VA-HRS Data Linkage Project

Record Matching Report

Author:
Elizabeth Tarlov, PhD, RN for the VA Information Resource Center
Elizabeth.Tarlov@va.gov

January 30, 2016

Introduction

The VA-HRS linked data files comprise the VA healthcare data of HRS respondents who self-identified as having served in the U.S. military. Two record matches, primarily using probabilistic methods, were conducted to identify their VA records. One match procedure identified the VA records of deceased HRS respondents and the other identified the VA records of living HRS respondents who had provided a signed authorization. Separate procedures were conducted to allow the use of year and month of death as additional matching data elements for the deceased group. A matching methodology similar, though not identical, to that employed by the National Center for Health Statistics in constructing the NHIS Linked Mortality Files and the National Death Index, was used.[1] Link Plus version 2.0 probabilistic data linkage software, developed by the CDC for use in its National Program of Cancer Registries, was used to identify and weight potential matches.[2,3] Using SAS statistical software and manual review, potential matches were classified based on the specific matching identifiers and decision rules were then developed for final classification of each potential match as true or false. VA data on enrollment, utilization, and costs from 1999-2013 for the matched individuals was identified and data files were constructed following removal of information protected by Title 38 U.S.C. § 7332, *Confidentiality of certain medical records*.[*]

Matching Algorithm

The following identifying information from HRS records and VA administrative records was used in the record linkage:

Social Security Number
First name
Middle name
Last name
Birth year
Birth month
Birth day
Gender
Death year
Death month

Table 1 lists the specific data sources and data elements used in the linkage from HRS and VA sources. To be eligible for the linkage, a record had to include either a 9-digit Social Security Number (SSN) or a last name; records from either HRS or VA without one of these data elements was ineligible for linkage. In addition, VA records had to include a year of birth. Veteran records with a birth year before 1914 and after 1985 were excluded from the linkage procedure. Table 2 shows availability of specific data elements in HRS data.

In a first step, Link Plus selected all HRS and VA records that matched on either SSN or last name (the "blocking" variables). Only record pairs satisfying that criterion were examined for agreement on

[*] http://uscode.house.gov/view.xhtml?path=/prelim@title38/part5/chapter73&edition=prelim. Accessed 1/7/2013.

other variables. Agreement on names was based on exact spelling matches or on the way a name sounds rather than how it is spelled, using the New York State Identification Intelligence System (NYSIIS),[4] which converts a name to a phonetic coding. Since one or more VA records could be matched to a given HRS record, the record selection process could return several potential matches for each HRS respondent, many of which will be non-matches or duplicate records.

Scoring and Classifying Potential Match Records

Link Plus assigned a score to each potential match reflecting the degree of agreement between the identifying information on the HRS record and the VA record. The score is based upon probabilistic weights assigned to each of the identifying data items used in the record match.[5] For example, a common first name, such as “John”, that has a higher probability of occurrence in the population has a lower weight than an uncommon name such as “Bartholomew”. Name weights assigned by Link Plus are based on the frequency of names in the National Death Index. The score for each potential match is the sum of the weights for each individual data item. VA-HRS record pairs with a score of 7 or higher were output as potential matches (based on the recommendation provided in the Link Plus guidance).

Using SAS, each potential match was categorized into one of six mutually exclusive classes. This class categorization takes into account which identifying items agree.[1] In this way, the match acknowledges the greater importance of some items compared to others (e.g., SSN compared to first name) for determining true matches. In additon, two additional identifying data elements, not used in the probabilistic procedure, were taken into consideration at this stage. For each potential match, the middle initial (HRS and VA) or name (VA) were flagged as: missing in one or both sources; present in both sources and agree; or present in both sources and disagree. In addition, vital status was flagged as in agreement or not based on the presence or absence of a date of death in HRS and VA data.

As SSN is a key identifier in the matching process, each VA-HRS record match was initially classified according to whether SSN was present and agreed (Class 1 or 2), was present but disagreed (Class 3b or 5) or was missing (Class 3a or 4). Tables 3a and 3b show, for the deceased and living groups, respectively, definitions for the six classes used to group potential matches.

Selecting Matches

For those HRS records for which Link Plus returned more than one potential VA match, a single best record was selected using a two-step strategy involving ranking by class and score. First, if the potential matches for an HRS record were categorized in more than one class, only potential matches in the higher class (where 1 is highest, 5 is lowest) were considered; potential matches in lower classes were dropped from consideration. Second, when there was more than one potential match in the higher class, the matched record pair with the highest score was selected as the single best match; other potential matches were dropped from further consideration.

Finally, each best record match was determined to be "true" or "false" based on the following criteria.

