National Center for Health Statistics
Data Linkage

Comparing Multiple SAS® Functions for Text Field Matching in Data Linkage: SOUNDEX, NYSIIS, COMPGED

Yu Sun and Cordell Golden
National Center for Health Statistics, Centers for Disease Control and Prevention

Background:

NCHS Data Linkage Program

- Links survey data with vital and administrative records
- Designed to maximize the scientific value of the NCHS population-based surveys

Motivation:

- Previous linkage algorithms relied heavily on 9 digits of Social Security Number (SSN9)
- Current algorithms are more dependent on name variables due to changes in the way personally identifiable information is collected, only last 4 digits of SSN (SSN4)
- This analysis will assess the value added by incorporating phonetic algorithms and string comparator functions for text field matching in SAS® rather than exact matches

Data Sources:

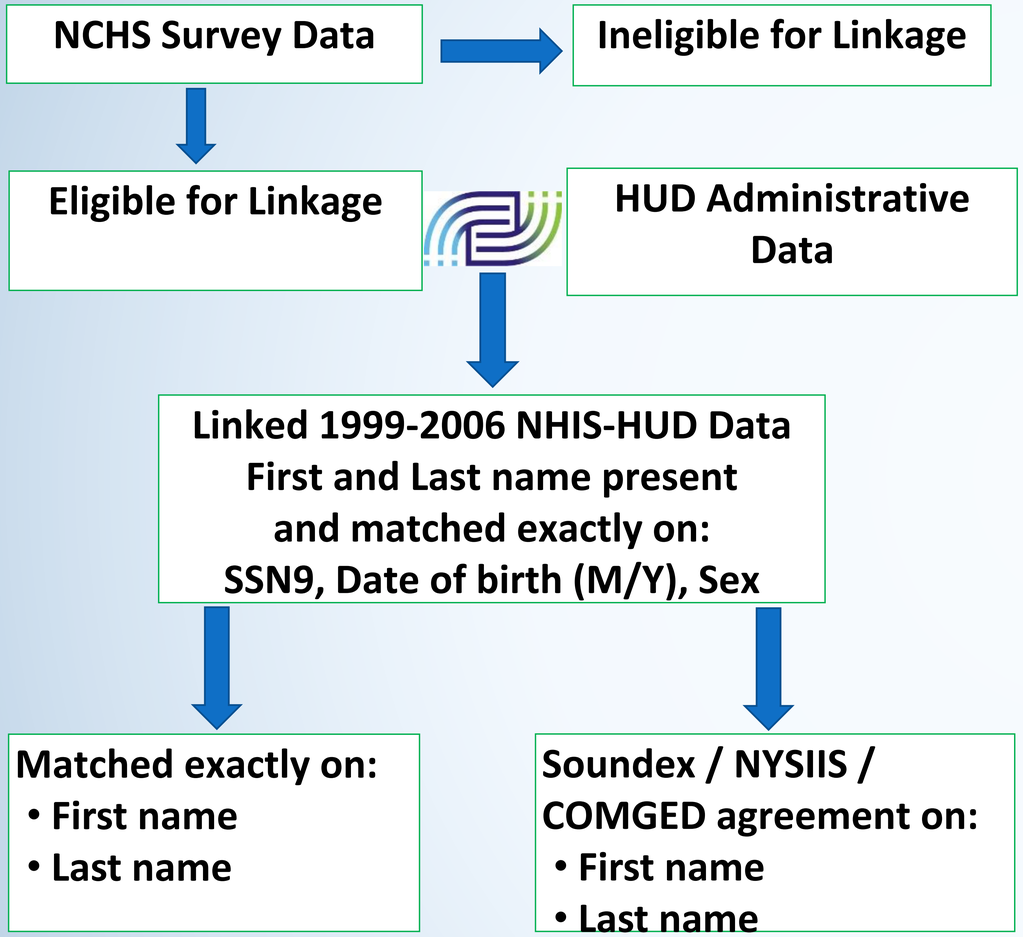
National Health Interview Survey (NHIS)

Nationally representative, cross-sectional household interview survey conducted by NCHS that serves as an important source of information on the health of the civilian, noninstitutionalized population in the U.S.

