THE USE OF PROBABILISTIC METHODS IN MATCHING CENSUS SAMPLES TO THE NATIONAL DEATH INDEX

Eugene Rogot, National Heart, Lung, and Blood Institute
Sidney H. Schwartz, U.S. Bureau of the Census
Karen V. O'Conor, U.S. Bureau of Labor Statistics
Christina L. Olsen, U.S. Bureau of the Census

In a pilot project, Census Bureau files containing 226,000 person records were matched to the 1979 National Death Index (NDI) (1). There were 5542 "possible matches" or "hits" of which about a third are true positives, representing deaths of Current Population Survey (CPS) sample persons in 1979, and about two-thirds are "false positives". A probabilistic method adapted for computer use is now being developed to separate true positives from false positives. Such a method could save much time in manual review of records and is deemed virtually essential with very large matching studies. (One now underway, for example, involved the matching of one million CPS records to the NDI resulting in more than 50,000 hits. More than 35,000 of these hits are expected to be false positives).

MATERIALS AND METHODS

The NDI matching items, matching algorithm, and retrieval report have been described by Bilgrad (2) and are set forth in the NDI User's Manual (3). As indicated, there are seven ways in which a hit may occur with the NDI algorithm but these are not mutually exclusive. For our purposes, hits were classified into mutually exclusive types according to agreement or disagreement on five key items - Social Security Number (SSN), last name (LN), first name (FN), year of birth (YB), and month of birth (MB). A finer break down of hits into sub-types according to whether agreement on names was exact or Soundex agreement only was also found useful. (Soundex is a method of coding names to take account of common spelling differences. The Soundex codes consist of the first letter of the name followed by three or more numerical codes corresponding to the consonant sounds represented in the name.)

Table 1 shows the distribution of hits by type and sub-type collapsed into seven mutually exclusive categories. Also shown in Table 1 is a breakdown of hits by category and final judgments of true positive and false positive for deaths occurring in five states that were selected for intensive review. The five states were: Arizona, Florida, Illinois, Maine and Washington.

The determination of true or false positive was based on a consensus of three raters using all available information including death certificates and CPS control cards (1).

As seen in Table 1, about a fifth of the hits were Category A--all five key identifiers agree. In the sample, all 141 hits in this category were true positives. Hits in this category are virtually certain to be true positives and therefore have been excluded from any probabilistic scheme.

About 8% of the hits fall into Category B--SSN agrees but one or more key identifiers disagree. In the sample all but four of the 54 hits in this category were true positives. In all four instances sex disagreed. Based on these results, judging whether a hit is a true or false positive for Category B can be accomplished simply by checking for agreement or disagreement on sex. This category was also excluded from the probabilistic scheme. The remaining categories, C-G, constituting about three-fourths of the hits, were studied with probabilistic methods.

Table 2 gives the frequencies of true and false positives and overall odds of true positives to false positives for hit categories C through G. Data here are based on the sample of five states mentioned earlier. The odds vary from (1:.9) for Category E to (1:110) for Category G. Also shown in Table 2 are binit weights (labelled W_T) which range from -6.8 for Category G to +.2 for Category E. A binit weight is defined as the logarithm to the base two of the odds.

In the probabilistic method espoused by Newcombe and others (4-9), a weight (W_i), which is positive for agreement and negative for disagreement, is assigned to every matching item. The weight for agreement is defined as the $\log_2$ of the ratio of the percentage agreement among good links to percentage agreement among non-links, and the disagreement weight is the $\log_2$ of the ratio of the percentage disagreement among good links to percentage disagreement among non-links. In this context, a "good link" refers to a pair of records, one from each of two different files, which are known to represent the same person, i.e., a true positive. A "non-link" is a pair of records, one from each of two different files, which do not represent the same person i.e., a false positive. In the table below the binit for agreement of item X is $\log_2 \left(\frac{a}{a+c} \div \frac{b}{b+d} \right)$ and the binit for disagreement of item X is $\log_2 \left(\frac{c}{a+c} \div \frac{d}{b+d} \right)$.

