PROBABILITY LINKAGE USING SOCIAL SECURITY ADMINISTRATION FILES

Bert Kestenbaum
Social Security Administration

ABSTRACT

The methods and results of a sophisticated application of probabilistic linkage to Social Security Administration administrative files is presented. The study objectives were to (1) identify persons issued more than one social security number, and (2) to identify twins.

The methodology incorporates the procedures outlined in Newcombe's HANDBOOK OF RECORD LINKAGE. This study is believed to be one of the largest applications of probabilistic linkage.

KEYWORDS

Multiple issuance, Twins, Cohort born in 1919

It gives me great pleasure to appear with colleagues who have contribued so much to record linkage: with Martha Fair, who has been at the forefront of the pioneering efforts of our neighbors to the north to make probabilistic linkage a viable and accessible tool for health research; with Fritz Scheuren, who engineered the very productive exact matching of the March 1973 Current Population Survey with administrative records of the Social Security Administration and the Internal Revenue Service, organized conferences for the sharing of experiences among persons doing record-linkage work and the promotion of good practice of these techniques, and interested me in such work; and to Bill Winkler, who has made important methodological contributions to the theory and practice of record linkage.

This talk will describe both the context and the substance of my experiences with probabilistic record linkage using Social Security Administration administrative files. Particularly because this paper is first in the session, I will begin by discussing the probabilistic method and describing some of its major features. Next I will describe the rather special issues to which I am applying probabilistic record-linkage techniques. Then I will describe the linkage itself, which I believe is one of the largest- ever applications of probabilistic linkage methods, and present some results and observations.

I. THE PROBABILISTIC METHOD

In brief, what is probabilistic linkage? Whenever a record from File A, the search file, is compared to a record from File B, the file being searched, with respect to a set of identifiers selected for linkage purposes, the probabilistic method computes a score which represents the odds that a linkage is a truematch, and pairs with scores high enough are accepted as truematches. More specifically, for each identifier with respect to which a pair of records is compared, one determines the frequency of the obtained outcome of the comparison among truematches with the frequency of the obtained outcome among random pairs, and calculates the ratio of these two frequencies. The product of the frequency ratios calculated for each comparison outcome across the set of identifiers is then the score for this particular record pair.

To illustrate, suppose one knows that if a record from File A and a record from File B form a truematch, then the frequency with which the surnames on the two records are identical is 90%, while if the records do not form a truematch, then the frequency with which the surnames are identical is 0.5%. Then, if a record from File A is compared to a record from File B with respect to surname, and they agree, the frequency ratio is 90/0.5, or 180; if they disagree, the frequency ratio is 10/99.5, or 0.10. If the linkage is based on surname, date of birth, and place of birth, and a pair of records, one from File A and the other from File B, agree on surname, disagree on date of birth, and agree on place of birth; then the score for this pair is the product of the frequency ratio for agreement on surname, the frequency ratio for disagreement on date of birth, and the frequency ratio for agreement on place of birth.

To approximate the frequencies of comparison outcomes among pairs of truematched records, one needs a file of truematches from which the requisite frequencies can be obtained. A similar file of pairs of randomly-linked records for such calculations is also recommended in the more sophisticated linkage applications.

Comparison outcomes need not form a dichotomy -- agreement and disagreement -- but, preferably, one or more levels of partial agreement will be recognized. If surnames, for example, are not identical, perhaps the Soundex phonetic codes are, or perhaps the first three letters agree.

Also, comparisons may be made on either a 'global' basis or -- preferably -- on a value-specific basis. If the surname on the record from File A is an uncommon one, like 'Kestenbaum', one would prefer, rather than the frequency among random pairs of overall agreement on surname, the much smaller frequency of agreement among random pairs on surname when one surname is 'Kestenbaum'. For the surname 'Smith', on the other hand, the value-specific frequency of agreement among random pairs will exceed the global frequency of agreement. The usual practice is to prepare lookup tables listing the specific values of identifiers with their frequency of occurrence in File B.

