Probabilistic and Deterministic Data Linkage between Kentucky Cancer Registry and Health Claims Data

Bin Huang, University of Kentucky

NAACCR Annual Meeting,
June 14, 2016

UNIVERSITY OF KENTUCKY
College of Public Health

Background

- Due to limited resource, cancer registry data do not capture complete treatment information.
- Report from IOM (Cancer Patient to Cancer Survival: Lost in Transition) recommended great focus on linking data between registry and administrative data.
- Provide empirical foundation for better studied to access quality of care and outcome.
- A lot registries have done such data linkage (GA, OH, NY, WV,…)

Background – Data Linkage Methods

- Deterministic matching
  - Exact matching on key variables
  - Matching is defined by a predetermined algorithm
  - No consideration on how likely values are to agree by chance

- Probabilistic matching
  - Linkage scores based on properties of fields being matched
  - M-probability and U-probability
  - Linkage score depends on probability that two records match
  - A cutoff value will be used to define potential true matches.
  - Involve manual review

Background – Previous Studies

- Previous studies
  - WV: Deterministic data linkage for Medicaid and Medicare data
  - NY: Probabilistic data linkage for Medicaid data.
  - OH: Deterministic data linkage for Medicaid and Medicare data
  - GA: Deterministic data linkage for Medicaid and commercial claims data
- The characteristics of probabilistic and deterministic data linkage have not been thoroughly examined between registry data and claims data.
- Kentucky Cancer Registry (KCR) is conducting a study to link registry data and health claims data, such as Medicare, Medicaid and private insurance group claims data. This provides opportunity to examine the characteristics of two data linkage approaches.

CDC Registry Plus Link Plus

- Many probabilistic data linkage software packages are available.
- Registry Plus Link Plus is free and developed for data linkage between registry data and data from other sources.
- Link Plus Version 3.0 (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp_tech_info.htm)
  - Handle very large data sets (over 4 millions records for file1 and no limit on file 2)
  - Provide one-to-many matching and many-to-many matching
  - Nice manual review interface