Table 1. Frequency of hits by category for complete file and final judgments of true+ and false+ for hits observed in a sample of 5 states: Census-NDI Pilot Study

Hit category	SSN	LN	FN	YB	MB	No. of hits	No. observed in 5-state sample		
							True+	False+	Total
A	+	(+ or S)	(+ or S)	+	+	1056	141	0	141
B	+	various combinations				463	50	4	54
C	-	+	+	+	+	774	13	89	102
D	-	S	+	+	+	868	3	113	116
	-	+	S	+	+	362	1	48	49
	-	S	S	+	+	430	0	58	58
E	?	+	+	+	+	323	16	14	30
F	?	S	+	+	+	196	2	32	34
	?	+	S	+	+	110	2	12	14
	?	S	S	+	+	137	1	16	17
G	Other hits[1]					823	1	110	111
Totals						5542	230	496	726

[1] Includes 79 hits based on father's surname agreement and 744 hits (among females) in which last name on the CPS record agreed with the decedent's father's last name on the NDI record. None of these 823 hits showed agreement on last name.

Table 2. Frequency of true positives and false positives, with overall odds and binit weights, for selected hits[1] from a sample of 5 states: Census-NDI Pilot Study

Hit category	No. of hits		Overall odds (true+:false+)	Binit Weight (W_T) ($\log_2$ (odds))	SSN	LN	FN	YB	MB
	True +	False +							
C	13	89	1:6.8	-2.8	-	+	+	+	+
D	4	219	1:54.8	-5.8	-	S on one or both		+	+
E	16	14	1:.9	+.2	?	+	+	+	+
F	5	60	1:12.0	-3.6	?	S on one or both		+	+
G C'	0	30	1:110.0	-6.8	-	+	+	+	+
G D'	0	62			-	S on one or both		+	+
G E'	0	2			?	+	+	+	+
G F'	1	16			?	S on one or both		+	+

[1] For type G, LN means father's surname comparison or, among females, a cross-match in which the surname on the CPS record is compared with the decedent's father's surname on the NDI record. Types C', D', E' and F' correspond to types C, D, E and F if "last name" is replaced with "father's surname" or with the "cross-match."

SSN =	Social Security Number	+ =	Exact agreement
LN =	Last name	S =	Soundex agreement (on names)
FN =	First name	- =	Disagrees
YB =	Year of birth	? =	Not available on CPS file or insufficient data on NDI file
MB =	Month of birth		

Agreement	Among good links	Among non-links
Agrees	a	b
Disagrees	c	d
	a+c	b+d

For each hit an overall score, W, is obtained by algebraically summing these weights over all matching items and then adding a weight for the probability of death (for a person of that age and sex) and a weight for the size of the death file. W may be interpreted as directly reflecting the absolute odds of the particular hit being a true positive to its being a false positive assuming that the W_i are independent.

In the present study this approach has been modified by

(1) treating hits separately by category of hit as shown in Table 2,

(2) distinguishing between key identifiers--those that affect hit status--and other identifiers that do not, and,

(3) replacing the weights for death and size of file with the weight for the hit type (W_T).

The formula used is:

$$W=W_T+W_{MI}+W_{DB}+W_S+W_R+W_{MS}+W_{SR}+W_{SB}+W_{SLN}+W_{SFN}+W_{SYB}$$

where

W_T = weight for the type of hit

W_{MI} = weight for agreement or disagreement on middle initial

W_{DB} = weight for agreement or disagreement on day of birth

W_S = weight for agreement or disagreement on sex (W_S = 0 for hits in which first name agrees exactly)

