## EXHIBIT 13: LAST NAME EXAMPLES OF NYSIIS RECODES

| **LAST NAME** = JOHNSON<br>**NYSIIS** = JANSAN* | **LAST NAME** = MORSE<br>**NYSIIS** = MARS* | **LAST NAME** = ANDERSON<br>**NYSIIS** = ANDARSAN |
|---|---|---|
| JOHNSON | MORSE | ANDERSON |
| ***JANZEN*** | MAERSCH | AMDERSON |
| ***JEANSON*** | MARSAU | ANDEERSON |
| JAENSEN | MARSAW | ANDERSEN |
| JAHNSEN | MARSCHE | ANDERSHIN |
| JAHNSON | MARSE | ANDERSIN |
| JAHNSSEN | MARSH | ANDERSOHN |
| JAMSEN | MARSHAE | ANDERSOM |
| JAMSION | MARSHAUS | ANDERSONS |
| JANNSOHN | MARSHAW | ANDERSSEN |
| JANSEN | MARSHIO | ANDERSSON |
| JANSENIUS | MARZIO | ANDERZHON |
| JANSHEN | MARZOA | ANDORSON |
| JANSON | MEARSE | ANMDERSON |
| JANSONIUS | MEHRZAI | |
| JANSSEN | MERSI | |
| JANSSON | MEYERSHAW | |
| JANSZEN | MORSA | |
| JANZAN | MORSCH | |
| JANZANO | MORSEAU | |
| JEANSONNE | MORSESR | |
| JEHNSEN | MORZO | |
| JEMSON | MOURSI | |
| JENSEMA | MUERSCH | |
| JENSEN | MURSAU | |
| JENSSON | MURSCH | |
| JENSYNN | MURSE | |
| JOHNSEN | MURSU | |
| JOHNSION | MURZI | |

*** This is not an exhaustive list of all the last names associated with the NYSIIS code.**

# APPENDIX A

## A PROBABILISTIC SCORING APPROACH FOR ASSESSING NATIONAL DEATH INDEX MATCH RESULTS

**Caution**: NDI users should be aware that submission records containing a significant number of missing data items (for example not collecting state of birth or state of residence) will have lower overall probabilistic scores and may as a consequence underestimate mortality for their cohort if additional clerical review of potential match records is not conducted.

The probabilistic scoring technique described in this Appendix is intended only to aid (or guide) NDI users in determining which NDI record matches are likely to be ***true*** matches. Please read this appendix carefully before attempting to use the scores in your assessment of NDI matches. Please note that the cut-off scores are fairly conservative, meaning that a ***status code*** of 1 implies that there is a high probability that the NDI record is a ***true*** match (i.e., the study subject is assumed to be deceased). On the other hand, some portion of those NDI matches assigned a status code of 0 (assumed alive) may in fact also be ***true*** matches. The final responsibility of determining ***true*** and ***false*** matches rests with the NDI user.

### NDI Matching Methodology

The NDI is designed to facilitate health related mortality studies. Researcher supplied submission files are matched to the NDI computerized index of death record information compiled from death certificates submitted by State Vital Statistics offices to the National Center for Health Statistics (NCHS). Matching user submission records to the NDI is a two-step process. In the first step, the NDI system selects potential death record matches based on a set of seven matching criteria. The second step consists of a scoring and classification procedure that results in the assignment of a probabilistic score and a suggested determination of final match status by NCHS. The approach taken to classify researcher-supplied records to the NDI matches is a modification of the probabilistic approaches developed by Fellegi and Sunter (1969) and Rogot, Sorlie, and Johnson (1986).

