Identifying a Cohort of US Twins Using Social Security Administration Records

Bert Kestenbaum

Social Security Administration, Office of the Chief Actuary, Baltimore, Maryland, USA

A sample of pairs of twins who were born in the United States in 1919 and survived to adulthood is identified through an innovative and large-scale application of the methodology of probabilistic linkage. The social security program began in the United States in November 1936, and the file of applicants for a social security number — which was used in this study — is the closest thing in the United States to a population register. The study results are very satisfactory, and demonstrate the superiority of probabilistic linkage to exact linkage. We estimate that about 33,000 twin pairs were born in the United States in 1919 and about 19,000 survived to age 17. Since the social security number was not then, at the inception of the program, the universal identifier that it is today, the number of *enumerated* twin pairs is somewhat less. Nonetheless, over 16,000 twin pairs can be identified by the method of probabilistic linkage. By comparison, only about half as many can be identified by straightforward exact linkage.

A large project was begun some time ago to describe the demography and social experience of the cohort born in 1919 in the United States (US), based mostly on person records of the Social Security Administration and other federal government agencies (Kestenbaum, 1988). The project evolved from a much more narrowly-focused investigation which is of significant interest to the Social Security Administration, namely how people fared, in terms of taxes paid and benefits received, under the program which it administers: who fares best and who fares worst. The broadened scope, however, includes health and living arrangements of cohort members in old age, their mortality, childbearing, educational attainment, lifetime earnings, geographic mobility, et cetera.

Given the fascination, both popular and scientific, of the extent of similarity in the experiences of adult twins, we thought it would be an appropriate subject to include in the project. The preliminary question, however, is whether a relatively large number of adult twins can be identified in government records, specifically Social Security Administration records. The question is answered in the affirmative by the present study, in which personal information for a 1 in 100 sample of persons born in the US in 1919 was compared to personal information for all other persons born in 1919, in a search for twins.

For several of the phenomena addressed in the project, the dataset for analysis is limited to Social Security Administration records for only a 1 in 100 sample of enumerated persons born in the US in 1919. A large clerical effort was needed to retrieve information which was not available electronically, rather only on microfiche, as well as to straighten out situations where ostensibly more than one social security number belonged to one person or one number belonged to more than one person. Such practical considerations explain the limitation to a 1 in 100 sample, roughly 25,000 records.

The search for twins is carried out with a methodology known as "probabilistic linkage", and the success of the search derives from the efficacy of this tool. The "twin" objectives of this paper are to measure the success of the search for twins and to describe this new methodology.

Materials and Methods

An Estimate of the Number of Twins

First, how many twin pairs born in the US in 1919 survived to November 1936, when the brand-new Social Security Administration, part of President Roosevelt's New Deal, began issuing account numbers?

There exists for the US an historical series of twinning rates which begins in 1922 (Heuser, 1967); extrapolating the series backwards 3 years produces an estimated twinning rate of 12 per thousand for 1919. Given that there were approximately 2.75 million births in the United States in 1919 (National Center for Health Statistics, 1994, Table 1-1), about 33,000 pairs of twins were born that year.

Cohort life tables (Social Security Administration, 1992) provide the cohort's age-specific mortality from

Received 27 November, 2003; accepted 9 December, 2003.

Address for correspondence: Bert Kestenbaum, Social Security Administration, Office of the Chief Actuary, Room 760, Altmeyer Building, 6401 Security Boulevard, Baltimore, MD 21235, USA. Email: Bert.M.Kestenbaum@ssa.gov

https://doi.org/10.1375/twin.7.2.134 Published online by Cambridge University Press

birth to age 17, denoted by ${}_{17}q_0$, but care must be taken not to repeat the mistake of calculating the expected number of twin pairs surviving to 1936 as $33{,}000*(1-{}_{17}q_0)^2$. The authors of the paper describing the establishment of the United States World War II veterans twin registry (Jablon et al., 1967), who used this simplification, expressed disappointment at finding many fewer twins than expected, as well as surprise at the tendency to find either both or neither of the twins, a tendency which they attributed to the supposed greater diligence of clerks in finding the second member of a pair when the first had already been found.

As was recognized later, (1) during infancy, particularly the neonatal period, the mortality of twins is significantly greater than the mortality of singletons, and (2) there is concordance in survival between twins in a pair. Thus, following Gittlesohn and Milham (1964), the expected number of pairs of twins among those born in 1919 who survive to age 17 should be calculated as $33{,}000*\{(1-{}_{17}Q_0)^2 + (1-{}_{17}Q_0)*{}_{17}Q_0*r\}$, where "Q" denotes twin mortality and "r" the correlation coefficient.