Now, if one had to compare every record in File A with every record in File B, record linkage would be impracticably time-consuming except for very small files. Therefore, it is necessary, in general, to partition or 'block' both files and limit the attempted linkages to records from A and B in the same block. To attempt linkages with records which fall in different blocks under one blocking strategy, one could repeat the linkage using alternative blocking strategies.

In the preprocessing stage steps are taken to facilitate the subsequent linkage. For example, names of persons and places may be standardized with respect to abbreviation and punctuation. Also, if a given identifier appears in more than one way, multiple records may be generated corresponding to the multiple ways.

The technique of probabilistic record linkage is not new; indeed the Canadians have had in place a sophisticated linkage system for more than a decade. Nonetheless, the technique has recently achieved greater accessibility and prominence. Howard Newcombe, whose name is associated with the practice of probabilistic linkage more than any other, published a "how to" text in 1988 (Newcombe, 1988) and, together with Martha Fair and a second coauthor, a review article in the Journal of the American Statistical Association in 1992 (Newcombe, et al, 1992). The National Highway Traffic Safety Administration used Matt Jaro's AUTOMATCH software to carry out a legislatively-mandated study of the benefits of safety belt use and motorcycle helmet use among crash victims (National Highway Traffic Safety Administration, 1995), and the National Center for

Health Statistics is experimenting with conducting searches to its National Death Index on a probabilistic basis (Bilgrad, 1995).

II. **The 1919 cohort**

Several years ago I undertook a study of how persons born in the United States in 1919 fared with respect to the social security program. I am, in fact, of the opinion that an elaborate demographic profile of a birth cohort, viewing from a longitudinal perspective its mortality and morbidity, fertility and mobility, labor force participation and career earnings, similar to cross-sectional Bureau profiles in Current Population Reports of special groups such as women and minorities, would both be of substantive interest and provide opportunity for challenging methodological work.

Let's begin by estimating the number of persons born in the United States in 1919 who survived to November 1936, when the brand-new Social Security Board, part of President Roosevelt's New Deal, began issuing social security numbers. Also, let's make a very rough estimate of the number of pairs of twins born here in 1919 surviving intact to that date.

The official estimate of 1919 births is Pascal Whelpton's estimate of 2,740,000 (National Center for Health Statistics, 1994, table 1-1). Applying to this estimate first infant mortality rates for that year and then mortality rates for children from appropriate decennial life tables, all on a sex-and-race-specific basis, we arrive at an estimate of 2,365,000.

Extrapolating backwards from a historical series of twinning rates prepared by Robert Heuser which begins in 1922 (Heuser, 1967), we estimate a twinning rate of 12 per thousand for 1919, hence 33 thousand twin births. Infant mortality among twins in 1919 cannot be estimated precisely; however, using as guides the earliest measurement of twin infant mortality in the U.S. in 1960 (Kleinman, et al, 1991, table 2) and data for developing countries collected in the World Fertility Survey (Rutstein, 1984), a rough but reasonable estimate of infant mortality among twins born in 1919 is 250 per thousand. Mortality among twins is estimated to be slightly higher than among singletons between ages 1 and 5, and equal to singleton mortality thereafter (Rutstein, 1984; Bulmer, 1970). Finally, a factor is required to recognize the concordance for survival in members of a twin pair (Jablon, et al, 1967; Gittelsohn and Milham, 1965, table 3). We arrive at an estimate of 19 thousand 1919 twin pairs surviving intact to November 1936.

A question of fundamental importance in social insurance policy, one which is often addressed to my organization, the Office of the Actuary in the Social Security Administration, concerns the extent to which people receive program benefits at least equal to the taxes they've paid during their worklives -- the "money's-worth" issue. In fact, just last week the Subcommittee on Social Security of the Senate Finance Committee held hearings, at which our Commissioner gave testimony, on this issue. We are constrained to frame a reply in terms of a hypothetical typical worker: a worker with typical life and worklife spans and a typical earnings history with a typical family will pay x dollars in taxes and receive y dollars in benefits. I wanted to address the money's-worth issue with the distribution of experiences of a cohort of real people. I chose the cohort born in 1919 for this project, and a sample design which selected a 1-in-100 sample of members of this cohort, the selection based on patterns of digits in the social security number.