W_R = weight for agreement or disagreement on race

W_{MS} = weight for agreement or disagreement on marital status

W_{SR} = weight for agreement or disagreement on state of residence

W_{SB} = weight for agreement or disagreement on state of birth

W_{SLN} = weight for agreement on specific last name (W_{SLN} = 0 for hits with Soundex agreement only on last name, and category G)

W_{SFN} = weight for agreement on specific first name (W_{SFN} = 0 for hits with Soundex agreement only on first name)

W_{SYB} = weight for agreement on specific year of birth

This score, W, is calculated for each hit except those in which the Social Security Number agreed. A positive score indicates that the hit is more likely to be a true positive than a false positive and a negative score indicates the reverse. The score, W, is in binits so that the odds of a true positive to a false positive are 2^W:1. For example, if W=10, the odds are 1024:1 that the hit is a true positive rather than a false positive. In this report, we have chosen +10 and -10 as cut off points. Hits in which $W \geq 10$ are taken as true positives, the odds being 1024:1 or greater that the hit is a true positive, and hits in which $W \leq -10$ are taken as false positives, the odds being 1:1024 or smaller that the hit is a true positive. The remaining hits, in which W lies in the interval -9 through +9, may be considered questionables requiring additional information before a decision is reached whether they are true positives or false positives.

The agreement and disagreement weights used for the non-key identifiers (MI,

Table 3. Counts of agreements and disagreements for selected identifiers by final judgments of true+ and false+ and binit weights for hits categorized as C, D, E and F combined [1] in a sample of 5 states: Census-NDI Pilot Study

Identifier	Agrees		Disagrees		Unknown		Binit Weights for:	
	True +	False +	True +	False +	True +	False +	Agreement	Disagreement
MI	12	18	2	148	24	216	+3.0	-2.6
DB	37	12	-	369	1	1	+5.0	-6.2
S[2]	9	216	-	61	-	2	+.3	-2.1
R	37	298	-	79	1	5	+.3	-4.0
MS	27	178	3	148	8	56	+.7	-2.2
SR	29	11	9	371	-	-	+4.7	-2.0
SB	9	-	1	21	28	361	+5.3	-3.3

Note: When calculating weights, a zero cell frequency was replaced by 1/2.

[1] All straight matches except those in which the Social Security Number agreed.

[2] Based on hits in categories D and F only.

MI = Middle initial S = Sex MS = Marital status
DB = Day of birth R = Race SR = State of residence SB = State of birth

DB, S, R, MS, SR and SB) are based on counts made from the five sample states and are shown in Table 3.

The weights used for agreement on specific last names, W_{SLN}, and specific first names, W_{SFN}, are based on relative frequencies of names in a large Census file--the CPS for April 1980 (see Appendix example). Weights for agreement on specific years of birth, W_{SYB}, are based on relative frequencies of individual years of birth in the 1979 NDI file.

RESULTS

For the original sample of 5 states, the observed frequency distribution for W is shown in Figure 1, separately for straight matches and indirect matches.

For the indirect matches, hit category G, W ranged from -25 to +2 for the 110 false positive hits with a skip to +6 for the single true positive hit. For the straight matches, W varied from -23 to +7 for 382 false positives and from +2 to +29 for 38 true positives. The overlap area, from +2 to +7, included 3 true positives and 5 false positives.

So far the probabilistic method has been developed and tested only on the first sample of five states. Our aim is to validate the method on the entire file of hits. Death certificates are being collected for hits from the remaining states so that final judgments of true positive or false positive can be made for all 5542 hits. At this time, a second sample consisting of 723 hits from Alabama, Alaska, Delaware, District of Columbia, Georgia and Pennsylvania was available for study.

A comparison of this second sample of 723 hits with the first sample of 726 hits, by counts in each hit category according to final determination of true positive and false positive as judged by manual review of death certificates, is shown below.

