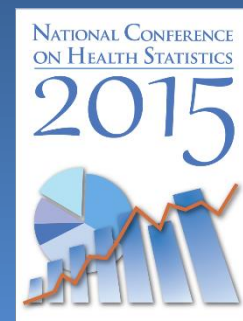


Evaluation of the 4-Digit Social Security Number Algorithm Used for the 2011 Linked Mortality Files

Frances McCarty, Ph.D.
National Center for Health Statistics



Background

Special Projects Branch (SPB) employed a matching methodology for the 2011 Linked Mortality Files similar, but not identical, to the method offered by the National Death Index (NDI)

- NCHS survey records were matched with NDI records using the following identifying information, as available:
 - Social Security Number (SSN), First name, Middle initial, Last name, Month of birth, Day of birth, Year of birth, Sex, Father's surname, State of birth, Race, State of residence, Marital status

Background: SSN

SSN is often a key identifier used in the matching process

- Increasing reluctance to provide a full 9 digit SSN
- Since 2007, National Health Interview Survey (NHIS) requests only the last 4 digits of SSN
 - Less identifying
 - Requires modified linkage method
- SPB developed a variant algorithm to determine "Class" and "Score" for respondents with only last 4 digits of SSN

Background: Score

SPB assigned a score to each potential match reflecting degree of agreement between the identifying information on the survey record and the NDI death record

- Score based on probabilistic weights assigned to each identifying data item used in the NCHS-NDI record match
- Weights could be positive, negative, or zero
- Score for SSN based on sum of the individual digit weights
- Total score for each potential match was sum of weights for each item

Background: Class

- After scoring potential matches, each was categorized into one of five mutually exclusive classes
 - Classes reflect that some items are more important for determining true matches than others (e.g. SSN vs state of birth) and that non-changing information is more important than information that can change over time (e.g. birth surname vs marital status)
- Class and score used to determine final mortality status

Class, Score, and Mortality Status

Class	Match	Score
1	At least 8 (of 9) or 4 (of 4) digits of SSN , FN, MI, LN, birth year (+/- 3 years), birth month, sex, and state of birth	All True Matches
2	At least 7 (of 9) or 4 (of 4) digits of SSN at least 5 more of the following items: LN, MI, LN, birth year (+/- 3 years), birth month, sex, and state of birth	True Match Score >=44
3	A: SSN is unknown , but LN matched and at least 7 of the following items agreed: FN, MI, LN, birth year (+/- 3 years), birth day, sex, race, marital status and state of birth. B: SSN was known but 3 or more (of 9) and 1 or more (of 4) digits did not agree , but at least 8 of the following items agreed: FN, MI, LN, birth year, birth day, sex, race, marital status, and state of birth. Switched from Class 5 to Class 3 - SSN was recorded incorrectly or spouse's SSN was recorded. Scores adjusted to reflect that SSN was missing (assigned value of 0).	True Match Score >=45
4	SSN was unknown on either the NCHS survey submission record or the NDI record and fewer than 8 of the items listed in Class 3 matched	True Match Score >=42
5	SSN was present but fewer than 7 (of 9) or 4 (of 4) digits on SSN agreed	None True Matches

Evaluation of 4-Digit SSN Processing

Compare Class, Score, and Final Status when 4 digit algorithm is used instead of the 9 digit algorithm