Housing and Urban Development (HUD)

Agency responsible for overseeing and managing domestic housing programs and policies, including specialized programs for high-needs populations (e.g., the elderly, homeless, and disabled) in the U.S.

Methods



Code used for phonetic and string comparison of First (FN) and Last names (LN):

SOUNDEX

SDX=1: SOUNDEX(FN1)=SOUNDEX(FN2) and
SOUNDEX(LN1)=SOUNDEX(LN2);

NYSIIS

NYS=1: %NYSIIS(name, name_NYS); FN1_NYS=FN2_NYS and
LN1_NYS=LN2_NYS;

COMPGED

Ged=1: (COMPGED(FN1, FN2)<=100 or COMPGED(FN2,FN1)<=100) and (COMPGED(LN1, LN2)<=100 or COMPGED(LN2,LN1)<=100) ;

x_ged=COMPGED(x, y); y_ged=COMPGED(y, x);

Goal: Compare the matches when requiring an exact match **vs.** using phonetic and string comparators

Results

Exact match	GED	SDX	NYS	N	%
1	1	1	1	13,188	86.62
	1	1	1	910	5.98
	1	1		396	2.60
	1		1	47	0.31
	1			455	2.99
		1	1	129	0.85
		1		81	0.53
			1	19	0.12

Value added: 13.4% (n=2,037) additional matches were captured with the phonetic and string comparators

Of the 2,037 additional matches:
44.7% were captured using by all 3 functions
45.7% were captured by SOUNDEX and/or COMPGED
9.6% were captured by NYSIIS only or a combination of NYSIIS and either SOUNDEX or COMPGED

An example of name matched by **NYSIIS only**

x	y	x_SDX	y_SDX	x_GED	y_GED	x_NYS	y_NYS
Catherine	Katherine	C365	K365	200	200	CATARAN	CATARAN

An example of name matched by **SOUNDEX only**

x	y	x_SDX	y_SDX	x_GED	y_GED	x_NYS	y_NYS
Maricela	Marisella	M624	M624	120	120	MARACAL	MARASAL

Example of names matched by **COMPGED only**

x	y	x_SDX	y_SDX	x_GED	y_GED	x_NYS	y_NYS
Hermon-Sisco	Hermon	H65522	H655	280	60	HARNAN-SASC	HARNAN
Pat	Patricia	P3	P362	50	250	PAT	PATRAC
Zheng	Zhen	Z52	Z5	50	10	ZANG	ZAN

An example of name matched by **ALL** three functions

x	y	x_SDX	y_SDX	x_GED	y_GED	x_NYS	y_NYS
Brian	Bryan	B65	B65	100	100	BRAN	BRAN

Conclusions

- **All three functions have their unique strengths and were able to identify matches not picked up when an exact match on name was required**
  - SOUNDEX and NYSIIS are good in matching names that sound alike and are spelled similarly
  - NYSIIS accounts for differences in the first letter, but SOUNDEX does not
  - COMPGED was better at handling multi-part last names; abbreviations and nicknames; and ethnic and non-traditional spelling variations
- **A combination of all three functions appears to work best**
- **NCHS will continue to research other name comparison functions and algorithms (e.g. Jaro-Winkler, SPEDIS, Perl) for data linkages using text fields**

National Center for Health Statistics
Data Linkage

REFERENCES

COMPGED Function
http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a002206133.htm
SOUNDEX Function
http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000245948.htm
NYSIIS
http://www.dropby.com/NYSIIS.html
Amanda Roesch, Matching data using Sounds-Like Operators and SAS compare functions, SAS Global Forum 2012 Paper 122-2012

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Lisa Mirel, Eileen Call, Jim Brittain and the Special Projects Branch for their contributions to this work.

CONTACT

NCHS Data Linkage Program: datalinkage@cdc.gov