EVALUATION OF THE MATCHING EFFECTIVENESS OF THE NATIONAL DEATH INDEX

John E. Patterson, National Center for Health Statistics

The National Death Index (NDI) is a central, computerized index of death record information compiled from magnetic tapes submitted to the National Center for Health Statistics (NCHS) by all the State vital statistics offices [1], [2]. These tapes (beginning with deaths occurring in 1979) contain a standard set of identifying information for each decedent to be used in identifying and locating death records filed in the State offices.

The NDI is primarily designed to provide health and medical investigators with a simplified mechanism for determining whether persons in their prospective statistical studies may have died, and if so, to indicate the States in which the deaths occurred and the corresponding death certificate numbers.

Investigators are then able to procure copies of death certificates from the appropriate States to obtain specific statistical information such as cause of death. In the past, investigators conducting such studies have often found it necessary to contact all or most State vital statistics offices, asking each to search its files to see if a death record may have been filed for any individual in their entire study group.

The extent to which the NDI is able to assist in the mortality ascertainment activities of such investigators is contingent on three factors:

1. The quality and completeness of the State death record information stored in the NDI file.
2. The quality and completeness of the information on study subjects submitted by investigators for searches against the NDI file.
3. The effectiveness of the NDI matching criteria used to determine which NDI records should be identified as possible matches with a user's records.

All three of these factors are interrelated and of major importance, but we can probably achieve the greatest short-term improvement in our NDI services by evaluating the effectiveness of the current NDI matching criteria and then making modifications as indicated by the evaluation. The purpose of this paper is to describe how the current criteria are being evaluated, to summarize the evaluation results produced thus far, and to report on some new matching criteria that are now being tested as a result of our evaluation.

1. CURRENT NDI MATCHING PROCESS

When submitting records to be searched against the NDI file, users should provide as many of the following NDI data items as possible for each name submitted:

Matching criteria data items:

1. First name
2. Last name
3. Father's surname (for females)
4. Social security number
5. Month and year of birth
6. Sex

Other data items:

7. Middle initial
8. Day of birth
9. State of birth
10. State of residence (last known)
11. Race
12. Marital status
13. Age at death (actual or estimated)

Only selected combinations of the six "matching criteria data items" are used in the searching procedure to determine whether a particular death record in the NDI file constitutes a "possible match" with a given user record. The remaining data items (items 7-13) are not used to determine possible NDI record matches, but they are matched against corresponding items in the NDI file in each case where a possible record match does occur on the basis of some combination of the six matching criteria data items. The user should examine matches or nonmatches on the items not used in the searching procedure to assess the quality of each possible record match.

As a minimum, each user request record must contain a FIRST and LAST NAME and either a SOCIAL SECURITY NUMBER or a MONTH and YEAR OF BIRTH.

Users should also provide the FATHER'S SURNAME for female subjects in their projects, if at all possible, since the FATHER'S SURNAME will remain constant. The LAST NAMES of female subjects, in contrast, may change as a result of marriage, divorce, widowhood, and remarriage.

The NDI computerized retrieval program is used to search the NDI file to determine whether an NDI death record qualifies as a possible match for a record in the user's file. For an NDI death record to qualify as a possible match with a user request record, both records must satisfy at least one of the following seven conditions, or current matching criteria:

1. SOCIAL SECURITY NUMBER and FIRST NAME match on both records.
2. SOCIAL SECURITY NUMBER and LAST NAME match on both records.
3. SOCIAL SECURITY NUMBER and FATHER'S SURNAME match on both records.
4. If the subject is female, SOCIAL SECURITY NUMBER matches and the LAST NAME on the user request record matches the FATHER'S SURNAME on the NDI record.
5. MONTH and YEAR OF BIRTH and FIRST and LAST NAME match on both records.
6. MONTH and YEAR OF BIRTH and FIRST NAME and FATHER'S SURNAME match on both records.
7. If the subject is female, MONTH and YEAR OF BIRTH and FIRST NAME match and the LAST NAME on the user request record matches the FATHER'S SURNAME on the NDI record.

In the case of FIRST NAME, LAST NAME, or FATHER'S SURNAME, the matching criteria are satisfied if a match occurs on either an "exact spelling" or a "soundex" basis [3].