### Selecting NDI records

The NDI matches user submission records to death records based on seven criteria:

1. Social Security number
2. Exact month and +/- 1 year of birth, first and last name
3. Exact month and +/- 1 year of birth, first and middle initials, last name
4. Exact month and day of birth, first and last name
5. Exact month and day of birth, first and middle initials, last name
6. Exact month and year of birth, first name, father's surname
7. *If the subject is female*: exact month and year of birth, first name, last name (***on user's record***) and father's surname (***on NDI record***)

Record matches between NDI records and user records are referred to as possible matches. An NDI record is selected as a possible match to a user record if it matches on any one of the seven criteria. In the case of multiple NDI records returned for a given user record, the potential for a large number of false positives may occur. Of those matches listed, one may be a ***true*** match – but it is also possible that none

may be a ***true*** match. Alternatively, it is also possible that no NDI record will be selected for a given user record.

Indications of agreement between the user record and the NDI record are returned to the user for each possible match record. In addition to the data items involved in the seven matching criteria, the NDI results return an indication of agreement for up to five additional data items: age at death; race; marital status; state of residence; and state of birth.

## Scoring and classification of potential matches

Assessing the quality of possible matches and determining the best match for each user submission record requires a consistent approach. Each NDI possible match record is assigned a probabilistic score. The probabilistic score is the sum of the weights assigned to each of the identifying data items used in the NDI record match, where the weights reflect the degree of agreement between the information on the submission record and the NDI death record. NCHS developed the weights, known as binit weights, based upon the frequency of occurrence of the identifying data items in the NDI files for years 1986 to 1991 and the 1988-1991 National Health Interview Survey (a nationally representative survey of the non-institutionalized U.S. population).

A weight is the base 2 logarithm of the inverse of the probability of occurrence of the characteristic based on the above files. For example, since males constitute about 46.3 percent of the population aged 18 and over, the weight is $\log_2(1/.463) = 1.11$. Weights are constructed in a similar manner for race, last name, father's surname, birth month, day, and year, state of residence, and state of birth. The last name weights have been modified for females. Since females have historically changed their surnames upon marriage, divorce, and remarriage, matching on surname only may produce a false non-match. The NDI returns an indication of either match or non-match on father's surname as well as last name. Since a person's father's surname does not change over time this is used as auxiliary information for females. If last name does not match on the two records (the last name weight is negative), the last name weight is replaced with the father's surname weight if positive (matches), otherwise the last name weight is retained. Since middle initials are sex-specific, sex-specific weights were constructed for middle initial. Weights for marital status were constructed to be jointly age and sex specific. Common first names, such as "John" that have a higher probability of occurrence, receive a lower binit weight than uncommon names, such as "Jonas". First name weights are both sex and birth year cohort specific (<1926, 1926 - 1935, 1936-1955, and >1955) since there are some secular trends in the assignment of first names. The weight assigned for SSN is a constant value of 30.

Weights are either positive or negative. If there is agreement between the user record and the NDI record for a particular identifying data item, the weight is positive. If there is no agreement, the weight is negative. Some items, such as year of birth, allow a tolerance (+/ - 3 years) and are still considered to agree. With the exception of middle initial, data items that are missing on the user's submission record, the NDI record, or both receive a weight of zero. A blank middle initial is considered a valid value and receives the appropriate weight. The score for each potential match is the sum of the weights for each individual data item.

$$\textbf{Score} = W_{SSN} + W_{firstname\ x\ sex\ x\ birthyear} + W_{middleinitial\ x\ sex} + W_{lastname} + W_{race} + W_{sex} + W_{maritalstatus\ x\ sex\ x\ age} + W_{birthday} + W_{birthmonth} + W_{birthyear} + W_{stateofbirth} + W_{stateof\ residence}$$

After scoring the potential matches, each is categorized into one of five mutually exclusive classes. Whereas weighting and scoring take into account the probability that the submission record and the NDI record share a particular value for the identifying items, the classes take into account which identifying items agree. They reflect the fact that some of the NDI identifying data items used in the matching criteria are more important for determining true matches than others. For example, as SSN is a key identifier in the matching process, each NDI record match is initially classified according to whether SSN is present and agrees (Class 1 or 2), is present but disagrees (Class 5) or is unknown (Class 3 or 4). Additionally, non-changing, identifying information is more important than information that can change

over time. Many women, for example, assume their spouse's name at marriage, a common example of legitimate change over time. Birth surname, however, does not change and is thus an important matching variable for women. By contrast, state of residence and m arital status may change over time and are, therefore, less important as classification variables.