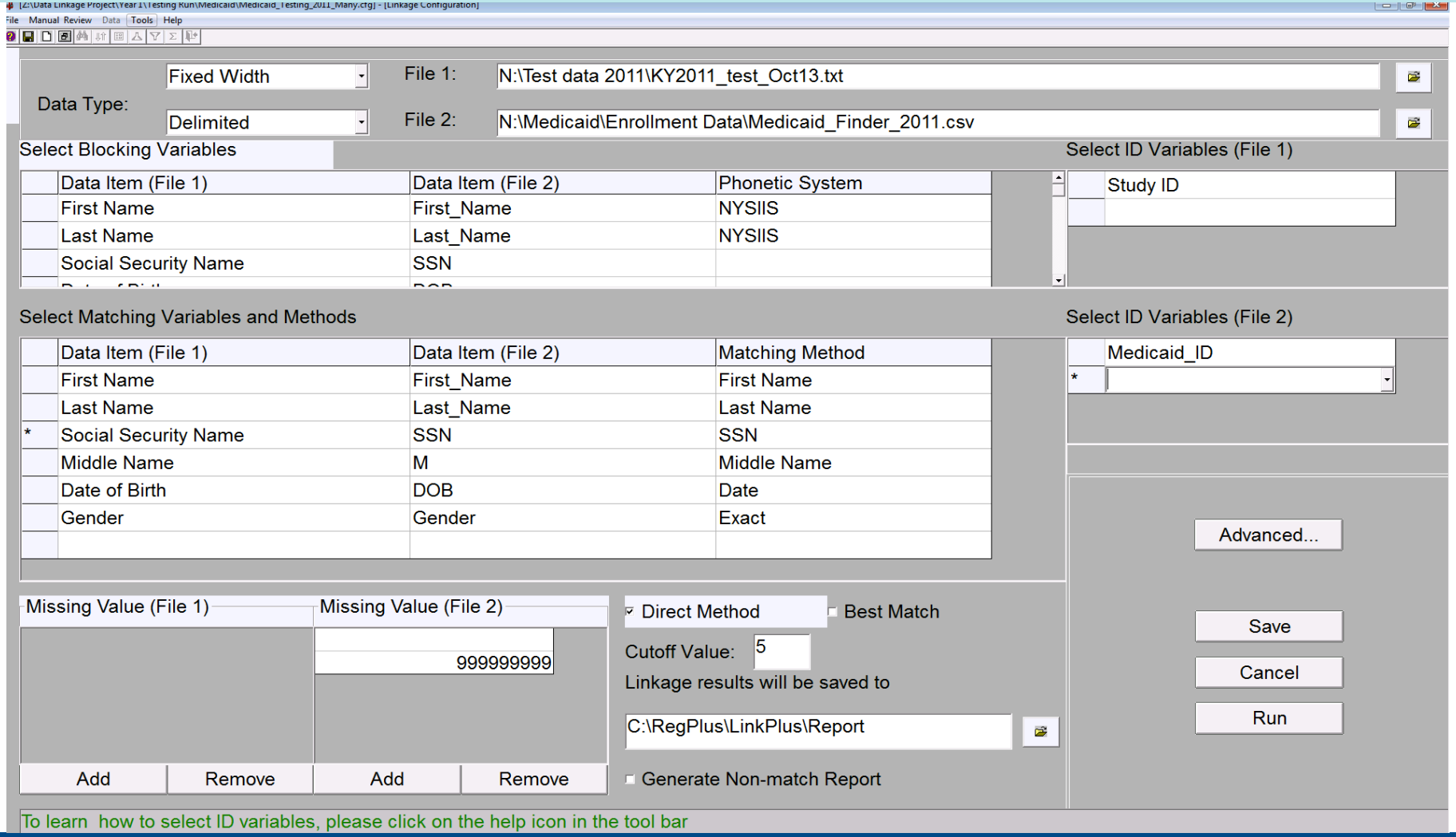
Data Involved in the Linkage

- KCR and Medicaid data linkage
  - KCR: 2011 KCR registry data for six cancer sites (10,887)
  - Medicaid: 2011 Medicaid enrollment data, including SSN, name, birth date, gender, county code. Race is not reliable. (1,051,987)
  - KCR has the access of the Medicaid enrollment data.
- KCR and Humana Data linkage
  - KCR: 2011 KCR registry data for six cancer sites
  - Humana: 2007-2012 Humana Enrollment data, including SSN, name, birth date, gender, specific address. Race is not reliable.
    - KY 1,714,465, IN 761,051, OH 1,339,385,
    - Rest neighboring states 3,696,381
  - KCR doesn't have the access of the Humana enrollment data, which make the linkage process very challenging.

