National Center for Health Statistics
Data Linkage

Evaluating Hispanic and Asian Name Algorithms as a Tool for Improving Linkage to the National Death Index

Dean H. Judson,
Eric Miller,
Jennifer D. Parker,
Hannah Day
National Center for Health Statistics (NCHS)

2013 Federal Committee on Statistical Methodology

Disclaimer: These analyses were performed on an “almost final” file; a small handful of cases still need final resolution.

Outline

- Describe goals and process
- Describe innovations
- Focus on name algorithms/SSN4's
- Describe operational results and evaluate effects
- Think about future evaluations

NCHS Linked Mortality Files

- Last released in 2009 (up to NDI data year 2006), linkage of survey respondents from NCHS surveys (now ~2.5M records) to the National Death Index (~70M records) to ascertain mortality status & causes of death
- Auxiliary data from the Social Security Death Master File (DMF) also used to ascertain status
- Used for passive follow-up epidemiological studies

Process

- Construct submission files
  - Up to ~50 alternative records per person accounting for:
    - Alternative names (nicknames, middle/first flip)
    - Alternative dates of birth (esp. with imputations)
    - Alternative identifying numbers
    - Unique Hispanic and Asian naming conventions
- Submit to NDI linkage
- Potential links returned

Process (continued)

- Reweight and rescore
  - Construct detailed weights based on name, dates, digits, and other characteristics' frequencies
  - $Wt_i = \log_2(1/\text{Proportion})$, prorate for inexact matches
  - Sum weights and select "best" (highest scoring) pair
  - Tentatively flag records above threshold
- Rule-based post-processing
  - Alternative to clerical review
- Clerical review

Innovations this time

- Hispanic and Asian name algorithms
- Hispanic nickname tables
- New use of phonetic algorithms:
  - NYSIIS, Soundex, Jaro-Winkler score, Edit distance, "within"
- Use of enhanced matching macro (citation: Dean Resnick) to search the DMF
- Enhanced weights
- Rule-based post-processing
- Expedited/Computerized clerical review
- 4-digit SSN NDI search (completely new)

Example of potential for linkage error - Hispanic Paradox

- Hispanics have lower mortality rates compared to non-Hispanic whites
  - Health selective immigration
  - Salmon bias (return migration)
  - Advantageous health behaviors and social support
  - **Data quality / Insufficient linkage**
    - Naming conventions for Hispanics differ from other US populations
      - Use of mother's and father's surname
      - May not have single middle name
    - Less likely to have social security number
      - Especially among older adults and foreign born

Methods – Name Clean-up

- Accounting for Hispanic and Asian naming conventions
  - Hispanic
    - Hispanic nickname lookup table
    - switch middle and last
  - Asian
    - switch first and last

Hispanic Nickname Table

Sex	Formal Name	Nicknames						
F	*Adelina*	*Deli*	*Lina*					
F	*Adelaida*	*Ade*	*Adela*					
M	*Adrián*	*Adri*						
F	*Adriana*	*Adri*						
M	***Alberto***	*Alber*	*Albertito*	Beto	Berto	Tico	Tuco	**Tito**
M	*Alejandro*	*Ale*	*Álex*	Alejo	Jandro	Jano	Sandro	
F	*Alejandra*	*Sandra*	*Ale*	Álex	Aleja	Jandra	Jana	
M	*Alfonso*	*Alfon*	*Fon*	Fonso	Fonsi	Poncho		
F	*Alicia*	*Ali*	*Licha*					

NDI Selection Algorithm

NDI Selection Criteria			
	Criterion one	Criterion two	Criterion three
Social Security Number	X		X
First name	X	X	
Middle initial			
Last name	X	X	
Day of birth		X	X
Month of birth		X	X
Year of birth		X	X
Sex			X
Father's surname			
Race			
State of residence			
State of birth			
Marital Status			

Basic Tabulations

```
-> eligible = NO
Rejected by |
        NDI |      Freq.     Percent        Cum.
------------+-----------------------------------
         NO |         169        0.12        0.12
        YES |     143,835       99.88      100.00
------------+-----------------------------------
      Total |     144,004      100.00
-> eligible = YES
Rejected by |
        NDI |      Freq.     Percent        Cum.
------------+-----------------------------------
         NO |   2,478,065       99.09       99.09
        YES |      22,862        0.91      100.00
------------+-----------------------------------
      Total |   2,500,927      100.00
```