The final five classification groups developed by the NDI are as follows:

Class 1: Exact match on SSN, (all nine digits), first name, middle initial, la st names, sex, state of birth, birth month and birth year.

Class 2: SSN matches on at least *seven* digits and one or more of the other items from Class 1 may not match.
*Note: Some matched cases are moved from Class 2 to Class 5 because of an indication that the reported SSN belongs to the spouse. This includes those cases for which the SSN is known and matches, but the first name and sex do not agree.*

Class 3: SSN unknown but eight or more of first name, middle initial, last name, birth day, birth month, birth year, sex, race, marital status, or state of birth match.

Class 4: Same as Class 3 but less than eight items match.

Class 5: SSN is known but doesn't match.
*Note: Some matched cases are moved from Class 5 to Class 3 because of an indication that one of the SSN's (on the user record or on the death certificate) may have been reported incorrectly but a significant number of other data items are in agreement.*

In this classification scheme all of Class 1 matches are considered to be true matches implyi ng that the individuals are deceased. All of the Class 5 matches are considered false matches. Assignment of records falling into one of Classes 2, 3, or 4 as either true matches or false matches are based on score cut-off points within each class, as sho wn in **Table 1**. Records with scores greater than the cut -off scores are considered true matches while records with scores lower than the cut -off scores are considered false matches. The recommended cut-off scores were determined on the basis of two calibra tion samples, with consideration given to jointly maximizing the proportion of records correctly classified and minimizing the number of records incorrectly classified.

NDI recommended and alternative cut-off scores are given in Table 1. It is suggeste d that the user adopt the recommended cutoff scores since they were chosen to provide overall optimal performance and are independent of any given study. If it is desired to use alternative cutoff scores to conduct sensitivity analyses, they should be chos en within each class. Table 1 provides estimates of the correct classification rates for each class under the assumption of 1,000 records within each class.

## Table 1 - The Impact of Using Alternative Cutoff Scores

(assuming 1,000 records in each class)

**Class 2**

| | Deaths | | Alive | |
|---|---|---|---|---|
| Score | N = 926 | Percent correct | N = 74 | Percent correct |
| 34.5 | 924 | 99.8 | 7 | 9.1 |
| 39.5 | 924 | 99.8 | 9 | 12.1 |
| **44.5** | **921** | **99.5** | **13** | **18.2** |
| 49.5 | 917 | 99.0 | 18 | 24.2 |
| 54.5 | 906 | 97.8 | 20 | 27.3 |

**Class 3**

| | Deaths | | Alive | |
|---|---|---|---|---|
| Score | N = 959 | Percent correct | N = 41 | Percent correct |
| 27.5 | 959 | 100.0 | 2 | 4.6 |
| 32.5 | 958 | 99.9 | 6 | 13.6 |
| **37.5** | **946** | **98.6** | **23** | **56.1** |
| 42.5 | 933 | 97.3 | 25 | 60.6 |
| 47.5 | 872 | 90.9 | 27 | 66.7 |

**Class 4**

| | Deaths | | Alive | |
|---|---|---|---|---|
| Score | N = 281 | Percent correct | N = 719 | Percent correct |
| 22.5 | 191 | 67.8 | 622 | 86.6 |
| 27.5 | 173 | 61.6 | 684 | 95.2 |
| **32.5** | **143** | **51.0** | **703** | **97.8** |
| 37.5 | 119 | 42.4 | 712 | 99.0 |
| 42.5 | 70 | 24.9 | 718 | 99.8 |

Notes: N is the number of presumed correctly classified deceased and living persons based on a hypothetical sample of 1,000 persons.

The suggested cutoff score is the middle score ( **in bold text**) within each class.

Any decision to use an alternative cutoff score should be made on the basis of both the proportion correctly classified and the numbers of persons correctly classified.

