# How Many Digits In A Handshake?
# National Death Index Matching With Less Than Nine Digits of the Social Security Number

Bryan Sayer, Social & Scientific Systems, Inc.
8757 Georgia Ave, STE 1200, Silver Spring MD 20910-3739
E-mail: Bsayer@s-3.com
Christine S. Cox, National Center for Health Statistics
3311 Toledo Rd, Hyattsville MD 20782
E-mail: Ccox@cdc.gov

**Keywords: Record Linkage; National Death Index; National Health Interview Survey; Probabilistic Matching**

**Abbreviations: SSN, Social Security Number; NDI, National Death Index; NHIS, National Health Interview Survey; SSA, Social Security Administration; NCHS, National Center for Health Statistics; NHEFS, NHANES Epidemiological Followup Study; NHANES, National Health and Nutrition Examination Survey**

## Introduction

Survey respondents are increasingly reluctant to provide full SSNs to interviewers for a variety of reasons. For example, in the NHIS for respondents age 18 years and over the refusal rate for SSN has risen steadily from 16% in 1993 to 39% in 2000, while the total missing rate (including refused, don't know and missing) has risen from 28% to 60% over the same time period. SSN is used to select potential matches when linking to administrative files, such as the NDI. Record linkage is a cost effective method of adding a longitudinal mortality component to cross sectional surveys. As a number of surveys at NCHS are periodically linked to the NDI, NCHS would like to increase the number of respondents reporting their SSN, or at least a portion of their SSN. There is some indication that respondents may be willing to disclose a portion of their SSN, such as the last six or last four digits. Because the current NDI record linkage process uses only the full nine digit SSN, we test the impact of using the last six digits of the SSN on matching to the NDI.

## Overview

### Elements of Social Security Numbers

In the U.S., an SSN represents a unique nine digit identification number. The SSN an individual receives is not random. The first three digits (the area number) are based on the zip code to which the card is mailed, while the second two digits (the group number) are assigned as an administrative convenience. The last four digits (the serial number) are essentially random, although numbers ending in '00' are not assigned. A more extensive discussion of the structure of SSN can be found on the SSA's public internet web site, http://www.ssa.gov/history/ssn/geocard.html. A decedent's SSN is included as part of the death certificate registration process and researchers searching the NDI for mortality records can use SSN in the matching process.

### NDI Structure

The NDI is an electronic compilation of all deaths reported from the vital statistics offices of the United States, including the 50 states, the District of Columbia, and the territories beginning with deaths occurring in 1979 (Bilgrad 1997). Researchers desiring to determine vital status and detailed death information (date, cause and state of death) of subjects in research studies may submit identifying information to the NDI for electronic matching. Notable uses include occupational health research, cancer registries, and other studies examining the causes of mortality (Boyle and Decoufle 1990;Calle and Terrell 1993;Curb, Ford, Pressel, Palmer, Babcock and Hawkins 1985;Horm and Wright 1993;Patterson and Bilgrad 1986).

### NDI Selection Process

In the first step of the matching process all NDI records that agree with the submission record on any one of nine selection criteria (Appendix I, Part A) are retrieved from the NDI database. This is the pool of *potential* matches. No identifying information such as the name, date of birth, or SSN from the death certificate is returned to the user. Of the nine selection criteria, only one uses SSN. All records that match exactly on all nine digits of the SSN are selected into the pool of potential matches. Except for errors in the reporting or recording of SSN, these records should be correct matches for decedents since SSN is a unique identifier.

### NDI Scoring Process

Once the pool of potential matches is selected, each record pair (called the combined record in the documentation) is scored using marks of agreement for the twelve items (Appendix I, Part B) in the user's record and probabilistic weights based on the likelihood of the value of the item on the user's record. Items that match get a positive value and items that do not match get a negative value. Missing items (on either the user's record or the NDI record) are zeros. The total score is the sum of the weights for all available items. Each potential match is then classified into one of five classes based on which and how many items match. Based on the fact that a nine digit SSN is a unique identifier, the classification system relies heavily, but not exclusively, on whether the SSN from the user's record matches the SSN on the NDI record. The combination of class and the total score assists users in determining whether a specific combined record represents a true match or not.

## Test Data and Method

We test the impact of matching to the NDI with only the last six digits of the SSN using a sample of 11,804 subjects all with self-reported nine digit SSNs[1], known vital status, and confirmed death certificates collected for the decedents. The test sample is from NHEFS, a longitudinal study that uses as its baseline those adults (25-74) who were initially interviewed as part of NHANES I in 1971-1975. The NHANES I sample is a national probability sample of the U.S. civilian noninstitutionalized population from 1 to 74 years of age. Followup interviews with the original participants were conducted in 1982-84, 1986 (elderly only), 1987 and 1992 (Bilgrad 1997;Cox, Mussolino, Rothwell, Lane, Golden, Madans and Feldman 1997). If a participant is identified as deceased during the interview re-contact, a proxy interview is conducted and a death certificate is collected to determine date and cause of death.