Unfortunately, the level of twin infant mortality in the US in 1919 is not known and must be approximated. The earliest measurements in the US are for 1960: infant mortality rates for Whites of 105 (per thousand) for twins compared to 20 for singletons, and for blacks of 150 for twins compared to 39 for singletons (Kleinman et al., 1991, Table 2). These figures, together with data for developing countries collected in the World Fertility Survey (Rutstein, 1984), suggest that for 1919, when US infant mortality stood near 80 per thousand, a crude but reasonable estimate of infant mortality among twins would be 250 per thousand.

There is conflicting evidence as to what extent the unfavorable mortality of twins relative to singletons continues beyond infancy (Bulmer, 1970; Rutstein, 1984). Seeking a middle ground, we assumed a 40% excess at age 1, 30% at age 2, 20% at age 3, 10% at age 4, and nondifferential mortality at ages 5 and above. When combined with the infant mortality rate of 250 per thousand, a value for ${}_{17}Q_0$ of 300 per thousand is the result.

A value for "r" of about 0.4 is consistent with the literature (Gittelsohn & Milham, 1965, Table 3; Jablon et al., 1967). Thus the estimated probability of a twin pair born in 1919 surviving to 1936 is: $([0.7]^2 + [0.7*0.3*0.4])$, or 0.574, and the expected number of surviving pairs is close to 19,000.

This estimate appears to be inconsistent with the Danish experience that only about 40% of like-sex twin pairs born between 1910 and 1930 remained intact to age 15 (Hougaard et al., 1992). However, part of the discrepancy is explained by the fact that like-sex twins, about half of whom are monozygotic twins, have higher mortality than unlike-sex twins (Botting et al., 1987; Fowler et al., 1991).

The Compilation of Twin Registries

Twin registries are compiled in several ways. First, twins may be solicited in an advertisement to come forward. Second, birth records may contain information on plurality and be sequenced so as to permit the identification of co-twins; the aforementioned US World War II veteran twin registry is an example. A third method uses *family*-based record systems which contain date of birth information for each family member. Although not the subject of this paper, we expect to soon compile a registry of twins ever entitled to social security child benefits as survivors of a deceased worker or dependants of a disabled or retired worker. This should be a straightforward exercise, because all survivor and dependant records are housed under the worker's account number and contain the beneficiary's date of birth.

Fourth, twin pairs can be assembled from records in an *individual*-based record system containing sufficient personal data, by computer-matching a twin with his/her co-twin. For example, the twin registry in Finland was compiled from its central population register 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 following one after the file has been sequenced according to the combined values of the identifiers used in linkage.

Once the registry is in place, the experiences of twins may be inquired after by questionnaire, if the twins' addresses are known, as they will be in a population register country. Alternatively, in countries where residents are assigned identification numbers, the experience of twins may be ascertained, without respondent burden, from administrative records in which their numbers appear, *by those with access to the records*. This qualification is very significant in a country like the US with powerful confidentiality strictures on access to microdata and a decentralized federal statistical system.

The foremost general twin registry in the US, the registry of World War II veterans, was mentioned earlier. The mortality and morbidity of twins in this registry can be tracked, using their military service identification numbers, in the administrative records of the Veterans Administration. They are also contacted by questionnaire periodically for important scientific inquiries; one example was a combined study of cancer, stroke, Parkinson's disease, and Alzheimer's disease (Elderly twin registry, 1993). This registry is limited to white males.

Not long ago a team of US researchers developed a plan to compile a twin registry of older persons of both sexes and all races (Aging twins offer, 1993). The plan called for assembling twin pairs by computer-linking individual records in the Social Security Administration's "NUMIDENT" file of applicants for an account number, and then for obtaining address information from another federal agency, the Centers for Medicare and Medicaid Services (then the Health Care Financing Administration), which maintains a file

https://doi.org/10.1375/twin.7.2.134 Published online by Cambridge University Press

of all enrollees in the federal health insurance program for the elderly, Medicare. A complete NUMIDENT record has extensive personal information: name, name at birth, date of birth, place of birth, parents' names, sex, and (if known deceased) date of death. However, the Social Security Administration refused the researchers access to the NUMIDENT because of confidentiality concerns with records leaving the agency, and only agency personnel may work with NUMIDENT records.

Probabilistic Linkage

An objective of this paper is to describe the practice of probabilistic linkage, which we believe can be of significant value in the compilation of twin registries from individual-based record systems, and our application of the methodology to the search for twins. Although the technique is not new — indeed the Canadians, who are at the forefront in applying this methodology, have had in place a sophisticated linkage system for more than a decade — it 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 "cookbook"-type text in 1988 (Newcombe, 1988) and, together with two co-authors, a review article in 1992 in the foremost American statistical journal (Newcombe et al., 1992).