## Example 1

This example of NDI weighting and scoring is based on a hypothetical person w ith the characteristics as given in the table below.

| *Item* | *Value* | *Frequency* | *Weight* |
|---|---|---|---|
| SSN | Unknown | | 0.00 |
| Last name | Robinson | 0.00193 | 9.02 |
| Middle initial | A | 0.07748 | 3.69 |
| First name | Leo | 0.00140 | 9.48 |
| Race | White | 0.83509 | 0.26 |
| Sex | Male | 0.46329 | 1.11 |
| Marital status | Married | 0.05913 | 4.08 |
| Birth day | 10 | 0.03349 | 4.90 |
| Birth month | October | 0.08597 | 3.54 |
| Birth year | 1940 | 0.01418 | 6.14 |
| State of birth | Florida | 0.01552 | 6.01 |
| State of residence | New York | 0.06652 | 3.91 |

Since this record has an unknown Social Security Number but at least eight of more of the following criteria are matched (first name, middle initial, last name, birth day, birth month, birth year, sex, race, marital status, or state of birth), this possible mat ch record would be classified as a Class 3 match. The probabilistic match score then is the sum of the individual item weights that is 52.14. Since 52.14 is greater than the Class 3 recommended cutoff score of 37.5, it would be assumed that this is a true match and the person is deceased.

## Example 2

Using example 1 above, assume that birth day, marital status, and state of residence do not match but the remaining items do match between the two records. This potentially matched record would be classified as a Class 4 match since less than eight of the following criteria are matched (first name, middle initial, last name, birth day, birth month, birth year, sex, race, marital status, or state of birth). In this example, the weights for birth day (4.90), mari tal status (4.08), and state of residence (3.91) would be negative and the score would be 26.36. This is less than the recommended cutoff score of 32.5 for Class 4 and it would be assumed that this is a false match and the person is not deceased.

## Evaluation Studies

The NDI probabilistic scoring system was evaluated using 2 calibration samples. A calibration sample needs to have vital status information such as date and location of death, and ideally, death certificate number on the sample subjects based on sources independent of the NDI. Two NCHS surveys met this criteria: NHANES I Epidemiologic Follow-up Survey (NHEFS) and the Longitudinal Study on Aging (LSOA)

The 14,407 persons who participated in the NHANES I Epidemiologic Follow-up Survey (NHEFS) (1971-75) were used as the first calibration sample. Active follow -up was conducted on this sample to ascertain the vital status of the participants. Death certificates were obtained for persons found to be deceased. The NHANES is a large nationally repres entative survey, and can be used as a calibration sample for developing a methodology for classification of potential NDI matches.

Since the NDI was not begun until 1979, persons who died prior to 1979 were eliminated from further consideration. Vital status was obtained independent of the NDI by interviewer followback in 1982, 1986, and 1987. The NHEFS sample was then matched to the NDI for years 1979 through 1986. This yielded 5,393 records with potential matches to the NDI and 6,672 records not involve d in any matches.

The Longitudinal Study on Aging (LSOA) dataset was used as a second calibration sample. The LSOA was based on a subset of the 1984 NHIS participants. The data used in this calibration sample are those participants aged 70 and over at the time of interview who were followed through August, 1988. Vital status was obtained independent of the NDI by interviewer followback in 1986 and 1988. Of the 7,541 persons originally interviewed in 1984, 3,466 had potential matches with the NDI (1984 -88) and 4,075 persons were not involved in any match.

## Subgroup biases in classification

The results of the evaluation study revealed biases in the classification of NDI match status for females and for non-whites. The correct classification rate for female s, who were known to be deceased, was about 2.5 percentage points poorer than for males (94.0 percent and 96.6 percent, respectively). This is due to linkage problems caused by changing surnames. Even though father's surname is being used to provide additional information there still remain problems of correctly reporting and recording surnames in both the survey and on the death certificates. Both males and females had the same correct classification rates for living persons.