There are 2,991 decedents and 8,813 non-decedents in the test sample of 11,804. For each of the 11,804 people in our test sample, we know a specific end date for the period of observation. For decedents, this is the date of death. For non-decedents, this is the date the participant (or their proxy if the participant was incapacitated), was last interviewed. The time period from 1/1/1979 (the beginning of NDI) to the end date represents the time during which we know accurately the true vital status of the respondents in the test sample.

Using the test sample, we perform a special NDI matching run where we are able to simulate standard NDI selection using the last six digits of each person's SSN instead of all nine digits for criterion 1[2]. All records selected by the remaining eight criteria are also selected. For every record selected we identify which of the nine selection criteria the record meets. Then we determine whether the record would be selected by all nine digits of the SSN, only by the last six digits, by any of the other eight criteria, or any combination of the three.

The issue examined here is what happens when we use only the last six digits of the SSN for the first criterion, as opposed to selection based on exactly matching all nine of the digits of the SSN. We also compare the results of selection based on SSN matching to selection from the other eight criteria. Throughout the remainder of this paper "Other Than SSN" refers to the collective impact of all eight of the selection criteria that do not use SSN.

## Test Results

### Selection

Matching on nine digits of the SSN generally results in the selection of one record for deceased subjects that matches SSN exactly. Using the last six digits results in about four to five potential matches per subject per year searched, although this varies greatly based on the value of the group number[3]. In our test, we have about 765,000 selected records for the 11,804 subjects. Since the last six digits of SSN is a subset of the full nine-digit number, included in the 765,000 are the records we would have selected using all nine digits. In fact, it is possible that we will select records that are correct matches but would be missed because of error in the recording of the first three digits. In addition, many of the records selected

[1] By using a test sample of subjects who have provided an SSN, there is the possibility that we have introduced an anti-conservative bias. In addition to the fact that the data were initially collected 30 years ago before concern developed over identity theft, there is the possibility that individuals willing to disclose their SSN may be more diligent about making sure the rest of the identifying information (name, date of birth) is recorded correctly.

[2] The authors wish to express their gratitude for the help given by Robert Bilgrad and the staff of the NDI in performing this task. This research would not have been possible without their insight and cooperation.

[3] The assignment of group number was designed before computerization when information was maintained on cards which were retrieved by hand. Within each area number, odd numbers from 01 to 09 are used before even ones. For numbers above 10, even numbers are assigned before odd ones, and low numbers before high. In the NDI data, the most common group number is 10 and it occurs 336 times more often than the least common group number, 93.

using SSN are also selected by one or more of the other eight selection criteria. In our test we select 12 additional correct NDI records using the last six digits of SSN that are missed using all nine digits, although ten of these twelve records also match one of the eight criteria that does not use SSN.

The total number of decedents that have their NDI record selected by each of the three methods (regardless of overlap) is shown in table 1. These results indicate that there is little net difference between the various methods, although 6.7% of decedents are missed if SSN is not used at all, showing that SSN does make a unique contribution in the selection process. The individual contributions of each method can be seen more clearly in the Venn diagram (figure 1) which shows both the overlap and unique contribution of using all nine digits, the last six digits, and all other selection methods.

Table 1 - Individual Selection Results For The 2,991 Decedents

| | Nine Digits | Six Digits | No SSN |
|---|---|---|---|
| SSN Selection | 2,702 | 2,714 | N/A |
| Other Selection | 2,790 | 2,790 | 2,790 |
| Net Total Selected | 2,968 | 2,970 | 2,790 |
| Not Selected | 23 | 21 | 201 |

In figure 1 we see that, compared to methods not using the SSN, that using the last six digits of SSN uniquely contributes about 6% of total mortality (N=180, 178 from nine digits plus 2 from six digits only). Methods other than matching on SSN uniquely contribute an additional 9% (N=256), while 84% of decedent records are found by both SSN and other than SSN selection criteria. 1% (N=21) of mortality is missed entirely, suggesting that SSN is incorrectly recorded on either the survey record or the NDI record, and at least one of the elements used other than SSN (name, date of birth) is also incorrectly recorded. In order to select the greatest number of correct NDI records for decedents, it is necessary to use at least part of SSN, in addition to using methods that exclude SSN.