NDI records involved in matches, based on any of the seven matching criteria, should be considered as only possible matches. The user must review the output generated from the search of the NDI file to assess the quality of each possible record match listed, and then decide which possible matches are worthy of followup

with the State offices.

In this regard, it is extremely useful to include MIDDLE INITIAL, DAY OF BIRTH, STATE OF BIRTH, and STATE OF RESIDENCE on the user request records. These items are highly discriminating in assessing the quality of possible record matches. Under the seven matching criteria described above, a given user request record may generate possible record matches with more than one NDI record, especially when the subject's name is common; e.g., John Smith, Mary Jones.

The Retrieval Report is the primary output of an NDI file search. This computer-generated report indicates which NDI records are involved in possible record matches with particular user request records. Figure 1 shows a sample of the format and content of the Retrieval Report. The information above the first horizontal line represents the data submitted by the user. The information between the two horizontal lines represents the output of the NDI search.

2. EVALUATION ACTIVITIES

In our evaluation, we are reviewing three matching conditions generated by the seven current matching criteria:

1. True Matches -- NDI records are correctly identified and listed with their corresponding user records; i.e., true positives.
2. Nonmatches -- Persons on the user's file are deceased; however, the corresponding NDI records are not found; i.e., false negatives.
3. Mismatches -- NDI records are incorrectly linked with user records; i.e., false positives.

One objective of the evaluation is to assess how effective the seven current matching criteria are in generating true matches as well as how many mismatches are generated in the process. Nonmatches are also being studied to determine what changes in the matching criteria should be considered in order to maximize the number of true matches .

We have begun our evaluation efforts with test data consisting entirely of records for persons known to have died during a specific time period. We were able to obtain such files from two registries of persons diagnosed as having cancer. (These cancer registries are not identified in this paper because of confidentiality considerations.) The file for cancer registry A included 2,598 records of persons known to have died in 1979 in State A. The file for cancer registry B included 7,058 records of persons known to have died in 1979 in State B. All cancer registry deaths occurring in other States were excluded because the out-of-state death certificate numbers were not available from the cancer registries and because such deaths would have complicated the selection of a sample of NDI records to be used in the test.

An important reason for selecting these cancer registry records for testing was that each cancer registry was able to provide the decedent's death certificate numbers which each registry routinely recorded in the course of its normal mortality ascertainment activities. After we completed the computerized matching process for this test, we compared the certificate numbers and the names on both the NDI and the cancer registry records to determine which NDI records were matched correctly with their corresponding cancer

FIGURE 1. **Retrieval Report**

User's request:

		Birth Date					
Possible decedent Regina V Hanes	Soc Sec No	Mo Dy Yr	Age	S M R	SOR	SOB
Father's surname Franklin	114-57 6493	08 11 1901	—	F M W	AL	ND

Possible NDI record matches:

State of Death	Certif Number	Date of Death	Name F M L	Father's Surname	Soc Sec No	Birth Date Mo Dy Yr	Age	S M R	SOR	SOB
Alabama	698637	04 09 79	X X X	X	X	X X X	—	X X X	X	X
North Dakota	421304	03 01 79	X X			X X	—	X X X		X
Georgia	307698	03 06 79	X	S	?	X X	—	X X		
New Hampshire	407635	12 11 79	X S			X X	—	X X		?

Symbols and abbreviations

S = Sex
M = Marital status
R = Race
SOR = State of residence
SOB = State of birth

X = User's data item exactly matched NDI data item
S = Soundex match on name
? = Insufficient information on NDI file
— = Data item not supplied by user
Blank = User data item and NDI data item did not match

registry or "user" records and also to determine which NDI records were involved in mismatches.

The existence of certificate numbers on the cancer registry records was extremely helpful in assessing problems resulting in nonmatches. Whenever a cancer registry record was not involved in a match with its corresponding NDI record, the certificate number was used to retrieve the NDI record to see why none of the current seven matching criteria was satisfied.

Both of the cancer registry test files were searched against an NDI test file consisting of 500,000 death records for 1979. The cancer registry files contained only 1979 decedents for States A and B, and all 70,000 NDI records for these two States for 1979 were included in the test file of 500,000 records to insure that all of the cancer registry test records had an opportunity to be matched with their corresponding NDI records. The remaining 430,000 NDI records represented a systematic sample of 1979 deaths from all other States to evaluate the extent to which the matching criteria generate mismatches. The completeness of the NDI data items in each of the cancer registry test files and in the NDI test file is shown in Table 1.