	First Sample		Second Sample	
	True+	False+	True+	False+
A	141	--	109	--
B	50	4	54	6
C	13	89	10	107
D	4	219	3	210
E	16	14	25	21
F	5	60	5	66
G	1	110	--	107
	230	496	206	517

Although the distributions differ somewhat by category of hit, there is a striking similarity in the ratios of true to false positives by type of hit. Thus, the weights for type of hit, W_T, are virtually the same for the two samples.

Figure 1.
Frequency Distribution of Overall Binit Scores, W, for 531 Hits from First Sample of 5 States
(Hits Include Categories C,D,E,F & G)

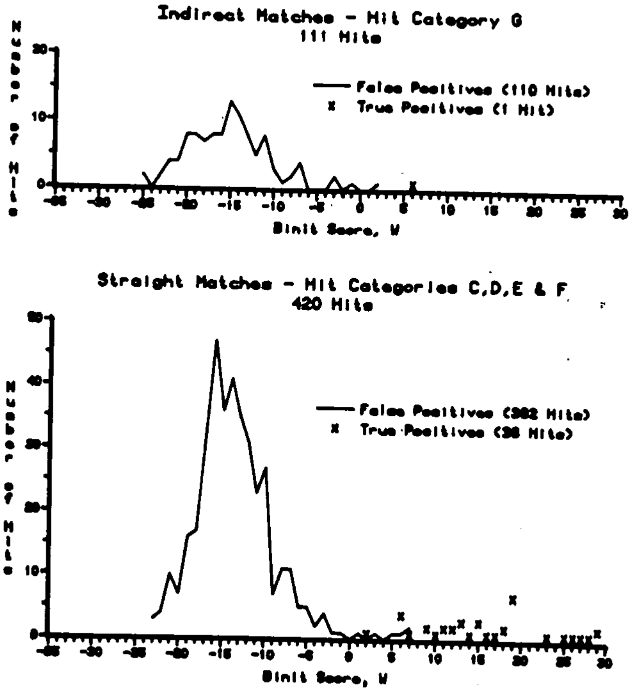


Figure 2.
Frequency Distribution of Overall Binit Scores, W, for 553 Hits from Second Sample.
(Hits Include Categories C,D,E,F & G)

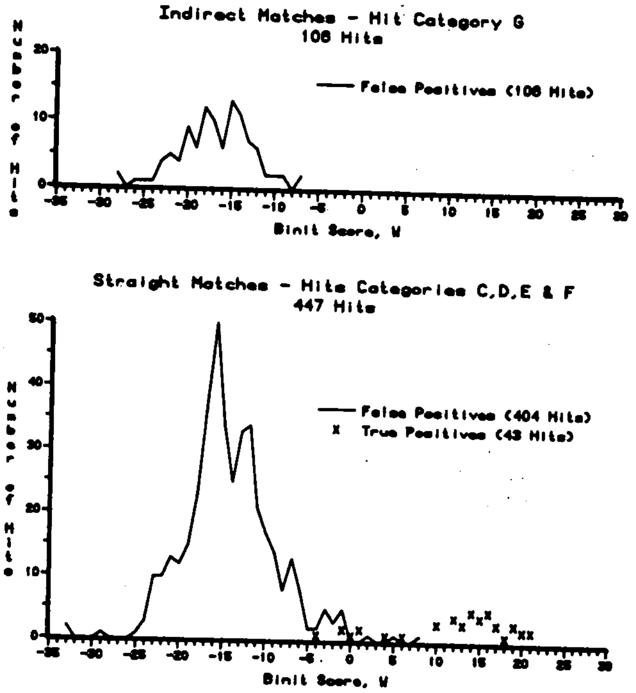
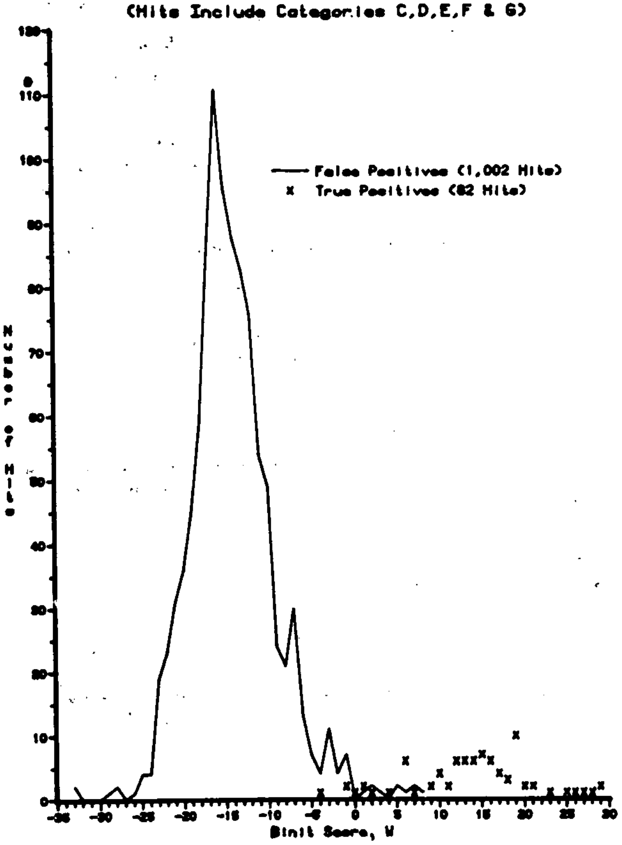