National Center for Health Statistics
Data Linkage

Eligible Record Dispositions

Final disposition class, among eligibles	Final NDI Status ALIVE	DEAD	Total
Assumed alive	2,087,704	0	2,087,704
	100.00	0.00	84.25
Dead by score/class	0	382,334	382,334
	0.00	97.94	15.43
Dead by rule	0	3,327	3,327
	0.00	0.85	0.13
Dead by manual review	0	1,848	1,848
	0.00	0.47	0.07
Dead by mult. sources	0	2,852	2,852
	0.00	0.73	0.12
Total	2,087,704	390,361	2,478,065
	100.00	100.00	100.00

Rescore

- Five “classes”
  - One: Full hit on 8 or 9 fields, incl. SSN
  - Two: SSN present almost exact
  - Three: SSN unknown, 8 or 9 other fields
  - Four: SSN unknown, <8 other fields
  - Five: SSN < 8 digits, not in classes one-four
- Scores = sum of fields weights with proration for partial matches
- Thresholds by class (NHEFS)

Eligible Record Dispositions by Class

Final disposition class, among eligibles	Rescore Class 1	2	3	4	5	.	Total
Assumed alive	4	1,228	23,083	802,162	474,065	787,162	2,087,704
	0.00	1.32	24.00	98.00	99.73	100.00	84.25
Dead by score/class	207,690	90,392	71,032	13,220	0	0	382,334
	100.00	97.01	73.87	1.62	0.00	0.00	15.43
Dead by rule	0	966	1,265	760	336	0	3,327
	0.00	1.04	1.32	0.09	0.07	0.00	0.13
Dead by manual review	0	137	543	1,159	9	0	1,848
	0.00	0.15	0.56	0.14	0.00	0.00	0.07
Dead by mult. sources	0	453	238	1,225	936	0	2,852
	0.00	0.49	0.25	0.15	0.20	0.00	0.12
Total	207,694	93,176	96,161	818,526	475,346	787,162	2,478,065
	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Innovations

- Rules
  - E.g. Rule 2: if social security matches on 8 or 9, last name matches, sex matches, exact DOB, and birth state matches or is missing, then match=yes (dead).
- Collectively referred to as “autodead”, “autoalive”, “autoreview”
- Clerical review screens to follow

Dispositions by “Auto” Rules

Final disposition class, among eligibles	Auto-dead rule fired N	Y	.	Total
Assumed alive	1,299,062	1,480	787,162	2,087,704
	99.56	0.38	100.00	84.25
Dead by score/class	979	381,355	0	382,334
	0.08	98.76	0.00	15.43
Dead by rule	0	3,327	0	3,327
	0.00	0.86	0.00	0.13
Dead by manual review	1,848	0	0	1,848
	0.14	0.00	0.00	0.07
Dead by mult. sources	2,852	0	0	2,852
	0.22	0.00	0.00	0.12
Total	1,304,741	386,162	787,162	2,478,065
	100.00	100.00	100.00	100.00

Dispositions by “Auto” Rules

Final disposition class, among eligibles	Auto-alive rule fired 0	1	.	Total
Assumed alive	797,632	502,910	787,162	2,087,704
	67.15	99.95	100.00	84.25
Dead by score/class	382,331	3	0	382,334
	32.19	0.00	0.00	15.43
Dead by rule	3,327	0	0	3,327
	0.28	0.00	0.00	0.13
Dead by manual review	1,828	20	0	1,848
	0.15	0.00	0.00	0.07
Dead by mult. sources	2,631	221	0	2,852
	0.22	0.04	0.00	0.12
Total	1,187,749	503,154	787,162	2,478,065
	100.00	100.00	100.00	100.00

Manual Review Form

Submission Record | DMF Record | String Comparators

Submission	Name	Intl.	Match	NDI	Name	Intl.
First:	D	D	X		Dean	D
Middle:	Harold	H	I		H	H
Last:	Judson		X		Judson	
Surname:					Judson	

Exact Match:

FSN-LN

LN-FSN

	Submission	Matching Digits:		NDI
9-Digit:	XXXXX2654	-----XXXX	4	XXXXX2654
4-Digit:	2654			2654

Birth Date	Sub.	Match	NDI	NDI Death Date
Month:	07	X	07	12
Day:	20	X	20	24
Year:	1962	-1	1963	2011