Among non-whites, there are multiple problems, including lower reporting of social security numbers and incorrect spelling/recording of ethnic names, that can lead to underestimated mortality (or incorrectly classifying a true match as a false match). The correct classification rates for known decedents who are non-white dropped to 86 percent (89 percent in the LSOA alone) while the classification rate for living persons remained high at about 97 percent. The classification rate for female non -whites known to be deceased was about three percent lower than the classification rate for non -white male decedents (84.7 percent and 87.8 percent respectively).

Differential reporting of SSN and correct name information results in a relatively large proportion of non - white potential matches classified as Class 4 matches. As discussed earlier, Class 4 consists of records with unknown SSN's and less than eight of the other items matching (due to errors or missing information). NDI users are urged to carefully evaluate the results of class 4 matche s, especially among matches for females and non-whites. Female and non-white matches assigned to Class 1, 2, 3, or 5 appear to have the same correct classification rates as those for white males.

## References


Calle E.E. and Terrell, D.D., Utility of the National Death Index for Ascertainment of Mortality among Cancer Prevention Study II Participants. American Journal of Epidemiology Vol 137, 235 -241, 1993.

Copas, J.B. and Hilton, F.J., Record Linkage: Statistical Models for Matching Computer Records . Journal of the Royal Statistical Society A Vol 153, 287 -320, 1990.

D'Andrea Du Bois, N.S. Jr., A Solution to the Problem of Linking Multivariate Documents. Journal of the American Statistical Association, 163 -174, 1969.

Fellegi, I.P. and Sunter, A.B., A Theory for Record Linkage. Journal of the American Statistical Association, Vol 64, 1183 -1210, 1969.

Keller, J.E., Howe, H.L. and Noak, J.R., An Algorithm for Matching Anonymous Hospital Discharge Records Used in Occupational Disease Surveillance: Anony mous Record Matching Algorithm. American Journal of Industrial Medicine Vol 20, 657 -661, 1991.

Newcombe, H.B., Kennedy, J.M., Axford, S.L., and James, A.P., Automatic Linkage of Vital Records. Science, Vol 130, 954 -959, 1959.

Newcombe, H.B., Fair, M.E., and Lalonde, P., The Use of Names for Linking Personal Records. Journal of the American Statistical Association Vol 87, 1193 -1208, 1992.

Rogot, E., Sorlie, P., and Johnson, N.J., Probabilistic Methods in Matching Census Samples to the National Death Index. Journal of Chronic Diseases Vol 39, 719 -734, 1986.

Tepping, B.J., A Model for Optimum Linkage of Records. Journal of the American Statistical Association Vol 63, 1321 -1332, 1968.

Williams, B.C., Demitrack, L.B., and Fries, B.E., The Accuracy of the Nat ional Death Index when Personal Identifiers other than Social Security Number are Used. American Journal of Public Health Vol 82, 1145 -1147, 1992.

Horm, J.W. and Wright, R.A.: A New National Source of Health and Mortality Information in the United States. Proceedings of the Joint Statistical Meetings, San Francisco, August, 1993. In Press.

National Death Index User's Manual. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statist ics, Hyattsville, Md. September, 1990.

Kovar, M.G., Fitti, J.E., and Chyba, M.M., The Longitudinal Study on Aging: 1984 -90. National Center for Health Statistics. Vital and Health Statistics 1(28). 1992.

Finucane, F.F., Freid, V.M., Madans, J.H., et al. Plan and Operation of the NHANES I Epidemiologic Followup Study, 1986. National Center for Health Statistics. Vital and Health Statistics 1(25). 1990.

厚生労働科学研究費補助金（政策科学総合研究事業（統計情報総合研究事業））
死亡統計データベースの作成とその研究利用のあり方に関する研究

平成20年度　総括・分担研究報告書（平成21年3月）

発行者責任者　研究代表者　安村誠司
発　　行　福島県福島市光が丘1番地
福島県立医科大学医学部公衆衛生学講座
電話　024－547－1180
FAX　024－548－4600