Figure 3.
Frequency Distribution of Overall Binit Scores, W,
for 1,084 Hits from Combined Samples
(Hits Include Categories C,D,E,F & G)



We note that for Hit Category A, all 109 hits in the second sample were true positives. For Category B, 5 of the 6 false positives showed disagreement on sex. Since sex agreed on one false+, we would have misclassified this hit by counting it as a true positive.

Using the same weights as before, an overall score, W, was calculated for each hit falling into categories C through G in the second sample (10). The observed frequency distribution for W for the second sample is shown in Figure 2. Figure 2 may be compared directly with Figure 1.

The results for the two samples appear to be quite similar. They have been combined and are shown in Figure 3. Of the 1084 hits studied, 871 false positives, or 80%, were clearly rejected as false positives by the method, 65 true positives, or 6%, were clearly accepted as true positives, and 148 hits or 14% were considered in the questionable range. Of the questionables, 88 were false positives with scores of -6 through -9.

DISCUSSION

The probabilistic method developed here on a sample of 5 states will ultimately be validated on the complete file of hits. To date, the method appears to be moderately successful in that 86% of the hits falling into categories C through G can be separated correctly into true positives and false positives. The remaining 14% of these hits would still require manual review.

For hits in which SSN agrees (categories A and B), for both samples combined, 363 hits were classified correctly by computer and one hit was incorrectly classified.

Of the total 1449 hits, 420 or 29% were accepted as true positives, 880 or 61% were rejected as false positives, and 149 or 10% were set aside for manual review. One hit was misclassified.

To improve our present methods, a number of factors still need to be taken into account. For the probabilistic methods, the most important ones at this stage appear to be include weights for specific agreements on middle initial, race, marital status, state of residence and state of birth. The present scheme includes only weights for general agreement for these items. Specific agreement weights for Soundex names also need to be studied.

NOTES AND REFERENCES

(1) Rogot, E, Feinleib, M, Ockay, KA, Schwartz, SH, Bilgrad, R and Patterson, JE: On the feasibility of linking Census samples to the National Death Index for epidemiologic studies. American Journal of Public Health. In press.

(2) Bilgrad, R: Overview of the National Death Index. Proceedings of the Section on Survey Research Methods, American Statistical Association (1983).

(3) National Center for Health Statistics: User's Manual: The National Death Index, DHHS Publication No. (PHS) 81-1148, September 1981.

(4) Newcombe, HB, Kennedy, JM, Axford, SJ and James, AP: Automatic linkage of vital records. Science 130:954-959, 1959.