	Sub.	Match	NDI
Sex:	M	X	M
ST Birth:	CT		MI
ST Res:	MD	X	MD
Race:	White	X	White
Marital:	Married	?	Unknown
Dth Age:		-	48

P_Dead_2011: 0.131400 ST Dth: VA

Linkage ID: 10000000000001

Reviewer: VVV6 Date:

Last Alive Year: 2008

Rec. # 1156

SCORE: 45

CLASS: 2

Review Notes:

Reason: DEAN'S SPECIAL PII

Review:

- No Match
- Unsure
- Match
- DMF Match

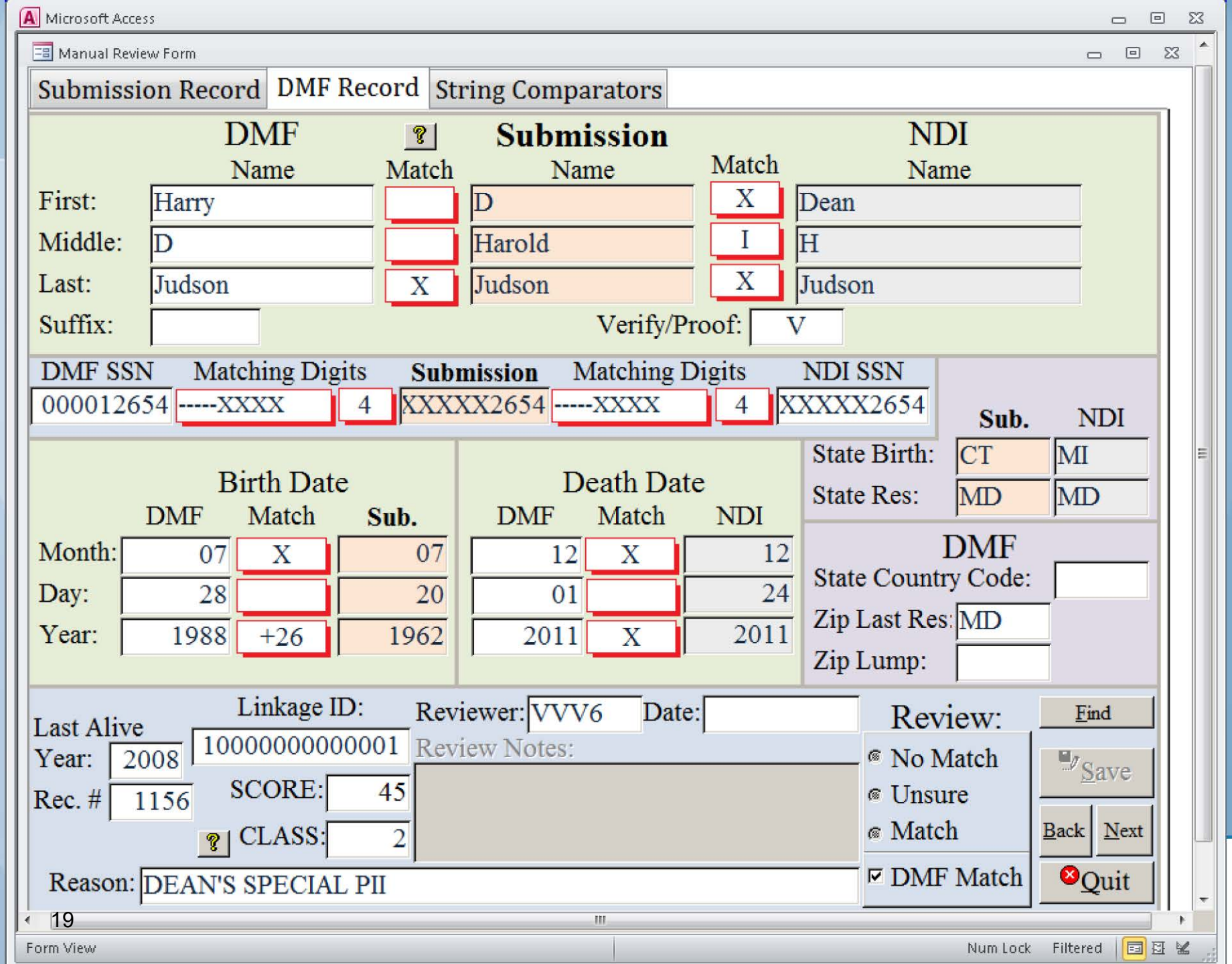
Find

Save

Back Next

Quit

Form View


Microsoft Access
Manual Review Form
Submission Record
DMF Record
String Comparators
DMF
Name
Match
Submission
Name
Match
NDI
Name
First: Harry D X Dean
Middle: D Harold I H
Last: Judson X Judson X Judson
Suffix:
Verify/Proof: V
DMF SSN
Matching Digits
Submission
Matching Digits
NDI SSN
000012654 -----XXXX 4 XXXXX2654 -----XXXX 4 XXXXX2654
Sub.
NDI
State Birth: CT MI
State Res: MD MD
Birth Date
DMF
Match
Sub.
Month: 07 X 07
Day: 28 20
Year: 1988 +26 1962
Death Date
DMF
Match
NDI
12 X 12
01 24
2011 X 2011
DMF
State Country Code:
Zip Last Res: MD
Zip Lump:
Last Alive
Year: 2008
Rec. # 1156
Linkage ID:
10000000000001
SCORE: 45
CLASS: 2
Reviewer: VVV6
Date:
Review Notes:
Review:
No Match
Unsure
Match
DMF Match
Find
Save
Back
Next
Quit
Reason: DEAN'S SPECIAL PII
19
Form View
Num Lock
Filtered

Some Operational Results

- Interrater reliability (clerical review)
- Overall link rates by race/ethnicity
- 4-digit SSN vs. 9-digit SSN match rate

Interrater Reliability Table