The main premise upon which these tests are based is that, in theory, 100 percent of each cancer registry's records should generate a true match with the corresponding NDI records. The overall effectiveness of the NDI file search (using the current matching criteria) was measured in terms of what percent of the cancer registry records did in fact match correctly. The effectiveness of each individual matching criteria was measured in the same way.

3. EVALUATION RESULTS

Table 2.A. shows the overall effectiveness of the seven current matching criteria. Column (5) shows that true matches were generated for 92.1 percent of the records searched for cancer registry A and 99.6 percent of the records searched for cancer registry B.

This table also presents the effectiveness of the individual matching criteria and two groupings of the criteria. It is important to emphasize, however, that many true matches generated by one of the seven criteria were also generated by one or more of the remaining criteria; i.e., the true matches shown for the various criteria are not mutually exclusive. The percentages of true matches resulting from criteria based on Social Security Number and names, i.e., criteria 1-4, were somewhat higher than the percentages of true matches based on month and year of birth and names, i.e., criteria 5-7.

At this point, it should be noted that the proportion of true matches generated for cancer registry B (99.6 percent) appeared to be too good to be true. After the test match was completed, we went back to the staff of that cancer registry and belatedly found that, to an unknown extent, information on names, Social Security numbers, and dates of birth from the death certificates had been used to complete, correct, or update some of the corresponding items in the cancer registry's data file.

Such file maintenance procedures would obviously tend to increase the proportion of true matches generated by our matching criteria, since the NDI file is compiled solely from death certificate data. However, we do not know

Table 1. **COMPLETENESS OF NDI DATA ITEMS**

NDI Data Items	Cancer Registry A (2,598 Records)		Cancer Registry B (7,058 Records)		NDI Test File (500,000 Records)	
	Number	Percent	Number	Percent	Number	Percent
Matching Criteria Data Items:						
First Name.....	2,598	100.0	7,058	100.0	499,069	99.8
Last Name.....	2,598	100.0	7,058	100.0	499,884	99.9
Father's Surname.....	646	24.9	0	0.0	444,330	88.9
Social Security No...	2,231	85.9	6,929	98.2	449,940	90.0
Birth Month.....	2,595	99.9	7,057	99.9	493,240	98.6
Birth Year.....	2,595	99.9	7,056	99.9	496,427	99.3
Sex.....	2,598	100.0	7,058	100.0	499,998	99.9
Other Data Items:						
Middle Initial.....	2,006	77.2	6,341	89.8	361,064	72.2
Day of Birth.....	2,596	99.9	7,057	99.9	492,654	98.5
Age at Death.....	2,596	99.9	6,821	96.6	499,875	99.9
Race.....	2,598	100.0	7,058	100.0	499,586	99.9
Marital Status.....	2,556	99.9	0	0.0	497,173	99.4
State of Residence...	2,598	100.0	6,049	85.7	499,981	99.9
State of Birth.....	2,598	100.0	6,572	93.1	497,518	99.5

anything about the extent to which such changes may have occurred. At the present time, we know only that (1) cancer registry A never used information from the death certificates in its file maintenance procedures, and (2) the proportion of true matches generated for its file amounted to 92.1 percent (as shown in Table 2.A.). We also know that three other users of the NDI have tested our matching criteria effectiveness using their data files for known decedents. These three independent evaluations have yielded proportions of true matches amounting to 98.4 percent, 96.5 percent, and 88.0 percent. The differences in these percentages appear to be due mainly to differences in the quality and completeness of data in these three user's files.

The matching criteria were also evaluated on the basis of the number of mismatches they generated. Column (7) of Table 2.A. shows the number of NDI records involved in mismatches, and column (8) shows the ratio of mismatched NDI records to the number of eligible user records. The overall ratios of mismatched NDI records were fairly similar for cancer registry A and cancer registry B. Matches based on criteria involving Social Security Numbers (criteria 1-4) generated a much smaller ratio of mismatches than did criteria involving month and year of birth (criteria 5-7). While these ratios appear small, it should be noted that they are based on searches of a test NDI file containing only 500,000 records. The ratios will increase if more NDI records are searched. We are not as concerned, however, about the problem of mismatches as we might be for reasons discussed in the next section of this paper.