(5) Kennedy, JM, Newcombe, HB, Okazaki, EA and Smith, ME: Computer methods for family linkage of vital health records. Atomic Energy of Canada Ltd., Report No. 2222, Chalk River, Ontario 1965.

(6) Fellegi, IP and Sunter, AB: A theory for record linkage. Journal of the American Statistical Association 64, 1183-1210, 1969.

(7) Smith, ME and Newcombe, HB: Methods for computer linkage of hospital admission--separation records into cumulative health histories. Methods of Information in Medicine 14, 118-125 (1975).

(8) Howe, GR and Lindsay, J: A generalized iterative record linkage computer system for use in medical follow-up studies: Computers and Biomedical Research 14, 327-340 (1981).

(9) Newcombe, HB, Smith, ME and Abbatt, JD: Linkage procedures for the Eldorado Mortality Searches--ENL-LINK-2. 1-93. Eldorado Nuclear Limited. Ottawa (1982).

(10) However, for W_{SLN} and for W_{SFN}, for males, relative frequencies of names in a veteran's cohort were used; and, for W_{SFN}, for females, a Canadian file was used.

APPENDIX

CALCULATION OF CONSTANT (C) NEEDED TO DETERMINE AGREEMENT WEIGHTS FOR SPECIFIC LAST NAMES (W_{SLN})

Suppose the following table represents observed frequencies of last names in frequency order, for hits in which last names agree. The relative frequency of these names in a large file--the April 1980 CPS--is also given for computation of true+ to false+ odds for each last name.

Last Name	Observed freq (f_i)	Rel. freq. of name in Apr '80 CPS (p_i)	Odds of true+: false+ ($1:p_iC$)	W_{SLN}
	(1)	(2)	(3)	(4)
Smith	f_1	p_1	$1:p_1C$	$\log_2(1/p_1C)$
Johnson	f_2	p_2	$1:p_2C$	$\log_2(1/p_2C)$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Rarest name	f_k	p_k	$1:p_kC$	$\log_2(1/p_kC)$
			1:1	

(1) Observed frequency of specified name in a given hit category.

(2) Relative frequency of name in a large file. In this report the April 1980 CPS file containing close to 200,000 persons was used.

(3) The odds of a true+ to false+, given agreement on last name and relative frequencies (in (2)), with the overall odds ratio set at 1:1.

(4) Associated binit weight for odds shown in (3).

Problem

Given: $\{f_i\}, \{p_i\}$ and overall odds set at 1:1

Find: C

We have

$$\frac{f_1(1:p_1C) + f_2\ (1:p_2C) + \ldots + f_k(1:p_kC)}{\Sigma f_i} = 1:1$$

since the overall odds (1:1) is a weighted average of the individual odds. This can be written as

$$\frac{(f_1+f_2+\ldots+f_k)\ :\ C\ (f_1p_1+\ldots+f_kp_k)}{\Sigma f_i} = 1:1$$

or

$$1:\frac{C}{\Sigma f_i}\ (\Sigma f_ip_i) = 1:1$$

Let $\bar{p} = \frac{\Sigma f_ip_i}{\Sigma f_i}$. Then $C\bar{p} = 1$ or $C = \frac{1}{\bar{p}}$.

Example showing how W_{SLN} was calculated for each of 102 hits in category C from a sample of 5 states.

Last Name	Observed freq (f_i)	Rel. freq. of name in Apr '80 CPS (p_i)	Odds of true+: false+ ($1:p_iC$)	W_{SLN}
Smith	17	.00971	1:2.3916	-1.2
Johnson	8	.00768	1:1.8916	- .9
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Rarest name	1	.00001	1:.0246	+5.3
	102		1:1	

$\bar{p} = \frac{\Sigma f_ip_i}{\Sigma f_i} = .00406$ $C = \frac{1}{\bar{p}} = 246.3054$