	NO MATCH	MATCH	TOTAL
NO MATCH	**108**	**11**	119
	49.1	5.0	54.09
	90.8	9.2	-
	83.7	12.1	-
MATCH	**18**	**79**	97
	8.2	35.9	44.09
	18.6	81.4	-
	14.0	86.8	-
TOTAL	126	90	216
	58.64	41.36	100
Cohen's Kappa:	0.727		

Disposition by Race/Ethnicity

Final disposition class, among eligibles	Race/Ethnicity						
	White-Non	Black-Non	API-Non	AIAN-Non	Hispanic	Other	Total
Assumed alive	1,306,474	306,850	66,917	13,488	313,848	80,127	2,087,704
	81.11	85.30	93.24	85.46	93.86	93.48	84.25
Dead by score/class	298,531	51,760	4,755	2,250	19,713	5,325	382,334
	18.53	14.39	6.63	14.26	5.90	6.21	15.43
Dead by rule	2,690	352	27	25	170	63	3,327
	0.17	0.10	0.04	0.16	0.05	0.07	0.13
Dead by manual review	1,209	305	26	9	229	70	1,848
	0.08	0.08	0.04	0.06	0.07	0.08	0.07
Dead by mult. sources	1,789	458	42	11	424	128	2,852
	0.11	0.13	0.06	0.07	0.13	0.15	0.12
Total	1,610,693	359,725	71,767	15,783	334,384	85,713	2,478,065
	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Nine-digit SSNs

Final disposition class, among eligibles	Full 9-digit SSN present: NO	YES	Total
Assumed alive	1,488,698	599,006	2,087,704
	94.74	66.07	84.25
Dead by score/class	77,896	304,438	382,334
	4.96	33.58	15.43
Dead by rule	1,923	1,404	3,327
	0.12	0.15	0.13
Dead by manual review	1,569	279	1,848
	0.10	0.03	0.07
Dead by mult. sources	1,315	1,537	2,852
	0.08	0.17	0.12
Total	1,571,401	906,664	2,478,065
	100.00	100.00	100.00