Table 2.B. presents the cumulative effectiveness of the current seven matching criteria in generating true matches. The small differences between most of the figures in column (2) reflect, in part, the fact that the true matches shown in this table are not mutually exclusive. Most of the true matches were generated by two or more of the seven criteria.

The marginal effectiveness of criteria 3, 4, 6, and 7, which involve the use of the father's surname, was extremely low for both cancer registries. As shown in Table 2.B., the addition of these criteria to criteria 1, 2, and 5 yields very few new true matches. We should not necessarily eliminate criteria 3, 4, 6 and 7, however. These criteria are intended to accommodate matching problems due to changes in the last names of women as a result of marriage, divorce, widowhood, and remarriage. Their low level of effectiveness in this study can almost certainly be attributed to the fact (1) cancer registry A provided few fathers' surnames and cancer registry B provided none (see Table 1); and (2) women who have been diagnosed as having cancer and are thus included in these two cancer registries are much less likely than other females to undergo name changes. These four criteria should be extremely effective in generating additional true matches for subjects in study groups that include a large proportion of women who are more likely to change their marital status.

4. ELIMINATING MISMATCHES

Column (7) of Table 2.A. shows that the seven current matching criteria generated 44 mismatches for cancer registry A and 99 for cancer registry B. A review of the records involved in these mismatches indicates that the typical NDI user would have had little difficulty in eliminating a very large majority of the mismatches on the basis of a visual inspection of the computerized output of the NDI search (see Figure 1). First, 140 of the 143 mismatches involved cases where two or three NDI records were linked with one cancer registry record. This would obviously have alerted the user to the problem. Second, visual inspection of the NDI output (to see if there was agreement or disagreement on the nonmatching data items such as middle initial, day of birth, State of birth and State of residence) made it fairly easy to eliminate at least three-fourths of the mismatches without eliminating true matches in the process. We need to look at the problems of mismatches in more detail, but the results of our evaluation, thus far, indicate that mismatches constitute a problem only for those NDI users who submit such large data files that visual inspection of the NDI output is impractical [4] or for those NDI users who submit only the minimum number of data elements required for a search.

5. ASSESSING NONMATCHES

We also assessed the reasons for the nonmatches; i.e., those cases where we did not find an NDI record for a known decedent in the cancer registry files. Column (3) of Table 2.A. shows that 204 records from cancer registry A were not linked with their corresponding NDI records, while only 26 records from cancer registry B were not linked. We retrieved each of the corresponding NDI records on the basis of the death certificate number provided on the cancer registry record. The pairs of NDI and cancer registry records were then inspected to see why these records did not satisfy any of the seven current matching criteria. The principal reasons for these nonmatches are summarized in Table 3.

The nonmatches involved discrepancies in Social Security numbers, names, and/or months and years of birth. Most of the nonmatching pairs of records had missing or incorrect Social Security numbers. This caused the records to be matchable only on the basis of criteria involving month and year of birth in combination with names (criteria 5-7). A very large proportion of the nonmatches for cancer registry A were due to discrepancies in the year of birth. The year of birth did not agree in 149 of the 204 nonmatching pairs. Of these 149 nonmatches, however, we found that 106 did agree on month and day of birth.

6. POTENTIAL REVISIONS IN THE MATCHING CRITERIA

On the basis of our evaluation of the seven current matching criteria, we have tested a number of new criteria involving different combinations of names and initials with either (1) Social Security number, (2) month and day of birth, or (3) month and year of birth + 1 year. Five new matching criteria were believed to have had the greatest potential payoff in increasing true matches. These new criteria are listed below and are numbered 8 through 12, to distinguish them from the seven current criteria. They generated possible NDI record matches if any of the five combinations of data items matched on both records.

8. Month and year of birth ±1 year, first name, last name

Table 2.A.

EFFECTIVENESS OF CURRENT MATCHING CRITERIA: TEST RESULTS

(Matches shown in this table are NOT mutually exclusive; i.e., many of the matches generated by one criteria were also generated by one or more of the remaining criteria.)