1. All Class 1 matches were considered true matches.
2. All Class 5 matches were considered false matches.
3. For Classes 2, 3, and 4, matches were sorted according to the specific identifying information on which they agreed and the minimum, median, and maximum match score

in each group were computed. After a detailed examination of the relationship of the summary data to the match patterns, a cut-off score was determined within each class. Scores above the cut-off were considered true matches and scores below the cut-off were considered false matches. The cut-off score was based on manual review and best judgment with the goal of maximizing the proportion of matches correctly classified and minimizing the proportion incorrectly classified.

Tables 4a and 4b show match score statistics by class for potential matches and for true matches in the deceased and living groups, respectively.

References

1. National Center for Health Statistics, Office of Analysis and Epidemiology. *The National Health Interview Survey (1986-2004) Linked Mortality Files, mortality follow-up through 2006: Matching Methodology*. Hyattsville, Maryland: Centers for Disease Control and Prevention; 2009 May. (Available at: http://www.cdc.gov/nchs/data/datalinkage/matching_methodology_nhis_final.pdf)
2. *Link Plus* [computer program]. Version 2.0. Atlanta, GA: Centers for Disease Control and Prevention; 2008. http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm
3. Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. *Health Informatics Journal*. 2008;14(1): 5-15. doi:10.1177/1460458208088855
4. Lynch B, Arends W. *Selection of a Surname Encoding Procedure for the Statistical Reporting Service Record Linkage System*. Washington, DC: United States Department of Agriculture; 1977.
5. Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association*. 1969;64:1183–1210

Table 1. HRS and VA Data Sources and Data Elements Used in Linkage

	HRS Data		VA Data		
	Source	**Variable Name**	**Source**	**Variable Name**	**Original data elements**
Social Security Number	HRS Field Data	SSN_ST	CDW SPatient.SPatient Table	REALSSN	REALSSN
	SSA	SSN_PERM			
	CMS	SSN_CMS			
	SS Death Master File	SSN_DMF			
Last Name		LAST	CDW SPatient.SPatient Table	VA_LNAME	PatientLastName
First Name		FIRST	CDW SPatient.SPatient Table	VA_FNAME	PatientFirstName
Middle Name		MIDDLE	CDW SPatient.SPatient Table	VA_MNAME	PatientFirstName
Birthdate		DOB	CDW SPatient.SPatient Table	VA_DOB	DateofBirth
Gender		GENDER	CDW SPatient.SPatient Table	VA_SEX	Gender
Death Year*	HRS Field Data	EX_YOD	VHA Vital Status File	CMS_DOD SSA_DOD BRL_DOD ENR_DOD FEE_DOD PTF_DOD	
	National Death Index	NDI_YOD	CDW SPatient.SPatient Table	VA_DOD	DateofDeath
	SS Death Master File	DMF_YOD			
Death Month*	HRS Field Data	EX_MOD	VHA Vital Status File	CMS_DOD SSA_DOD BRL_DOD ENR_DOD FEE_DOD PTF_DOD	
	National Death Index	NDI_MOD	CDW SPatient.SPatient Table	VA_DOD	DateofDeath
	SS Death Master File	DMF_MOD			

*Used in linkage of deceased respondents only

Table 2. HRS Data Element Availability

	Deceased Respondent Linkage (N= 4,055 HRS Subjects)			Living Respondent Linkage (N = 1,875 HRS Subjects)		
	Populated N	**Missing N**	**Missing %**	**Populated N**	**Missing N**	**Missing %**
Social Security Number*						
Any source	3,946	109	2.7	1,666	209	11.1
HRS Field Data	3,589	466	11.5	1,546	329	17.5
SSA	3,688	367	9.1	1,616	259	13.8
CMS	3,006	1,049	25.9	1,153	722	38.5
SSA Death Master File	3,106	949	23.4	NA	NA	
Last Name	4,055	0	0.0	1,875	0	0.0
First Name	4,055	0	0.0	1,875	0	0.0
Middle Name						
Birthdate						
Year	4,054	1	0.02	1,875	0	0.0
Month	4,047	8	0.20	1,875	0	0.0
Day	4,032	23	0.57	1,875	0	0.0
Gender	4,055	0	0.0	1,875	0	0.0
Death Year**						
Any source	4,041	14	0.35	NA	NA	NA
HRS Field Data	4,055	0	0.0	NA	NA	NA
NDI	2,782	1,273	31.4	NA	NA	NA
SSA Death Master File	3,105	950	23.4	NA	NA	NA
Death Month**						
Any source	4,041	14	0.35	NA	NA	NA
HRS Field Data	4,055	0	0.0	NA	NA	NA
NDI	2,782	1,273	31.4	NA	NA	NA
SSA Death Master File	3,105	950	23.4	NA	NA	NA
Death Day**						
Any source	3,503	552	13.6	NA	NA	NA
HRS Field Data	3,503	552	13.6	NA	NA	NA
NDI	NA	NA	NA	NA	NA	NA
SSA Death Master File	NA	NA	NA	NA	NA	NA

* 42 living respondents had more than one SSN.

** Used in linkage of living respondents only; NDI and SSA DMF data not available for 771 recently deceased respondents; death day not used in match since day was not available in NDI or DMF data.