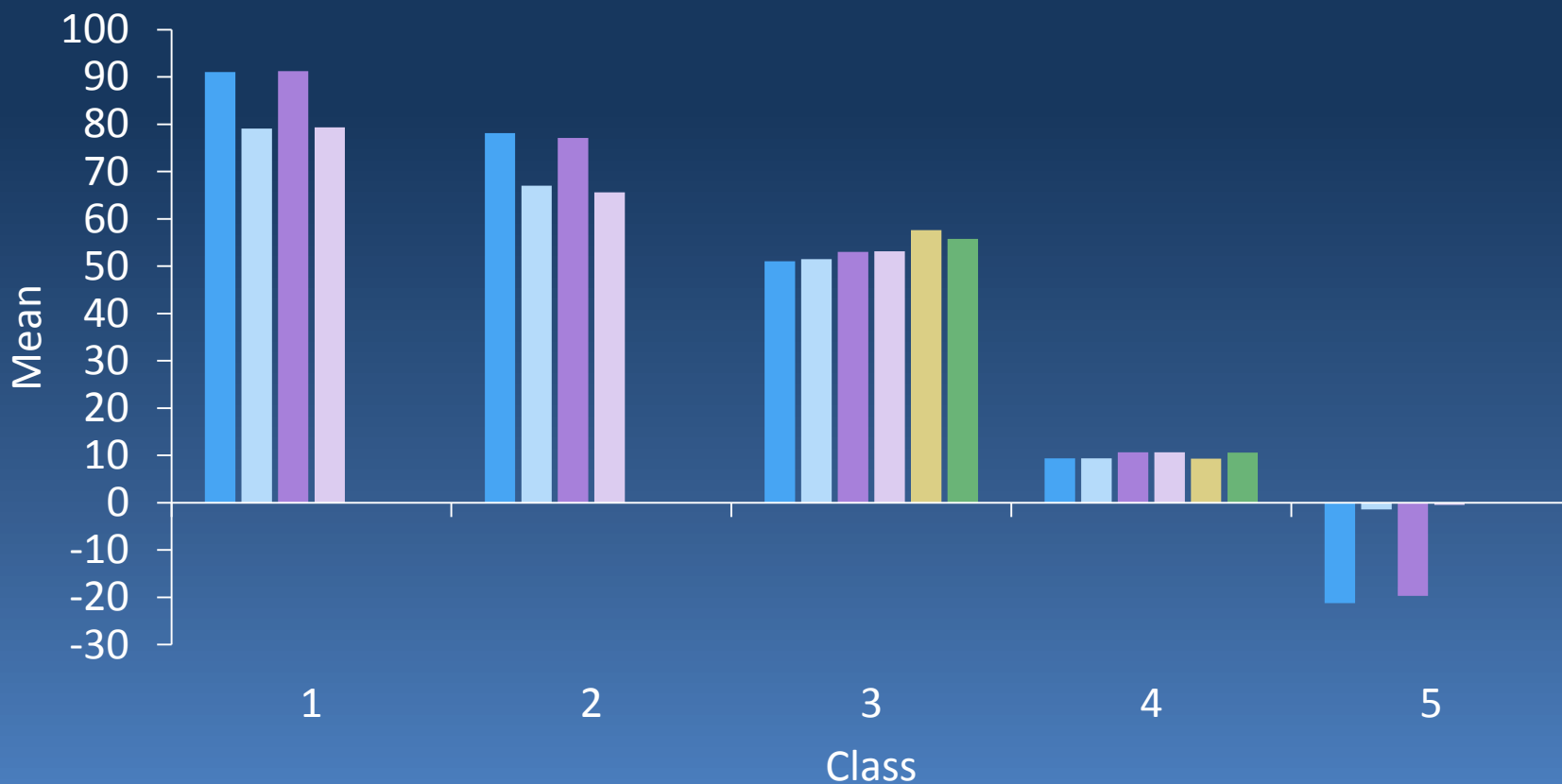
Methods

- Data from NHIS 1999-2006 and NHANES 1999-2010
 - All 9 digits were requested
- Censored first 5 digits and recalculated class and score as though only the last 4 digits were collected
- Censored all 9 digits and recalculated class and score
- 398,518 records for NHIS and 37,864 records for NHANES (Total n=436, 382)
 - 106,286 records for NHIS and 23,187 records for NHANES (total n=129,473) had SSN (29.7%)
- Agreement assessed using percent agreement and Kappa statistics