The first question considered was, what fraction of persons born in the United States in 1919 and surviving to the inception of the Social Security program in mid-November 1936 had any interaction

with the program, if not as taxpayers or beneficiaries, at least as issuees of a social security number.

We have earlier determined the denominator of this fraction to be 2,365,000. The numerator, the number with some connection to the program, ought to be derivable from the count of social security numbers issued to cohort members, after taking account of the information in agency records that, because persons have sometimes been issued more than one number, the count of numbers issued to the 1919 cohort exceeds the count of cohort members by about 4 percent. An agency task force report concluded that the extent of multiple issuance of which the agency is unaware is likely to be relatively small (Social Security Administration, 1971). The numerator derived by this count and adjustment was 2,530,000, however, about 7 percent bigger than the denominator!

Part of the explanation for this impossible result, we think, lies in an understatement of the denominator. The official estimate of U.S. births in 1919 was estimated by Pascal Whelpton by multiplying the number of births in the Birth Registration Area, after adjustment for underregistration, by the ratio of infants in the U.S. counted in the 1920 census to the number counted in the census in those States comprising the Registration Area. Now, the 1920 census, taken on January 1, the only census in our history taken in the winter, is recognized to be a particularly deficient census. It is reasonable to surmise that the deficiency would be greater in States outside the Registration Area, which tended to be less urbanized and less developed; if true, the ratio calculated by Whelpton is too low.

To assess this hypothesized downwards bias in Whelpton's estimate, we tabulated a public-use file of the 1980 census to produce a distribution by State of birth of persons born in 1919. The value of Whelpton's ratio was indeed moderately higher, as hypothesized.

Furthermore, decennial life tables are derived from counts of deaths unadjusted for underregistration and counts of population unadjusted for underenumeration, that is, under the implicit assumption that the relative underregistration of deaths and the relative underenumeration of population are equal. If this assumption holds for the typical census, it cannot hold for the atypical 1920 census, leading to the conclusion that probabilities of death in the 1919-21 decennial tables are too high, and more 1919 births survived to 1936 than we calculated.

Although we conclude that our earlier estimate of the denominator, the number of persons born here in 1919 surviving to 1936, is too low, we estimate the magnitude of the understatement to be only about 50 thousand. We agree with Gregg Robinson of the Bureau that the larger problem is the overstatement of the numerator, probably because yet-undetected multiple issuance of social security numbers is a problem of substantial magnitude, contrary to the reassurances of the agency task force report. There is evidence, in fact, that in the early days of the program many persons misunderstood and thought that a job change to a new employer required a new social security number. Also, I have seen number applications submitted in 1937 by <u>employers</u> on behalf of persons no longer in their employ so that these persons' earnings could be posted to some account.

The multiple issuance phenomenon does not really create a problem for our sample design. We've adopted the strategy to identify a person issued more than one social security number with his "main" number. For example, if he has received benefits under one number, we choose that one; if he hasn't received benefits, but paid taxes for 25 years on one account and for 2 years on the other account, we'll use the account number with the 25 years of payments.

On the other hand, the phenomenon of undetected multiple issuance does present a problem, indeed a twofold one. That is, we need to identify those accounts in our 1-in-100 sample belonging to persons issued another number which is their main number; these accounts should be deleted from our 1-in-100 sample. We also need to identify those accounts outside our 1-in-100 sample belonging to persons also issued a 1-in-100 number which is their main number; these persons remain in our sample, but, information from their other accounts should be added.

III Hidden multiple issuance

Let me at this point briefly describe two of the major administrative files of the Social Security Administration. One is the NUMIDENT file of initial applications for a social security number and card, reapplications to replace lost cards or to correct identifying information (such as changes in surname upon marriage), and certain records of claims to benefits. The NUMIDENT file is also used to house death information which became available to the Social Security Administration. Census Bureau staff, particularly in the Population Division, have used this file. The NUMIDENT record has extensive personal information: name, name at birth, date of birth, place of birth, sex, race, parents' names, and date of death, if known deceased.