Current Matching Criteria	Number of User Records			Percent			Number of NDI Records Involved in Mismatches (7)	Ratio of NDI Mismatches to Total Eligible User Records (8)
	Total Eligible for a Match (1)	True Matches (2)	Non-Matches (3)	Total Eligible for a Match (4)	True Matches (5)	Non-Matches (6)		
CANCER REGISTRY A								
All Criteria 1-7.....	2,598	2,394	204	100.0	92.1	7.9	44	.0169
SSN and Names: 1-4.....	2,231	1,874	357	100.0	84.0	16.0	7	.0031
1.....	2,231	1,796	435	100.0	80.5	19.5	0	.0000
2.....	2,231	1,861	370	100.0	83.4	16.6	7	.0031
3.....	553	368	185	100.0	66.5	33.5	0	.0000
4.....	406	61	345	100.0	15.0	85.0	1	.0025
Date of Birth (Month/Year) and Names: 5-7.....	2,596	2,069	527	100.0	79.7	20.3	37	.0143
5.....	2,596	2,064	532	100.0	79.5	20.5	30	.0116
6.....	645	458	187	100.0	71.0	29.0	5	.0078
7.....	512	83	429	100.0	16.2	83.8	3	.0059
CANCER REGISTRY B								
All Criteria 1-7.....	7,058	7,032	26	100.0	99.6	0.4	99	.0140
SSN and Names: 1-4.....	6,929	6,663	266	100.0	96.2	3.8	25	.0036
1.....	6,929	6,515	414	100.0	94.0	6.0	0	.0000
2.....	6,929	6,623	306	100.0	95.6	4.4	25	.0036
3.....	0	NA	NA	NA	NA	NA	NA	NA
4.....	3,079	280	2,799	100.0	9.1	90.9	15	.0049
Date of Birth (Month/Year) and Names: 5-7.....	7,056	6,710	346	100.0	95.1	4.9	74	.0105
5.....	7,056	6,710	346	100.0	95.1	4.9	59	.0084
6.....	0	NA	NA	NA	NA	NA	NA	NA
7.....	3,159	269	2,890	100.0	8.5	91.5	19	.0060

Table 2.B.

EFFECTIVENESS OF CURRENT MATCHING CRITERIA: TEST RESULTS

Current Matching Criteria	Number of User Records			Percent			Number of NDI Records Involved in Mismatches (7)	Ratio of NDI Mismatches to Total Eligible User Records (8)
	Total Eligible for a Match (1)	True Matches (2)	Non-Matches (3)	Total Eligible for a Match (4)	True Matches (5)	Non-Matches (6)		
CANCER REGISTRY A								
1.....	2,231	1,796	435	100.0	80.5	19.5	0	.0000
1 and 2.....	2,231	1,873	358	100.0	84.0	16.0	7	.0031
1,2 and 3.....	2,231	1,874	357	100.0	84.0	16.0	7	.0031
1,2,3 and 4.....	2,231	1,874	357	100.0	84.0	16.0	7	.0031
1,2,3,4 and 5.....	2,598	2,393	205	100.0	92.1	7.9	37	.0142
1,2,3,4,5 and 6.....	2,598	2,393	205	100.0	92.1	7.9	42	.0162
1,2,3,4,5,6 and 7.....	2,598	2,394	204	100.0	92.1	7.9	44	.0169
CANCER REGISTRY B								
1.....	6,929	6,515	414	100.0	94.0	6.0	0	.0000
1 and 2.....	6,929	6,660	269	100.0	96.1	3.9	25	.0036
1,2, and 3.....	6,929	6,660	269	100.0	96.1	3.9	25	.0036
1,2,3 and 4.....	6,929	6,663	266	100.0	96.2	3.8	25	.0036
1,2,3,4 and 5.....	7,058	7,032	26	100.0	99.6	0.4	84	.0119
1,2,3,4,5 and 6.....	7,058	7,032	26	100.0	99.6	0.4	84	.0119
1,2,3,4,5,6 and 7.....	7,058	7,032	26	100.0	99.6	0.4	99	.0140

Table 3

REASONS FOR NONMATCHES RESULTING FROM CURRENT MATCHING CRITERIA

(The frequencies and percentages presented in this table are not mutually exclusive.)

