICE		403248771	30.3
100455	01	1	19750212	BUSHMAN	CEDRIC		424378611	
92314	01	1	19750212	BUSHMAN	CEDRIC		525378611	30.3
76878	02	2	19890510	DOW	ROSIE		408513079	
100376	02	2	19890510	DOW	ROSA		408513079	30.2
100229	01	1	19771109	CARDENAS	CHARLEEN	DALEMBERTE	000551298	
48759	01	2	19771109	CARDENAS	CHARLEEN	DALEMBERTE	405551298	29.7
100483	98	1	19870717	MUSSELMANN	DARREN	J	412742447	
96497	98	1	19870717	MUSSELMAN	DARREN	JO	412742447	29.5
36534	01	2	19790708	TREADWAY	IRIS	KYLE	404314481	
100163	01	2	19790708	TREADAWAY	IRIS	K	404314481	29.4

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Detail Report (Continued)

100301	01	1	19510710	HEIM	MILES	A	407499225	
62262	01	1	19510712	HEIM	MYLES	A	407499225	24.6
83038	01	1	19580922	KNOW	LEONARDO		409243619	
100410	01	2	19580922	KNOX	LEONARDO		409243619	24.3
26096	01	2	19520210	MCDOWDRA	PATTY		403523106	
100110	01	1	19520210	MCDOWRA	PATTY	S	403523106	23.7
58077	01	1	19891226	MIZE	GRACE	MEYER	406135964	
100270	01	1	19891226	MIXE	GRACIE	MEYER	406135964	23.4
45923	01	1	19731213	SCARLET	RIOS	HARDY	405470782	
100216	01	1	19731213	RIOS	SCARLET	HARDY	405470782	23.4
94855	01	1	19610407	WILLIAMS	CORNS	THOMAS	412422620	
100469	01	1	19610407	WILLIAMS	THOMAS	C	412422620	23.1
83687	01	1	19660825	GOLACHUCK	YOUNG		409626424	
100413	01	1	19669999	GOLACHUK	YOUNG	KINCH	409626424	22.6
52686	01	1	19719999	SOTO	SERGIO	DOBLE	406087816	
100252	01	1	19711110	SOTO	SERGIO	DOBLE	999999999	22.2
54380	01	1	19770928	IMGRUND	WM	JULIUS	406935213	
100256	01	1	19770928	IMGRUND	WM	JULIS	306935213	21.8

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Detail Report (Continued)

30316	01	2	19710212	CARDENAS	KELLI	T	999999999	
41791	01	2	19850108	CARDENAS	KELLI	T	405846977	18.7
23073	02	2	19390106	HOWERON	SOMER	MATTSON	999999999	
100090	01	2	19390106	HOWERTON	SOMER	MATTSON	403051908	17.8
100325	01	1	99999999	JIMENEZ	JOSEPH	WILMOT	407068436	
66255	01	1	19300904	JIMENEZ	JOSEPH	WILMOT	999999999	17.7
78836	01	1	99999999	THOMPSON	EDWIN	MCCOY	408807317	
100385	01	1	19330909	THOMPSON	EDWIN	MCCOY	999999999	17.6
30974	01	2	99999999	WASHINGTON	ELIZABETH	CORTEZ	999999999	
100136	01	2	19490603	WASHINGTON	ELIZABETH	CORTEZ	404129883	17
44464	01	2	19500312	MARANGONI	LILIAN	M	405160724	
100208	01	2	19501203	MARANGONI	MARIE	L	405160724	17
79257	01	1	19840619	CAIN	MATTHEW	N	999999999	
88352	01	1	19601209	CAIN	MATTHEW	N	409394246	16.1
42871	01	2	19361123	GRAY	MELISSA	J	999999999	
36270	01	2	19620619	GRAY	MELISSA	J	404520105	14.7
70790	01	1	19850407	SMITH	JAMES	N	408792016	
63245	01	1	19850407	SMITH	JAMES	O	999999999	14.6

For More Information

David Gu: dfg2@cdc.gov

Joe Rogers: jdr0@cdc.gov

www.cdc.gov/cancer/npcr

DEPARTMENT OF HEALTH AND HUMAN SERVICES
CENTERS FOR DISEASE CONTROL AND PREVENTION
Atlanta, Georgia, USA