Another major administrative file is the Master Beneficiary Record, the file of entitlements to retirement, survivorship, and disability program benefits. This file is similar to, but more inclusive than, the Medicare enrollment file which Population Division staff use in their work on measuring undercount through demographic analysis and in intercensal and postcensal estimates of migration among the elderly. The personal information in this file, which includes name, date of birth, sex, race, and date of death, if deceased, is not quite as extensive as in the NUMIDENT.

The idea of probabilistic matching was not our first idea for detecting multiple issuance to members of the 1919 cohort. Our first initiative was to exact-match numberholders in the 1-in-100 sample who had not applied for benefits with a file of 2 million MBR records not in the 1-in-100 sample with year-of-birth of 1919; the 144 numberholders (representing 14,400 cases) successfully matched were dropped from the sample.

The next initiative was to select suspect numberholders, in the sense that the numberholder had not applied for benefits through 1992 (73 years after birth), was not working, and was not known from our databases to be deceased. There is in place a NUMIDENT query system, by which one manually keypunches a name and date of birth (with certain tolerances) and receives a printout of matching NUMIDENT records. Using this manual facility, we discovered 325 multiple numberholders which could be purged because a newly-discovered number outside the 1-in-100 sample was their main number, as well as 45 multiple numberholders whose newly-discovered number was less active, but whose activity on this number will be included for describing their program experience.

Next we obtained a tape file of 6 million NUMIDENT records with year-of-birth of 1919, sorted them according to the set of values for certain combinations of identifiers, and identified and examined pairs of records, one of which was in the 1-in-100 sample the other of which was not, matching exactly on the combination of identifiers. This process netted 350 multiple numberholders, resulting in 127 purgings and 223 instances of combining experiences.

At this point we turned to probabilistic matching of the combined NUMIDENT and Master

Beneficiary Record files. Please note that the modest success from probabilistic linkage which we will report would have been more impressive had we not previously uncovered several hundred truematches.

The technique of probabilistic matching is especially valuable when the identifiers used in the linkage are prone to change, as is the case for surnames of females, and when the identifiers are frequently missing or in error. The NUMIDENT, in particular, has serious data quality shortcomings, largely for two reasons. First, the creation of the NUMIDENT in the mid-1970's constituted an enormous data entry undertaking stretching over several years with the keying-in of information from hundreds of millions of paper forms; the sheer magnitude of the operation made conditions less than conducive to quality control. Second, it was the practice in the old paper environment, that in the adjudication of claims for benefits, the account number application form would be physically removed from the file to become part of the packet of documents provided to the adjudicator, with its place taken in the file by a claims form which generally lacked much of the information on the original form. Often even the sex code was missing, causing us to depart from the standard practice of splitting the linkage application into two, one for males, the other for females. Separate-sex treatment cuts down on the size of the linkage, as well as allows for disagreement on surname to be of less importance for females than for males, as it should be.

For my application I used all these identifiers: month of birth, day of birth, surname (including surname at birth), given name, middle name, sex, race, first 5 digits of account number (which reflect geographic and temporal information as of the time of issue), State and place of birth, vital status and date of death, mother's maiden name, mother's given name, mother's middle initial, father's given name, and father's middle initial. For each social security number, I produced a record for every unique combination of these identifiers: beginning with approximately 6 million NUMIDENT records and approximately 2 million Master Beneficiary Records for persons born in 1919, close to 30 3/4 million unique records were produced for linkage.

In the preprocessing stage we expended considerable effort on standardizing place names, at least more common ones: for example, before standardizing, Brooklyn, NY could appear as BROOKLYN, BKLYN, or KINGS (COUNTY). Another interesting preprocessing initiative was to attempt to identify which female middle names were actually maiden names, by comparing to a list of common female given names.

Records with social security numbers in the 1-in-100 sample constituted file A; all others became File B. File B contained about 30 1/2 million records, with the remaining 1/4 million going to File A. I believe that this may be one of the largest ever applications of probabilistic linkage.

Record pairs which were already identified in agency files as instances of multiple issue formed our reference file for calclating the frequencies of comparison outcomes among truematches. Using random numbers we fashioned a very large file of randomly linked record pairs to provide frequencies of comparison outcomes among random pairs.