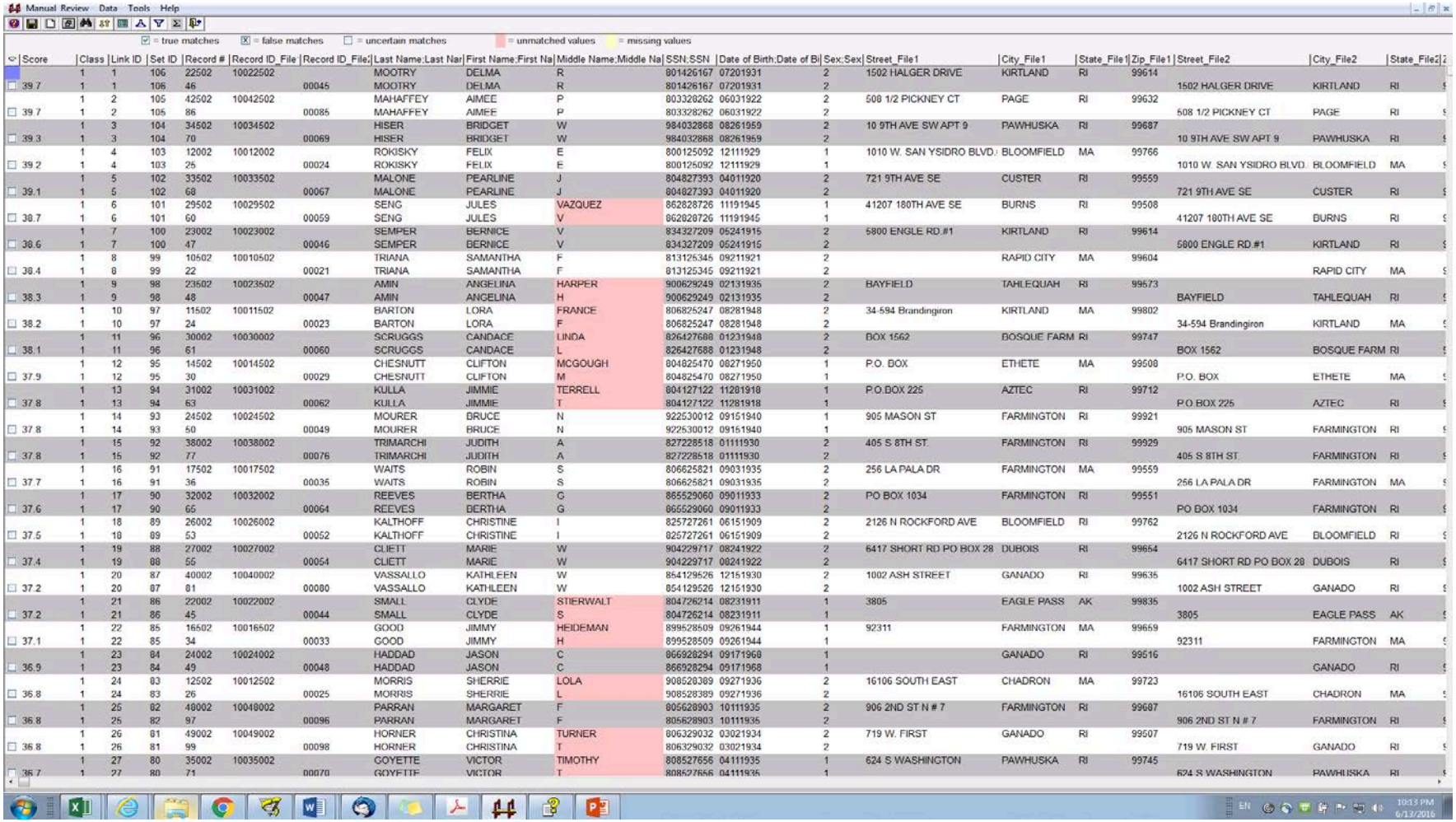
Process for Probabilistic Data Linkage

- Linking criteria:
  - Blocking variables: SSN, birthdate, first name, last name, middle name
  - Matching variables: SSN, birthdate, first name, last name, middle name and sex
  - Direct method: calculate M and U-probabilities from the distribution of file 1.
  - Cutoff values: lowest linkage score to be considered as a match (5)
  - 1-to-Many: Records in File 1 can match multiple records in File 2.
  - Many-to-Many: Records in both File 1 and File 2 can match multiple records.
- Manually review the potential matches to identify true matches