Total Score

4-Digit vs 9-Digit SSN

■ NHANES SSN-9 ■ NHANES SSN-4 ■ NHIS SSN-9
■ NHIS SSN-4 ■ NHANES SSN-0 ■ NHIS SSN-0



Class Agreement: NHANES and NHIS

Class SSN4	Class SSN 9					Total	% in class
	1	2	3	4	5		
1	17003	72	10	0	1	17086	3.92
2	0	5238	4	0	7	5249	1.2
3	262	200	29991	0	0	30453	6.98
4	0	0	1	278155	0	278156	63.74
5	16	167	0	0	105255	105438	24.16
Total	17281	5677	30006	278155	105263	436382	
% in class	3.96	1.3	6.88	63.74	24.12		

Agreement on Class: 99.8% , 435642 out of 436382

Kappa = .9956 (95% CI: .9953, .9960)

Final Status Agreement: NHANES and NHIS

	SSN 9			
SSN 4	Alive	Dead	Total	Status%
Alive	388,529	383	388,912	89.12
Dead	3	47,467	3,761	10.88
Total	388,532	47,850	436,382	
Status%	89.03	10.97		

Agreement on Status: 99.9% , 435996 out of 436382

Kappa = .9955 (95% CI: .9950, .9959)

Class Agreement: NHANES and NHIS Only Records with SSN Present

	Class SSN 9					
Class SSN4	1	2	3	5	Total	% in class
1	17003	72	10	1	17,086	13.20
2	0	5238	4	7	5,249	4.05
3	262	200	1238	0	1,700	1.31
5	16	167	0	105255	105,438	81.44
Total	17,281	5,677	1,252	105,263	129,473	
% in class	13.35	4.38	0.97	81.30		

Agreement on Class: 99.4% , 128734 out of 129473

Kappa = .9903 (95% CI: .9896, .9911)

Final Status Agreement: NHANES and NHIS Only Records with SSN Present

	SSN 9			
SSN 4	Alive	Dead	Total	Status%
Alive	106,010	383	106,393	82.17
Dead	3	23,077	3,761	17.83
Total	106,013	23,460	129,473	
Status%	81.88	18.12		

Agreement on Status: 99.7% , 129087 out of 129,473

Kappa = .9899 (95% CI: .9889, .9909)

Number of Deaths using 9 digit SSN, 4 digit SSN and No SSN

- Consider deaths identified by 9 digit SSN as true matches
- Number of false positives and false negatives is considerably smaller for 4 digit SSN algorithm compared to no SSN [numbers comparing no SSN to 4 digit SSN]

	All Records			Only Records with SSN		
	SSN 9	SSN 4	No SSN	SSN 9	SSN 4	No SSN
Number of Deaths	47,850	47,470	46,550	23,460	23,080	22,601
False Positive (Dead by Alternative)		3	629 [693]		3	629 [693]
False Negative (Alive by Alternative)		383	1929 [1613]		383	1488 [1172]

Summary

4 digit SSN algorithm performed well compared to the 9 digit algorithm

- Overall agreement on final status was high
- In general, agreement was not different based on survey, respondent sex, or race/ethnicity
- A relatively small number of deaths (383) identified using 9-digit SSN were not identified using 4-digit algorithm

Compared to status determined without SSN, the 4 digit algorithm resulted in fewer false negatives and false positives

Conclusions

The 4 digit SSN algorithm provided results very similar to those obtained using the 9 digit algorithm

The additional matching information provided by the last 4 SSN digits could be beneficial in determining true matches

Thanks!

Dean Judson

Eric Miller

Jim Brittain

Keith Zevallos

Jesse Bassich

Cordell Golden

Eileen Call

Deborah Ingram

Jennifer Parker

Contact Info:

Frances McCarty

FMcCarty@cdc.gov

Percent of Records with SSN Present by Survey, Sex, and Race/Ethnicity