Comparisons were made first on a global basis, and then -- for pairs scoring above a chosen threshold on the global basis -- on the more discriminating value-specific basis, as well. Partial agreements were recognized to a very large extent. The linkage was repeated under eight blocking strategies.

To make the transition from global to specific, lookup tables were constructed containing the frequencies of identifier values in File B. I had an idea, which is not found in Newcombe's book, to speed up the processing by eliminating from the lookup table any values which do not appear in File A. With respect to given names, for example, the lookup table was thereby reduced from about 135 thousand entries to less than 5 thousand.

Probabilistic linkage allowed us to purge another 227 numberholders from our sample because they were persons issued multiple numbers whose main number was not their sample number, and to identify 371 accounts outside the sample belonging to multiple numberholders whose 1-in-100 sample number was their main one.

I was pleased with the power of probability linkage to uncover duplication which escaped detection under other scrutinies. Of course, while we have made progress in reducing the excess in the count of persons born in the United States in 1919 who were issued social security numbers relative to the count of eligibles calculated from vital statistics and census data, there remains an excess which still must be dealt with: the numerator is currently 2,453,000, and the denominator 2,415,000.

IV Twins

Many pairs scored very high in the probability linkage but had different given names: these were often twins, who, of course, would have in common date of birth, surname, parents names', and several other identifiers. Unless the two numberholders in the pair were of different sexes, or one was alive and the other deceased, it was sometimes difficult to judge whether the two were twins or whether they were the same person. I assumed that the pair were twins if the given names were quite dissimilar; otherwise I looked at other information, such as the lifetime earnings patterns on the two accounts, to arrive at a decision.

The probability linkage yielded 324 sets of twins in which one of the twins had a number in the 1-in-100 sample and the other did not. By inference, an estimated 16,400 twin pairs born in the United States in 1919 and issued social security numbers beginning in late 1936 can be identified through probabilistic linkage, since almost 1 in 50 (actually 0.0198) sets of twins are expected to have exactly one number in the 1-in-100 sample. Please recall that we had earlier estimated the number of twin pairs born in the U.S. in 1919 surviving intact to 1936 to be 19 thousand.

It should also be borne in mind that the strategy here for probability linkage was far from optimal for the objective of finding twins. Because the focus was on associating records belonging to one person, comparison disagreements on given name, on sex, and on vital status/date of death argued against a match. Such disagreements have different significance when searching for twins, of course.

I also determined that, by comparison, exact linkage in the file of 30 3/4 million records on a combination of date of birth, surname (Soundex), mother's maiden name (Soundex), mother's given name, and father's given name yielded 172 twin sets in the 1919 U.S. birth cohort with one number in the 1-in-100 sample and the other not. Thus the probabilistic method in this application uncovered nearly twice as many twins as an exact linkage method.

The international leaders in twin research are small countries, particularly in Scandinavia, which maintain population registers that are used both to identify twins and to contact them to inquire

about their experiences. For example, the twin registry in Finland was compiled from its central population registry by computer-matching records with common date of birth, surname at birth, place of birth, and sex (Kaprio, et al, 1981). The computerized linkage is typically limited to comparing each record with the one following after the file has been sequenced according to the combined values of the identifiers used in linkage. Our results suggest that computerized linkages founded on a probabilistic basis could be significantly more successful in finding twins.

As for the United States, the NUMIDENT is the closest thing we have to a central population register, and while it does not contain address information, the social security number can be used to link to certain files which do have address information to enable contact with twins, or to track the experience of twins, without respondent burden, in administrative files which use the social security number. Had the NUMIDENT ever been considered for the support of U.S. twin research?

In fact, a team of researchers recently undertook a major project, the Black Elderly Twin Study, or BETS, which begins with the compilation of a twin registry of older persons of both sexes and all races -- earlier U.S. twin registries were restricted, completely or almost completely, to white males (Science, 1993). The ambitious initial plan for BETS was to assemble twin pairs by computer-linking NUMIDENT records, and then to obtain current address information from other administrative files which use the social security number.