Link Plus – Screen Shot



Link Plus – Screen Shot



Manual Review Data Tools Help

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

Score	Class	Link ID	Set ID	Record #	Record ID_File	Record ID_File2	Last Name:Last Nar	First Name:First Na	Middle Name:Middle Na	SSN:SSN	Date of Birth:Date of Bi	Sex:Sex	Street_File1	City_File1	State_File1	Zip_File1	Street_File2	City_File2	State_File2
	1	1	106	22502	10022502		MOOTRY	DELMA	R	801426167	07201931	2	1502 HALGER DRIVE	KIRTLAND	RI	99614			
39.7	1	1	106	46		00045	MOOTRY	DELMA	R	801426167	07201931	2					1502 HALGER DRIVE	KIRTLAND	RI
	1	2	105	42502	10042502		MAHAFFEY	AIMEE	P	803328262	06031922	2	508 1/2 PICKNEY CT	PAGE	RI	99632			
39.7	1	2	105	86		00085	MAHAFFEY	AIMEE	P	803328262	06031922	2					508 1/2 PICKNEY CT	PAGE	RI
	1	3	104	34502	10034502		HISER	BRIDGET	W	984032868	08261959	2	10 9TH AVE SW APT 9	PAWHUSKA	RI	99687			
39.3	1	3	104	70		00069	HISER	BRIDGET	W	984032868	08261959	2					10 9TH AVE SW APT 9	PAWHUSKA	RI
	1	4	103	12002	10012002		ROKISKY	FELIX	E	800125092	12111929	1	1010 W. SAN YSIDRO BLVD.	BLOOMFIELD	MA	99766			
39.2	1	4	103	25		00024	ROKISKY	FELIX	E	800125092	12111929	1					1010 W. SAN YSIDRO BLVD.	BLOOMFIELD	MA
	1	5	102	33502	10033502		MALONE	PEARLINE	J	804827393	04011920	2	721 9TH AVE SE	CUSTER	RI	99559			
39.1	1	5	102	68		00067	MALONE	PEARLINE	J	804827393	04011920	2					721 9TH AVE SE	CUSTER	RI
	1	6	101	29502	10029502		SENG	JULES	VAZQUEZ	862828726	11191945	1	41207 180TH AVE SE	BURNS	RI	99508			
38.7	1	6	101	60		00059	SENG	JULES	V	862828726	11191945	1					41207 180TH AVE SE	BURNS	RI
	1	7	100	23002	10023002		SEMPER	BERNICE	V	834327209	05241915	2	5800 ENGLE RD.#1	KIRTLAND	RI	99614			
38.6	1	7	100	47		00046	SEMPER	BERNICE	V	834327209	05241915	2					5800 ENGLE RD.#1	KIRTLAND	RI
	1	8	99	10502	10010502		TRIANA	SAMANTHA	F	813125345	09211921	2		RAPID CITY	MA	99604			
38.4	1	8	99	22		00021	TRIANA	SAMANTHA	F	813125345	09211921	2						RAPID CITY	MA
	1	9	98	23502	10023502		AMIN	ANGELINA	HARPER	900629249	02131935	2	BAYFIELD	TAHLEQUAH	RI	99673			
38.3	1	9	98	48		00047	AMIN	ANGELINA	H	900629249	02131935	2					BAYFIELD	TAHLEQUAH	RI
	1	10	97	11502	10011502		BARTON	LORA	FRANCE	806825247	08281948	2	34-594 Brandingiron	KIRTLAND	MA	99802			
38.2	1	10	97	24		00023	BARTON	LORA	F	806825247	08281948	2					34-594 Brandingiron	KIRTLAND	MA
	1	11	96	30002	10030002		SCRUGGS	CANDACE	LINDA	826427688	01231948	2	BOX 1562	BOSQUE FARM	RI	99747			
38.1	1	11	96	61		00060	SCRUGGS	CANDACE	L	826427688	01231948	2					BOX 1562	BOSQUE FARM	RI
	1	12	95	14502	10014502		CHESNUTT	CLIFTON	MCGOUGH	804825470	08271950	1	P.O. BOX	ETHETE	MA	99508			
37.9	1	12	95	30		00029	CHESNUTT	CLIFTON	M	804825470	08271950	1					P.O. BOX	ETHETE	MA
	1	13	94	31002	10031002		KULLA	JIMMIE	TERRELL	804127122	11281918	1	P.O.BOX 225	AZTEC	RI	99712			
37.8	1	13	94	63		00062	KULLA	JIMMIE	T	804127122	11281918	1					P.O.BOX 225	AZTEC	RI
	1	14	93	24502	10024502		MOURER	BRUCE	N	922530012	09151940	1	905 MASON ST	FARMINGTON	RI	99921			
37.8	1	14	93	50		00049	MOURER	BRUCE	N	922530012	09151940	1					905 MASON ST	FARMINGTON	RI
	1	15	92	38002	10038002		TRIMARCHI	JUDITH	A	827228518	01111930	2	405 S 8TH ST.	FARMINGTON	RI	99929			
37.8	1	15	92	77		00076	TRIMARCHI	JUDITH	A	827228518	01111930	2					405 S 8TH ST.	FARMINGTON	RI
	1	16	91	17502	10017502		WAITS	ROBIN	S	806625821	09031935	2	256 LA PALA DR	FARMINGTON	MA	99559			
37.7	1	16	91	36		00035	WAITS	ROBIN	S	806625821	09031935	2					256 LA PALA DR	FARMINGTON	MA
	1	17	90	32002	10032002		REEVES	BERTHA	G	865529060	09011933	2	PO BOX 1034	FARMINGTON	RI	99551			
37.6	1	17	90	65		00064	REEVES	BERTHA	G	865529060	09011933	2					PO BOX 1034	FARMINGTON	RI
	1	18	89	26002	10026002		KALTHOFF	CHRISTINE	I	825727261	06151909	2	2126 N ROCKFORD AVE	BLOOMFIELD	RI	99762			
37.5	1	18	89	53		00052	KALTHOFF	CHRISTINE	I	825727261	06151909	2					2126 N ROCKFORD AVE	BLOOMFIELD	RI
	1	19	88	27002	10027002		CLIETT	MARIE	W	904229717	08241922	2	6417 SHORT RD PO BOX 28	DUBOIS	RI	99664			
37.4	1	19	88	55		00054	CLIETT	MARIE	W	904229717	08241922	2					6417 SHORT RD PO BOX 28	DUBOIS	RI
	1	20	87	40002	10040002		VASSALLO	KATHLEEN	W	854129526	12151930	2	1002 ASH STREET	GANADO	RI	99636			
37.2	1	20	87	81		00080	VASSALLO	KATHLEEN	W	854129526	12151930	2					1002 ASH STREET	GANADO	RI
	1	21	86	22002	10022002		SMALL	CLYDE	STIERWALT	804726214	08231911	1	3805	EAGLE PASS	AK	99835			
37.2	1	21	86	45		00044	SMALL	CLYDE	S	804726214	08231911	1					3805	EAGLE PASS	AK
	1	22	85	16502	10016502		GOOD	JIMMY	HEIDEMAN	899528509	09261944	1	92311	FARMINGTON	MA	99659			
37.1	1	22	85	34		00033	GOOD	JIMMY	H	899528509	09261944	1					92311	FARMINGTON	MA
	1	23	84	24002	10024002		HADDAD	JASON	C	866928294	09171968	1		GANADO	RI	99516			
36.9	1	23	84	49		00048	HADDAD	JASON	C	866928294	09171968	1						GANADO	RI
	1	24	83	12502	10012502		MORRIS	SHERRIE	LOLA	908528389	09271936	2	16106 SOUTH EAST	CHADRON	MA	99723			
36.8	1	24	83	26		00025	MORRIS	SHERRIE	L	908528389	09271936	2					16106 SOUTH EAST	CHADRON	MA
	1	25	82	48002	10048002		PARRAN	MARGARET	F	805628903	10111935	2	906 2ND ST N # 7	FARMINGTON	RI	99687			
36.8	1	25	82	97		00096	PARRAN	MARGARET	F	805628903	10111935	2					906 2ND ST N # 7	FARMINGTON	RI
	1	26	81	49002	10049002		HORNER	CHRISTINA	TURNER	806329032	03021934	2	719 W. FIRST	GANADO	RI	99507			
36.8	1	26	81	99		00098	HORNER	CHRISTINA	T	806329032	03021934	2					719 W. FIRST	GANADO	RI
	1	27	80	35002	10035002		GOYETTE	VICTOR	TIMOTHY	808527656	04111935	1	624 S WASHINGTON	PAWHUSKA	RI	99745			
36.7	1	27	80	71		00070	GOYETTE	VICTOR	T	808527656	04111935	1					624 S WASHINGTON	PAWHUSKA	RI