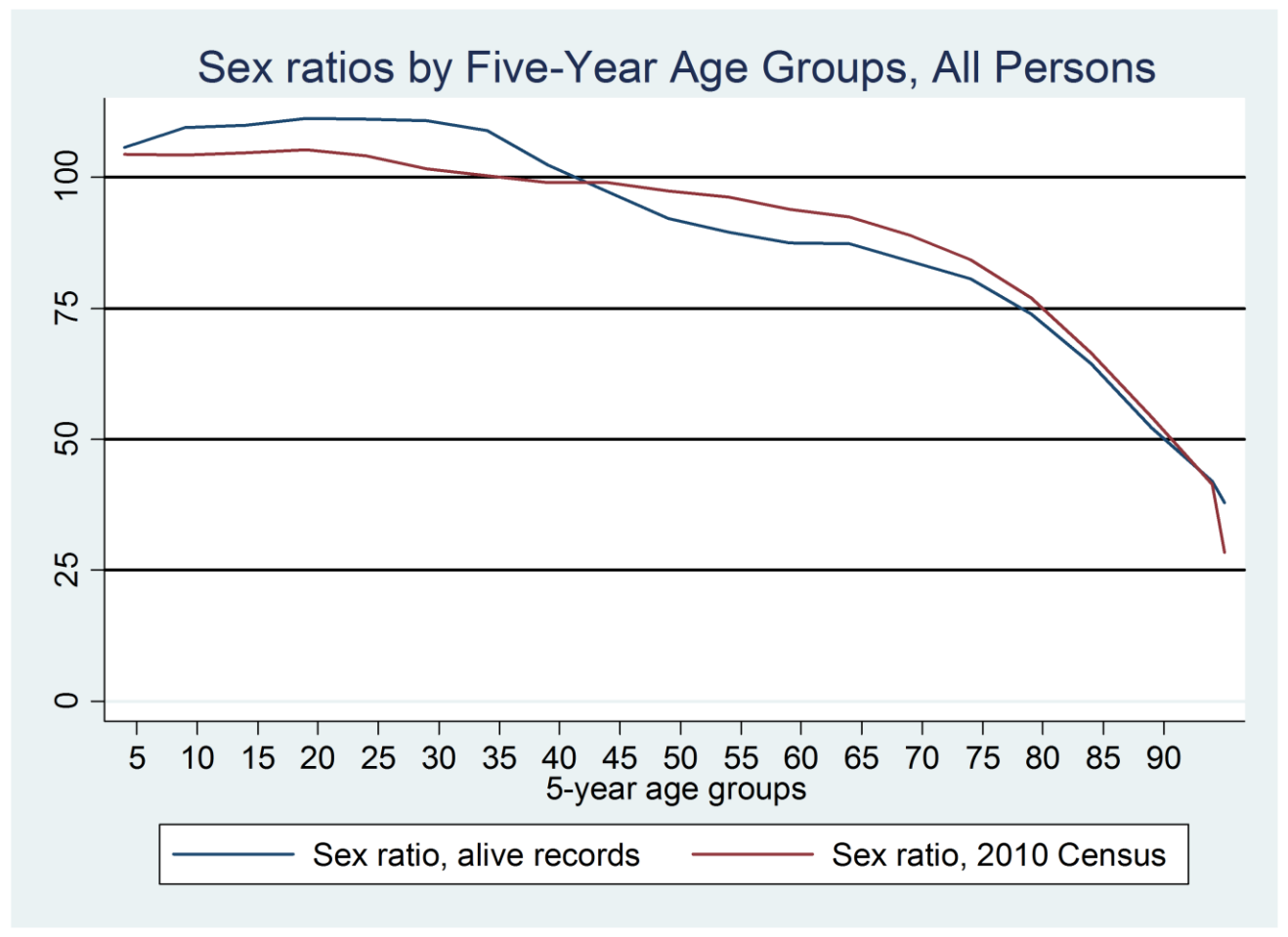
Four-digit SSNs, 2007+

Final disposition class, among eligibles	last four SSN digits only: NO	YES	Total
Assumed alive	2,064,166	23,538	2,087,704
	84.15	93.86	84.25
Dead by score/class	380,806	1,528	382,334
	15.52	6.09	15.43
Dead by rule	3,319	8	3,327
	0.14	0.03	0.13
Dead by manual review	1,843	5	1,848
	0.08	0.02	0.07
Dead by mult. sources	2,852	0	2,852
	0.12	0.00	0.12
Total	2,452,986	25,079	2,478,065
	100.00	100.00	100.00