However, the Social Security Administration refused the researchers access to the NUMIDENT because of privacy concerns. With this avenue closed, the researchers contracted with the Health Care Financing Administration to assemble linked pairs and provide current addresses from its file of elderly persons enrolled in Medicare, though the Medicare file is much less rich in personal identifiers than the NUMIDENT. The charter of the Health Care Financing Administration specifically allows that its data be made available for health-related research. It is, however, ironic that some of the data the Health Care Financing Administration makes available are data it received from the Social Security Administration.

The Medicare enrollment file was indeed processed to produce male-male pairs with common date of birth, surname, race, and State in which the application for a number was made. For females, surname is of little utility in record linkage; accordingly, females were paired if they had in common date of birth, race, and the first 7 digits of the social security number. In my opinion, the algorithm for females, which selects only twins who applied for a social security number in the same place and at about the same time, should be evaluated for a potential bias towards selecting twins whose members are more alike one another than other twins.

The algorithm for males produces a preponderance of "false positives": a pilot study indicated that there were more than 8 false positives for every truematch. The researchers are understandably interested in evaluating correlates of truematching, such as uncommonness of surname, similarity of given names, and similarity of account numbers, to enable a more efficient targetting of their mailing to candidate pairs likely to be twins.

The privilege of access to the NUMIDENT which Social Security Administration employees enjoy allows me to assist the research team in two ways. First, if the team submits a sample of candidate pairs for a check to the NUMIDENT, in identifying the factors which correlate with truematching. Second, in exposing possible biases in the team's study design, through a comparison of the characteristics of twins identified in its study with the characteristics of twins identified through

NUMIDENT.

V In conclusion

Probabilistic linkage appears to be an idea whose time has come. Results like those reported here illustrate its power and efficacy.

REFERENCES

Bilgrad, Robert. 1995. Personal Communication.

Bulmer, M.G. 1970. The Biology of Twinning in Man. Oxford: Clarendon.

Gittelsohn, Alan M. and Samuel Milham, Jr. 1965. "Observations on Twinning in New York State." British Journal of Preventive and Social Medicine 19:8-17.

Heuser, Robert L. 1967. Multiple Births: United States - 1964. Vital and Health Statistics series 21 no. 14. Washington: Public Health Service.

Jablon, Seymour, James V. Neel, Henry Gershowitz, and Glenn F. Atkinson. 1967. "The NAS-NRC Twin Panel: Methods of Construction of the Panel, Zygosity Diagnosis, and Proposed Use." American Journal of Human Genetics 19:133-61.

Kaprio, Jaakko, Markku Koskenvuo, Seppo Sarna, and Ilari Rantasalo. 1981. "The Finnish Twin Registry: A Preliminary Report". In Sarnoff A. Mednick, Andre E. Baert, and Barbara P. Bachmann, Prospective Longitudinal Research: An Empirical Basis for the Primary Prevention of Psychosocial Disorders. New York: Oxford.

Kleinman, Joel C., Mary Glenn Fowler, and Samuel S. Kessel. 1991. "Comparison of Infant Mortality among Twins and Singletons: United States, 1960 and 1983." American Journal of Epidemiology 133:133-43.

National Center for Health Statistics. 1994. Vital Statistics of the United States, 1990, volume I -- Natality. Washington: Public Health Service.

National Highway Traffic Safety Administration. 1995. "Crash Outcome Data Evaluation System (CODES)." Technical Report.

Newcombe, Howard B., Martha E. Fair, and Pierre Lalonde. 1992. "The Use of Names for Linking Personal Records." Journal of the American Statistical Association 87:1193-203.

Newcombe, Howard B. 1988. Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business. Oxford Medical Publications. New York: Oxford.

Rutstein, Shea Oscar. 1984. "Infant and Child Mortality: Levels, Trends, and Demographic Differentials." Revised edition. WFS Comparison Studies no. 43. Voorburg, Netherlands: International Statistical Institute.

Science. 1993. "Elderly Twin Registry in the Works." 260:1239.

Social Security Administration. 1971. Social Security Number Task Force: Report to the Commissioner. Baltimore: Social Security Administration.