Deterministic Data Linkage

- Following variables are included in the deterministic data linkage: SSN, date of birth, gender, last name, first name (both truncated to the first 6 letters).*
- Following process is used to identify matches:
  - Step 1: SSN, Last Name, First Name (Type 1)
  - Step 2: SSN, Last Name, Date of Birth (Month), Gender (Type 2)
  - Step 3: SSN, First Name, Date of Birth (Month), Gender (Type 3)
  - Step 4: First Name, Last Name, Date of Birth (Month and Year), Gender (type 4)

*SM Koroukian. Linking the Ohio Cancer Incidence Surveillance System with Medicare, Medicaid, and Clinical Data from Home Health Care and Long Term Care Assessment Instruments: Paving the Way for New Research Endeavors in Geriatric Oncology. J Registry Manag. 2008 Winter; 35(4): 156–165.

Identifying True Matches

- Medicaid linkage
  - Combine “true” matches resulted from manual review process: one-to-many matches and many-to many.
  - Further check matches from the deterministic approach
- Humana linkage
  - Identify “true” matches resulted from manual review process (many-to many matches) by state
  - Further check matches from the deterministic approach
  - Combine “true” matches from all states

Results – Medicaid (Probabilistic)

- There are total 2644 true matches. One-to-many and Many-to-Many only add one extra true case individually.

Score	1 to M N (Matches)	1 to M N (True Matches)	1 to M % of True Matches	M to M N (Matches)	M to M N (True Matches)	M to M % of True Matches
20+	2591	2591	100.0%	2591	2591	100.0%
18-19.9	19	19	100.0%	19	19	100.0%
15-17.9	23	23	100.0%	23	23	100.0%
13-14.9	18	6	33.3%	18	7	38.9%
11-12.9	123	4	3.3%	119	3	2.5%
10-10.9	180	0	0.0%	175	0	0.0%

Results – Medicaid (Deterministic)

- Total 2650 matches identified by deterministic matching with 2631 true matches. Sensitivity = 0.995, specificity = 0.998, PPV = 0.980, NPV=0.997