**Evaluating Potential Matches**

The difficulty with using less than all nine digits of the SSN is distinguishing between true matches and false matches, particularly with the large increase in the number of NDI records selected. Not using the first three digits of SSN represents the loss of about 10% of the total information available in the current NDI probabilistic scoring process (based on the average weight)[4], assuming all twelve items are present. Correctly identifying true positives depends upon having sufficient unique information on each record pair to distinguish correct matches from non-matches. This is the principal used by the current scoring algorithm, which has five distinct classes based on which matching items are present *and* which items match on both records. Each class has a separate score cut-off that can be used to separate true matches from false matches. Other tests conducted by the authors indicate that some of the twelve items are more important than others. Three items (state of residence, marital status, and last name for females) can legitimately change over the interval from when the information is collected in a survey to the point of time in the future that an NDI match is performed. Other items such as sex, are not particularly useful for determining matches, but are very useful for determining non-matches (that a record pair agrees on sex is not particularly informative for matches, but a pair that disagrees on sex is not a match, unless there is recording error). The current classification system makes substantial use of SSN and the fact that in total, it represents a unique identifying number. Consequently, the present classification system cannot be used in our test of the last six digits.

However, we can score the resulting matches using similar binit ($\log_2$) weights, incorporating digit specific weights for the six digits of SSN instead of a single weight for the total. Unlike the present system, these weights are based on the specific value of each digit in the SSN, and are developed from the values observed in the total NDI (1979-2000) representing about 48 million records. We graph the cumulative matches and non-matches by total score. From this we can determine the amount of overlap between the matches and non-matches. We use all records for decedents only, in order to keep the scale simple. However, the relationship of non-matching records for decedents is the same as the relationship of non-matching records for non-decedents.

Figure 2 shows the cumulative number of true non-matches (left side) and matches (right side) at each score over the range of total scores. These can also be interpreted as cumulative density functions (1-cdf for non-matches), since the left and right axes are scaled to their respective totals. Included are all records selected for decedents, by six digit SSN plus all other selection criteria. The overlap portion of the two lines, sometimes called the clerical review region, represents the amount of uncertainty of the matches (Fellegi and Sunter 1969). Overall, there is substantial separation of matches and non-matches

[4] Matching weights in the NDI are $\log_2$ (binit) weights based on the inverse of the probability of occurrence. Common items have smaller values and uncommon items have larger values. The first three digits represent about 14.7% of the weight for the most common 12 items and 8.5% for the least common, with 10.1% being the average of the weights.

based on score alone without creating classes similar to the current system. The cross-over point is the score at which the proportion of non-matching records with scores greater than or equal to the cross over score is equal to the proportion of matching records with scores less than or equal to the cross-over score. For this sample of decedents the cross-over score is approximately 15 (the vertical reference line). However the tail on the non-match distribution does run further past the 15 than the tail of the match distribution suggesting that there may be a greater risk of a false match than a false non-match. A false match can occur when record pairs with common values agree by chance rather than belonging to the same person. By using only the last six digits of SSN, we have eliminated the unique identifying property of SSN which is the basis of the current classification system. To fully implement a matching system using only six digits of SSN, a completely new classification system would have to be researched and developed.

## Conclusion

The increasing rate of refusal by survey respondents to disclose their full SSN has led to interest in requesting a partial SSN instead for the purpose of matching to administrative records such as the NDI and Medicare. However it has not been clear whether a partial SSN would be as effective in matching, or even whether it is needed. These results indicate that for matching to the NDI a partial SSN is needed in order to maximize the number of correct matches. Without SSN, 7% of correct matches would be missed under the present procedures. This simulation demonstrates that using the last six digits of SSN would be as effective as using all nine in the selection process. Due to the methods used for distinguishing true matches under the current system, it is not possible to make a direct comparison of the rate of false positives when using the last six digits of SSN. However, the overall separation of matches from non-matches for decedents by score in the data is generally very good, suggesting that it is possible to develop such a classification system.

Another alternative to using the last six digits of SSN is to use the last four digits. Credit card receipts often show only the last four digits of the credit card number, and students are used to seeing the last four digits of their SSN, suggesting that the public might be more willing to disclose four digits to survey interviewers than six or nine digits. However, using only the last four digits would require additional changes to the NDI selection process (which we cannot simulate easily) to limit the number of records selected. These changes would require careful study in order not to reduce the number of correct selections. Whereas using the last six digits of SSN results in four to five NDI records selected per submission record per year searched, using four digits (with no other blocking criteria) would result in selecting about 215 records per submission record per year searched. Since there are only 9,900 valid combinations of the last four digits, a sample of 10,000 subjects could result in every record in the NDI being selected (about 2.2 million per year searched). One alternative to the present system in this case would be to score records as they are selected and only keep those with positive scores, rather than scoring all selected records.