Table 3a. Class Definitions for Deceased Cohort

Class	Agreement on Matching Elements*										
	SSN Digits	**Last Name**	**First Name**	**DOB Year (+/- 3 Year)**	**DOB Month**	**DOB Day**	**MI**	**DOD Year**	**DOD Month**	**Sex**	**Vital Status**	**Other**
Class 1	9	1	1									
	9	0	1	1	1	1	≠0				≠0	
	9	0	1				≠0	1	1			
	9	1	0	1	1	1	≠0				≠0	
	9	1	0				≠0	1	1			
	7 or 8	1	1	1	1	1	≠0				≠0	
	7 or 8	1	1	1			≠0	1	1			
Class 2	>=7											Does not meet criteria for Class 1
Class 3a	SSN Missing	1	1'	1'	1'	1'	1'	1'	1'	1'	≠0	At least 7 of 8 elements must agree
Class 3b	<7	1	1	1	1	1						*OR*
		1'	1'	1'	1'	1'	1'	1'	1'	1'		At least 8 of 9 elements must agree
Class 4	SSN Missing											Does not meet criteria for Class 3A
Class 5	<7											Does not meet criteria for Class 3B

*SSN and last name (including NYSIIS phonetic coding match) were blocking variables. Therefore, all possible matches were, at a minimum, matched on either SSN or last name.

1 = Agree 0 = Disagree NA = Information not available +/-3 = Birth year agrees within 1 year 1' = Agree *or* (a prescribed number of these elements must agree, see Other column)

Abbreviations: SSN Social Security Number; DOB date of birth; MI middle initial; DOD date of death

Table 3b. Class Definitions for Living Cohort

Class	Agreement on Matching Elements*									
	SSN Digits	Last Name	First Name	DOB Year (+/-3 Year)	DOB Month	DOB Day	MI	Sex	Vital Status	Other
Class 1	9	1	1							
	7	0	1	1	1	1	≠0		≠0	
	7	1	0	1	1	1	≠0		≠0	
	7 or 8	1	1	1	1	1	≠0		≠0	
Class 2	>=7									Does not meet criteria for Class 1
Class 3a	SSN Missing	1	1'	1'	1'	1'	1'	1'	≠0	At least 5 of 6 elements must agree
Class 3b	<7	1	1	1'	1'	1'	1'	1'	≠0	At least 4 of 5 elements must agree
Class 4	SSN Missing									Does not meet criteria for Class 3A
Class 5	<7									Does not meet criteria for Class 3B

*SSN and last name (including NYSIIS phonetic coding match) were blocking variables. Therefore, all possible matches were, at a minimum, matched on either SSN or last name.

1 = Agree 0 = Disagree NA = Information not available +/-3 = Birth year agrees within 1 year 1' = Agree *or* (a prescribed number of these elements must agree, see Other column)

Abbreviations: SSN Social Security Number; DOB date of birth; MI middle initial; DOD date of death

Table 4a. Match Score Statistics by Class of 3,360 Potential Matched Record Pairs: Deceased Cohort

	Potential Matched Record Pairs					"True" Matches*	
	N	Min	Max	Mean (SD)	Median	N	Mean (SD)
Class I	1,404	15.6	47.2	37.1 (4.3)	37.9	1,404	37.1 (4.3)
Class 2	55	8.1	34.9	14.8 (6.8)	12.3	9	18.2 (11.8)
Class 3a	11	14.8	30.7	25.0 (4.3)	24.9	10	26.1 (2.7)
Class 3b	60	7.6	25.6	12.3 (4.0)	11.2	9	19.6 (3.1)
Class 4	82	7.5	25.1	12.9 (4.0)	12.2	13	20.5 (2.9)
Class 5	1,736	7.0	17.7	10.1 (1.6)	10.0	Not applicable	

*True matches: N 1,445 Mean (SD) 36.8 (4.9); By definition, Class 5 record pairs are false matches.

Table 4b. Match Score Statistics by Class of 1,655 Potential Matched Record Pairs: Living Cohort

	Potential Matched Record Pairs					"True" Matches*	
	N	Min	Max	Mean (SD)	Median	N	Mean (SD)
Class I	793	10.9	42.0	34.1 (3.2)	34.5	793	34.1 (3.2)
Class 2	17	7.8	24.6	14.0 (5.2)	13.8	3	22.7 (1.7)
Class 3a	114	11.3	28.5	21.2 (3.4)	21.5	110	21.5 (3.0)
Class 3b	45	7.3	26.1	12.4 (4.7)	11.1	10	19.8 (3.9)
Class 4	81	7.8	21.0	11.8 (2.3)	11.9	Not applicable (no true matches)	
Class 5	605	7.0	15.2	10.1 (1.5)	10.1	Not applicable	

*True matches: N 916, Mean (SD) 32.4 (5.4); By definition, Class 5 record pairs are false matches.