Reasons for Nonmatches	Nonmatches			
	Cancer Registry A (204 nonmatches)		Cancer Registry B (26 nonmatches)	
	Number	Percent of Non-matches	Number	Percent of Non-matches
SSN DISCREPANCIES:	195	95.7	25	96.0
No SSN on user's record.....	90	44.1	3	11.5
No SSN on NDI record.....	20	9.8	6	23.1
No SSN on either record.....	13	6.4	5	19.2
Entirely different SSN's.....	42	20.6	3	11.5
SSN's 1 digit off.....	22	10.8	5	19.2
SSN's 2 to 3 digits off.....	4	2.0	2	7.7
Transposition of digits.....	4	2.0	1	3.8
NAME DISCREPANCIES:.....	52	25.6	17	65.3
First and middle names reversed.....	19	9.3	2	7.7
Other combinations of names reversed..	4	2.0	3	11.5
Spelling or keying errors.....	13	6.4	6	23.1
Nickname on one record.....	11	5.4	2	7.7
Two first names on one record.....	3	1.5	3	11.5
Use of initial for first name.....	2	1.0	1	3.8
YEAR OF BIRTH DISCREPANCIES:.....	149*	73.0	8**	30.8
Birth year off by ± 1.....	58	28.4	6	23.1
Birth year off more than ± 1.....	91	44.6	2	7.7
MONTH OF BIRTH DISCREPANCIES:.....	30	14.7	3	11.5

* Of the 149 records which did not match on year of birth, 106 matched on month and day of birth.

** Of the 8 records which did not match on year of birth, all 8 matched on month and day of birth.

Table 4.A.

EFFECTIVENESS OF CURRENT AND POTENTIAL NEW MATCHING CRITERIA: TEST RESULTS

(Matches shown in this table are NOT mutually exclusive; i.e., many of the matches generated by one criteria were also generated by one or more of the remaining criteria.)

Matching Criteria	Number of User Records			Percent			Number of NDI Records Involved in Mismatches (7)	Ratio of NDI Mismatches to Total Eligible User Records (8)
	Total Eligible for a Match (1)	True Matches (2)	Non-Matches (3)	Total Eligible for a Match (4)	True Matches (5)	Non-Matches (6)		
CANCER REGISTRY A								
Current Criteria: 1-7.....	2,598	2,394	204	100.0	92.1	7.9	44	.0169
Potential Criteria: 8-12...	2,596	2,351	245	100.0	90.6	9.4	108	.0416
8.....	2,596	2,192	404	100.0	84.4	15.6	34	.0131
9.....	2,596	2,168	428	100.0	83.5	16.5	65	.0250
10.....	2,596	1,420	1,176	100.0	54.7	45.3	6	.0023
11.....	2,596	1,445	1,151	100.0	55.7	44.3	15	.0058
12.....	2,596	1,489	1,107	100.0	57.4	42.6	6	.0023
CANCER REGISTRY B								
Current Criteria: 1-7.....	7,058	7,032	26	100.0	99.6	0.4	99	.0140
Potential Criteria: 8-12.....	7,056	6,875	181	100.0	97.4	2.6	200	.0283
8.....	7,056	6,764	292	100.0	95.9	4.1	61	.0086
9.....	7,056	6,707	349	100.0	95.1	4.9	99	.0140
10.....	7,056	5,714	1,342	100.0	81.0	19.0	27	.0038
11.....	7,056	5,700	1,356	100.0	80.8	19.2	29	.0041
12.....	7,056	5,755	1,301	100.0	81.6	18.4	30	.0043

Table 4.B.

EFFECTIVENESS OF CURRENT AND POTENTIAL NEW MATCHING CRITERIA: TEST RESULTS

(Matches shown in this table are NOT mutually exclusive; i.e., many of the matches generated by one criteria were also generated by one or more of the remaining criteria.)