In summary, it appears from this analysis that conducting record linkage with non-unique partial SSNs is a viable alternative to matching with a nine digit unique SSN. Linkage systems that currently rely on nine-digit SSN matching would need to be re-engineered to include additional blocking criteria to remove obvious non-matches based on random agreement on partial SSN. Other research needs to be conducted to ascertain whether reluctant survey respondents are willing to provide partial SSNs in lieu of an outright refusal to provide SSN.

**Figure 1: Correct Certificates by Selection Source**

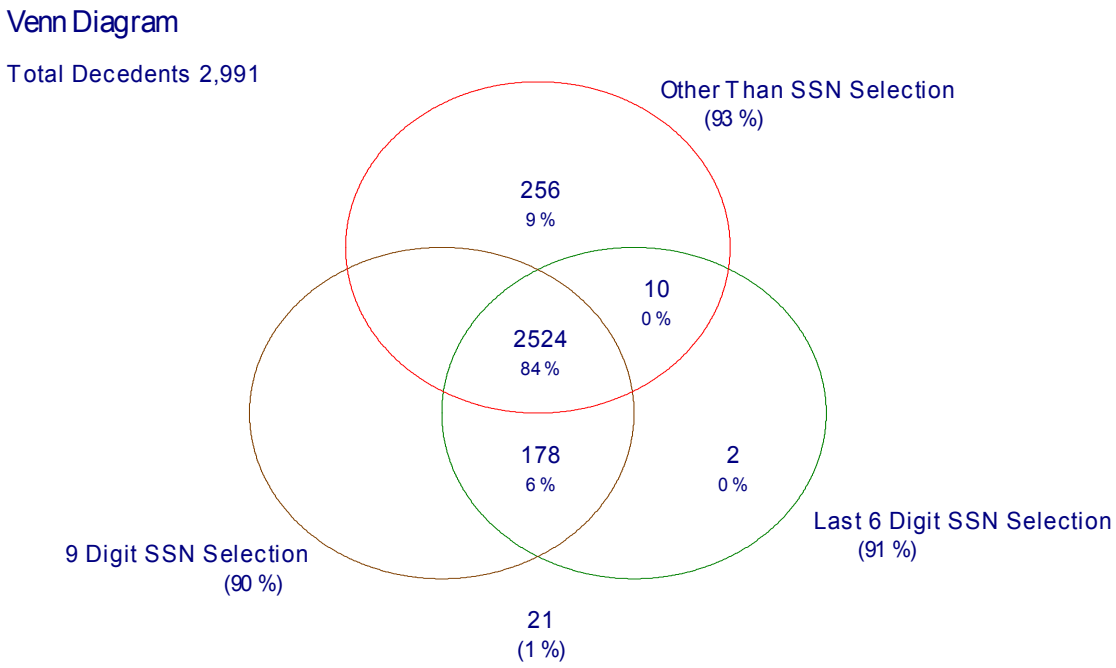
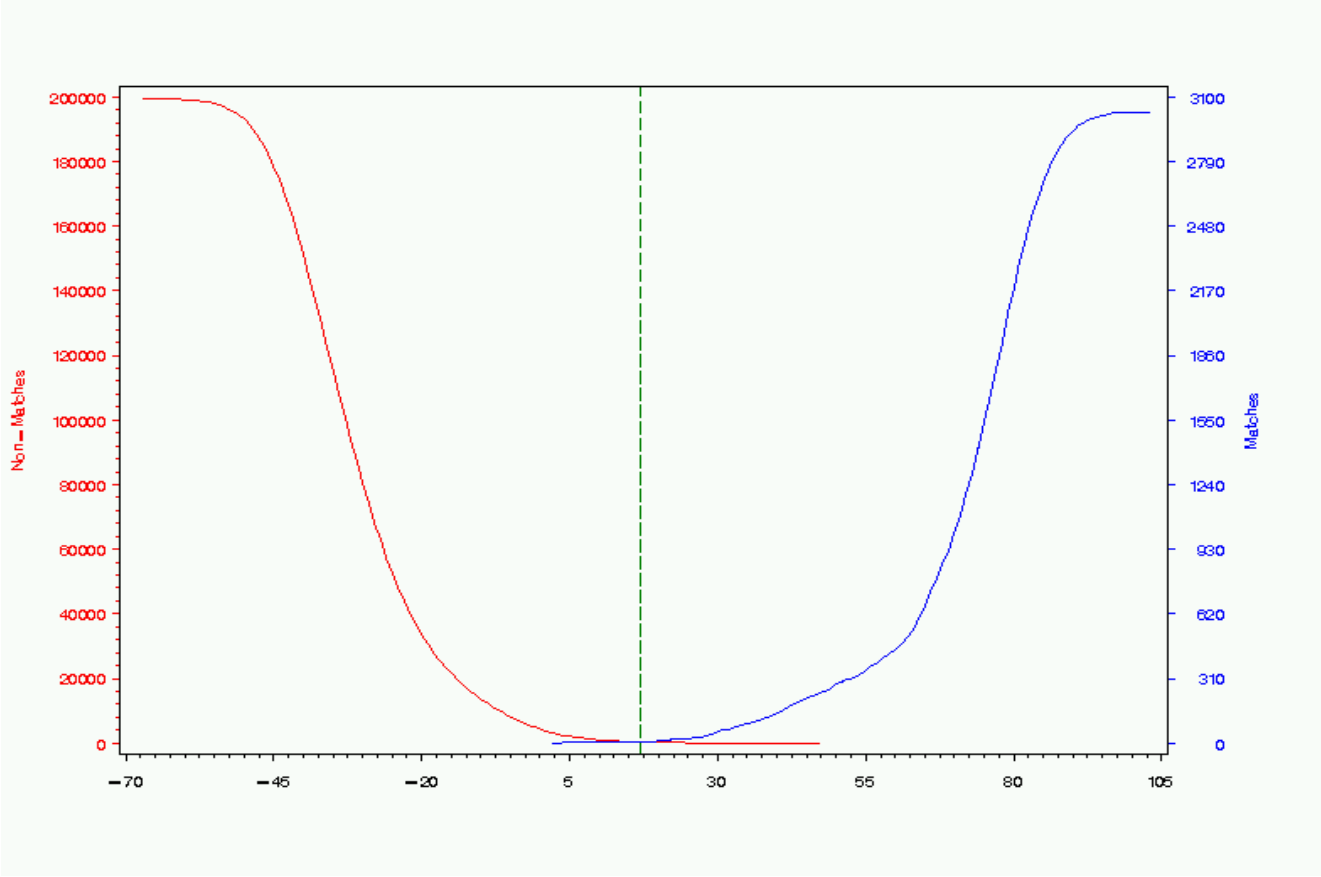