Match Method	N (Matches)	N (True Matches)	% of True Matches
SSN, Last Name, First Name	2485	2485	100.0%
SSN, Last Name, Month of Birth, Gender	71	71	100.0%
SSN, First Name, Month of Birth, Gender	40	40	100.0%
Last Name, First Name, Month of Birth, Year of Birth, Gender	54	35	64.8%
Total	2650	2631	99.3%

Results – KY Humana (Probabilistic)

- There are total 3825 true matches. 3784 from KY claims, 12 IN, 13 OH and 16 neighboring states. One-to-many from KY claims adds 5 extra true matches, and Many-to-Many from KY claims adds 17 extra true case separately.

Score	1 to M N (Matches)	1 to M N (True Matches)	1 to M % of True Matches	M to M N (Matches)	M to M N (True Matches)	M to M % of True Matches
20+	3416	3414	99.9%	3416	3416	100.00%
18-19.9	93	90	96.8%	93	93	100.00%
15-17.9	176	159	90.3%	176	165	93.8%
13-14.9	211	88	41.7%	204	85	41.7%
11-12.9	694	25	3.6%	686	27	4.0%
10-10.9	978	4	0.4%	979	4	0.4%
9-9.9	2193	2	0.0%	2215	4	0.0%

Results – Other States Humana (Probabilistic)

	IN			OH			Neighboring States		
Score	N Matches	N True Matches	% of True Matches	N Matches	N True Matches	% of True Matches	N Matches	N True Matches	% of True Matches
20+	13	13	100.0%	12	12	100.0%	16	16	100.0%
18-19.9	3	1	33.3%	0	0	0.0%	0	0	0.0%
15-17.9	7	0	0.0%	10	1	10.0%	10	0	0.0%
13-14.9	43	0	0.0%	73	0	0.0%	73	0	0.0%
11-12.9	323	0	0.0%	473	0	0.0%	473	0	0.0%
10-10.9	642	0	0.0%	718	0	0.0%	718	0	0.0%
9-9.9	1598	0	0.0%	1648	0	0.0%	1648	0	0.0%

Results – Humana (Deterministic)

- Deterministic matching identified 3761 true matches out of total 3825. Excluding type4 matches from out of state claims, Sensitivity = 0.994, specificity = 0.990, PPV = 0.981, NPV=0.990

Match Method	Kentucky		Indiana		Ohio		Neighboring States	
	N	N_True	N	N_True	N	N_True	N	N_True
SSN, Last Name, First Name	3262	3262	13	13	11	11	16	16
SSN, Last Name, Month of Birth, Gender	112	112	1	1	0	0	0	0
SSN, First Name, Month of Birth, Gender	41	41	0	0	1	1	0	0
Last Name, First Name, Month of Birth, Year of Birth, Gender	418	346	50	0	98	0	273	0
Total	3833	3761	64	14	110	12	289	16

Summary

- Link Plus is relatively stable and fast.
- One to Many and Many to Many matching generated almost identical results. Many to Many may provide slight more matches.
- Cut-off values for linkage score can be set at 10 or higher. It is very rare a true cases can be identified with a value below 10. This depends on the quality of data, particularly the quality of SSN.
- Deterministic matching provides comparable result compared to the probabilistic data linkage.
- Choosing the best algorithm to use depends on many interacting factors, such as time, resources, data access, quality of matching variables.

If Manual Review is Not Possible

- "True" match if linkage score > 18 or it is a deterministic match
  - Medicaid: sensitivity = 0.997, specificity = 0.998
    PPV = 0.993, NPV = 0.999
  - Humana: sensitivity = 0.999, specificity = 0.990
    PPV = 0.982, NPV = 0.999
- This algorithm provides better linkage than the deterministic matching.

Acknowledgements

Drs. Eric Tai, Blythe Ryerson, David Butterworth, Centers for Disease control and Preventions

Drs. Kevin Ward, Joseph Lipscomb, Emory University

Drs. Thomas Tucker, Quan Chen, Jaclyn Nee, University of Kentucky

Richard Maiti, Pete Landfield, Kentucky Family Services Department for Medicaid Service

Nirmal Subramanian, Dr. Robert Dufour, Humana

Funding Support:

CDC U48DP005014-01 SIP14-017

NCI SEER HHSN261201000031

Thanks!

Bin Huang

Kentucky Cancer Registry

bhuang@kcr.uky.edu