Matching Criteria	Number of User Records			Percent			Number of NDI Records Involved in Mismatches (7)	Ratio of NDI Mismatches to Total Eligible User Records (8)
	Total Eligible for a Match (1)	True Matches (2)	Non-Matches (3)	Total Eligible for a Match (4)	True Matches (5)	Non-Matches (6)		
CANCER REGISTRY A								
Current Criteria: 1-7.....	2,598	2,394	204	100.0	92.1	8.5	44	.0169
Current and Potential Criteria: 1-7 and 8-12.....	2,598	2,500	98	100.0	96.2	3.8	122	.0470
1-7 and 8.....	2,598	2,439	159	100.0	93.9	6.1	48	.0185
1-7 and 9.....	2,598	2,496	102	100.0	96.1	3.9	108	.0416
1-7 and 10.....	2,598	2,398	200	100.0	92.3	7.7	48	.0185
1-7 and 11.....	2,598	2,446	152	100.0	94.1	5.9	58	.0223
1-7 and 12.....	2,598	2,425	173	100.0	93.3	6.7	48	.0185
CANCER REGISTRY B								
Current Criteria: 1-7.....	7,058	7,032	26	100.0	99.6	0.4	99	.0140
Current and Potential Criteria: 1-7 and 8-12.....	7,058	7,047	11	100.0	99.8	0.2	240	.0340
1-7 and 8.....	7,058	7,038	20	100.0	99.7	0.3	101	.0143
1-7 and 9.....	7,058	7,040	18	100.0	99.7	0.3	196	.0278
1-7 and 10.....	7,058	7,039	19	100.0	99.7	0.3	119	.0169
1-7 and 11.....	7,058	7,043	15	100.0	99.8	0.2	127	.0180
1-7 and 12.....	7,058	7,041	17	100.0	99.8	0.2	122	.0173

9. Month and day of birth, first name, last name
10. Month and year of birth, first and middle initials, last name
11. Month and day of birth, first and middle initials, last name
12. Month and year of birth ±1 year, first and middle initials, last name

Table 4.A. shows the extent to which each of the five new criteria generated true matches, nonmatches, and mismatches using data from the two cancer registries. Column (5) shows that, in terms of generating true matches, criteria 8 and 9 were by far the most effective. Column (5) also shows that criteria 8, which involves month and year of birth ±1 year, was somewhat more effective than criteria 9, which involves month and day of birth.

Table 4.B. indicates the effectiveness of each of the five new criteria when used in conjunction with the seven current criteria. In contrast to the preceding table, column (5) of this table shows that criteria 9 is the most effective in generating additional true matches and is somewhat more effective than criteria 8 for Cancer Registry A. This apparent anomaly reflects the fact that a larger proportion of the true matches generated by criteria 8 were also generated by the seven current criteria. This, of course, reduces the effectiveness of adding criteria 8 to the seven existing criteria.

We also found that a very large proportion of the true matches generated by criteria 8, 10, 11 and 12 were also generated by criteria 9, which involves month and day of birth. As a result, column (5) of Table 4.B. indicates that the addition of only criteria 9 to the seven existing criteria was almost as effective as the addition of all five of the new criteria.

However, column (7) shows that the addition of criteria 9 does result in a doubling of the number of NDI records involved in mismatches. We are currently reviewing the reasons for these additional mismatches to see if most of them can be readily eliminated by a visual inspection of the NDI search output.

7. CONCLUSIONS

In summary, our evaluation indicates that we can reduce the number of nonmatches by about 50 percent by adding five new matching criteria to the seven current criteria. In fact, we can achieve virtually the same reduction in nonmatches by adding only criteria 9. However, the addition of this criteria substantially increases the number of mismatches, and we need to evaluate the implications of this problem more fully.

We also need to carefully review the effectiveness of phonetic matching techniques since our current matching criteria permit matching on names on the basis of either exact spelling or Soundex codes. Our results thus far indicate that matches generated on the basis of Soundex codes increase the proportion of true matches by only 1 percent, although about 50 percent of the mismatches involve Soundex codes. We have also begun testing another phonetic matching technique called the New York State Identification and Intelligence System (NYSIIS) [5]. Our preliminary results indicate that this system generates even fewer additional true matches than Soundex, but it does generate substantially few mismatches.

Finally, we need to work with other test files which contain known decedents having different characteristics than those represented in the cancer registry files.

Any changes which we propose to implement in our matching criteria will also need to be tested to determine if they cause a significant increase in the amount of computer time required for NDI file searches. This is important because of the large number of users which we expect to serve and the increasing size of the NDI file which is updated annually with 2 million records.

NOTES AND REFERENCES

[1] User's Manual: The National Death Index, September 1981, National Center for Health Statistics.

[2] Bilgrad, Robert, "Overview of the National Death Index," 1983 Proceedings, American Statistical Association, Section on Survey Research Methods.

[3] 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 names.

[4] For more information on matching large data files against the NDI, refer to Rogot, E. et al., "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," in this volume.

[5] Lynch, Billy T. and Arends, William L., "Selection of a Surname Coding Procedure for the SRS Record Linkage System," Statistical Reporting Service, U.S. Department of Agriculture, February 1977.