9-digit SSNs, 2007+, NHANES

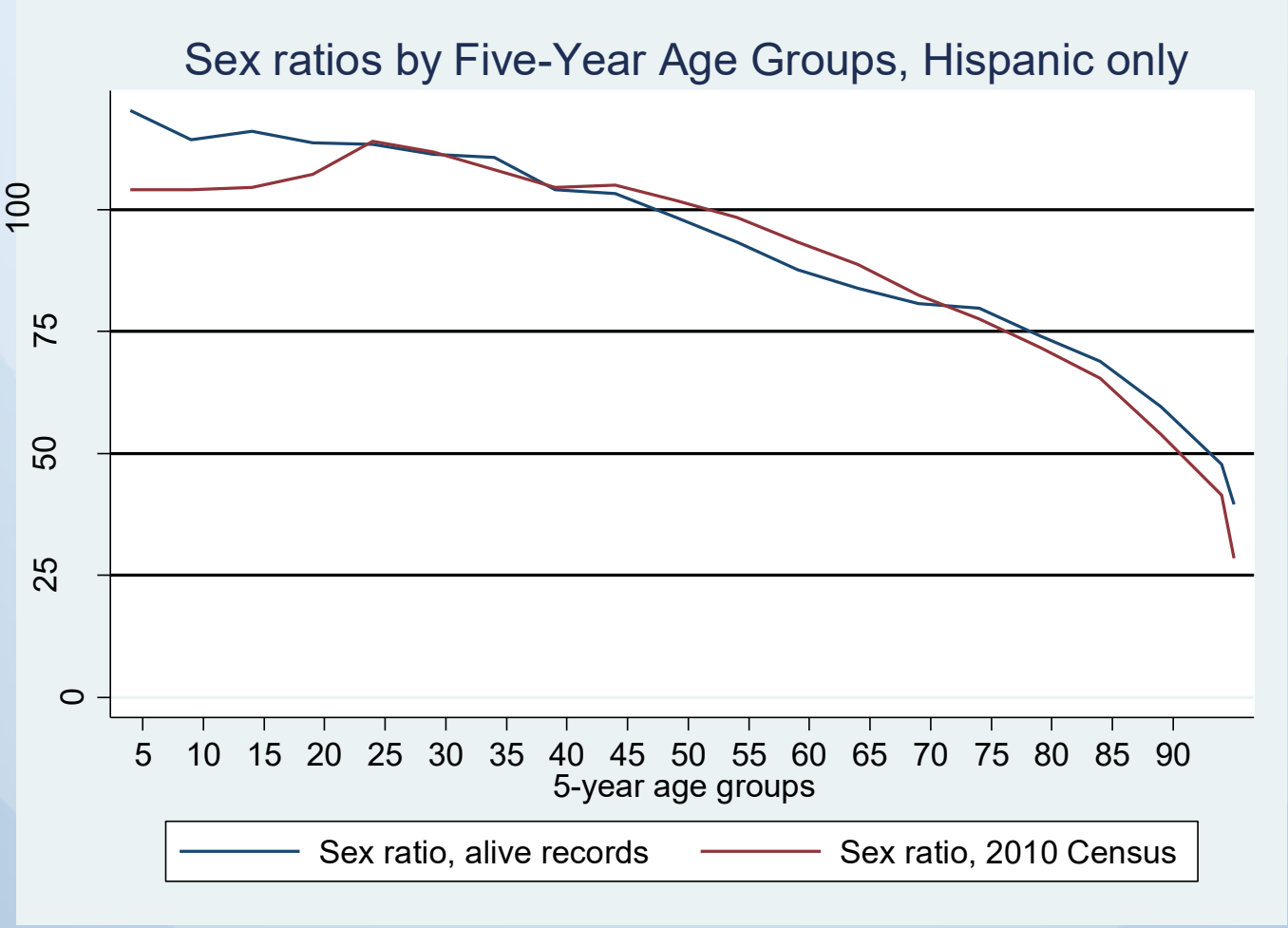
Final disposition class, among eligibles	9-digit SSN, 2007+ NO	YES	Total
Assumed alive	2,080,957	6,747	2,087,704
	84.22	94.59	84.25
Dead by score/class	381,949	385	382,334
	15.46	5.40	15.43
Dead by rule	3,326	1	3,327
	0.13	0.01	0.13
Dead by manual review	1,848	0	1,848
	0.07	0.00	0.07
Dead by mult. sources	2,852	0	2,852
	0.12	0.00	0.12
Total	2,470,932	7,133	2,478,065
	100.00	100.00	100.00

Demographic Analysis



National Center for Health Statistics
Data Linkage

Demographic Analysis, Hispanics



National Center for Health Statistics
Data Linkage

Conclusions and Future Research

- Limited success with Hispanic and Asian name algorithms
- Some success with four-digit SSN algorithm
- Good success with demographics
- More mysteries, more evaluation!
- Future: Latent class modeling, error bars, etc.