**Figure 2: Cumulative Non-Matches and Matches by Total Score for Decedents**



# Appendix I – Matching Criteria

Part A - Nine Selection Methods

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

Note: All matches on last name and father's surname are performed on the basis of either exact spelling or NYSIIS phonetic codes (New York State Identification and Intelligence Systems).

Part B - Twelve Weighting Items

1. First Name
2. Middle Initial
3. Last Name
4. Social Security Number
5. Month of Birth
6. Day of Birth
7. Year of Birth
8. Race
9. Sex
10. Marital Status
11. State of Birth
12. State of Residence

Reference List

Bilgrad, R. National Death Index User's Manual. 10-1-1997. Hyattsville, Maryland, U.S. DHHS, Centers for Disease Control and Prevention, National Center for Health Statistics.

Boyle, C. A. and Decoufle, P. (1990), "National sources of vital status information: extent of coverage and possible selectivity in reporting," *American Journal of Epidemiology*, 131, 160-168.

Calle, E. E. and Terrell, D. D. (1993), "Utility of the National Death Index for ascertainment of mortality among cancer prevention study II participants," *American Journal of Epidemiology*, 137, 235-241.

Cox, C. S., Mussolino, M. E., Rothwell, S. T., Lane, M. A., Golden, C. D., Madans, J. H., and Feldman, J. J. Plan and Operation of the NHANES 1 Epidemiologic Followup Study, 1992. 35. 1997. Hyattsville, Maryland, U.S. Department of Health and Human Services, CDC-NCHS. Vital and Health Statistics.

Curb, J. D., Ford, C. E., Pressel, S., Palmer, M., Babcock, C., and Hawkins, C. M. (1985), "Ascertainment of vital status through the National Death Index and the Social Security Administration," *American Journal of Epidemiology*, 121, 754-766.

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

Horm, J. W. and Wright, R. A. (1993), "A New National Source of Health and Mortality Information In The United States," in *Proceedings of the Social Statistics Section, 1993 Joint Statistical Meetings*, pp. 256-261.

Patterson, B. H. and Bilgrad, R. (1986), "Use Of The National Death Index In Cancer Studies," *Journal of the National Cancer Institute*, 77, 877-881.