The technique is especially valuable when the identifiers used in the linkage are prone to change, as is true 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-1970s was an enormous data entry undertaking over several years with the keying-in of information from hundreds of millions of paper forms, and apparently the sheer magnitude of the operation was not conducive to adherence to strict quality control. Second, it was the practice in the old paper environment, that in the adjudication of claims to retirement and survivor benefits, the account number application form would be physically removed from the file to become part of the package of documents for the adjudicator, with its place in the file taken by a more brief claims form from which was missing much of the information on the original form.

In brief, what is probabilistic linkage? Whenever a record from File A is compared to a record from File B with respect to the set of identifiers chosen for linkage, the probabilistic method computes a score which represents the odds that the linkage is a match, and pairs with high enough scores are accepted as matches. More specifically, for each identifier compared on the two records, the user determines the frequency of the outcome of the comparison among matches together with the frequency of the outcome among random pairs, and calculates the ratio of these frequencies. The product of the frequency ratios calculated for each comparison outcome is the score for the particular linkage.

To illustrate, suppose one has computed that if two records belong to one individual then the frequency with which the surnames on the two records are identical is 90%, while if two records belong to two individuals the frequency with which the surnames are identical is 0.5%. If a record from File A is compared to a record from File B with respect to surname, and they agree, then the frequency ratio is 90/0.5, or 180; if they disagree, then 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.

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 unusual one like "Rumplestilskin", one would prefer, rather than the overall frequency among random pairs of agreement on surname, the much smaller frequency among random pairs of agreement on surname when the surname on one record is "Rumplestilskin". For the surname "Smith", on the other hand, the value-specific agreement frequency among random pairs will be larger than the global agreement frequency.

Also, comparison outcomes need not be limited to a dichotomy — agreement and disagreement — but, preferably, one or more levels of partial agreement may be recognized. If surnames, for example, are not identical, perhaps the Soundex phonetic codes are, or perhaps the first three letters are.

If one had to compare every record in File A with every record in File B, linkage would be impracticably time-consuming except for very small files. It is therefore necessary to "block" the files and limit the linkages to records from A and B in the same block. To link records which occupy different blocks under one blocking strategy, one could repeat the linkage using other blocking strategies.

Results

In our application, File A was the set of records for persons born in the United States in 1919 with account numbers falling in the 1 in 100 sample, and File B the set of all other records for persons born in 1919. Comparisons were made first on a global basis, and then — for pairs scoring above some threshold on the global basis — on the more discriminating value-specific basis, as well. Partial agreements were recognized to a large extent. The linkage was repeated under eight blocking strategies.

We used records for persons born in 1919 from both the NUMIDENT and another Social Security

https://doi.org/10.1375/twin.7.2.134 Published online by Cambridge University Press

Administration file, the Master Beneficiary Record of persons ever entitled to program benefits. Because a NUMIDENT record is created not only for an initial application for a social security card, but also for an application for a replacement either for a lost card or to reflect a change of identifying information such as a change in surname upon marriage, for a death, and for certain claims to social security benefits, many persons have several NUMIDENT records.

These identifiers were used in the linkage: month of birth, day of birth, surname (including surname at birth), given name, middle name, sex, race, first five digits of account number, 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 account number, a record was produced for every combination of these identifiers: starting with about 6 million NUMIDENT records and about 2 million Master Beneficiary Records, close to 30.75 million unique records were produced for linkage, with about 30.5 million going to file B and the remaining quarter of a million to file A. We believe that this is one of the largest ever applications of probabilistic linkage.

The record pairs with very high linkage scores generally belonged either to one person or to two twins. The following criteria were used to decide which pairs belonged to one person and which to twins. First, if the given names are identical, it may be assumed that the two numbers belong to one person. On the other hand, if the given names are quite dissimilar, and certainly if one record belongs to a male and the other to a female, or one belongs to a living person and the other to a deceased person, then it may be assumed that the pair of numbers belong to twins. Lastly, if the given names are similar and there are no other immediate clues, other information, such as the lifetime earnings patterns for the two account numbers (using earnings records maintained by the Social Security Administration), is used to arrive at a decision.

The probabilistic linkage produced 324 sets of twins in which one of the twins had an account number in the 1 in 100 sample and the other did not. The inference is that an estimated 16,400 pairs of twins born in 1919 who were issued account numbers can be identified in this way, since almost 1 in 50 (actually 0.0198) sets of twins are expected to have exactly one number in the 1 in 100 sample. Given the earlier estimate that 19,000 twin pairs born in 1919 in the US survived to 1936 when account numbers were first issued, and bearing in mind that significantly fewer pairs were actually issued numbers since many people with no need for a number because they were not employed did not apply for one immediately, it is clear that the great majority of twins with account numbers can be identified by probabilistic linkage.

Discussion

How does probabilistic linkage compare to exact linkage? We found that *exact* linkage in the file of 30.75 million records on a combination of date of birth, State of birth, surname (Soundex), mother's maiden name (Soundex), mother's given name, and father's given name yielded 172 twin sets born in 1919 with one number in the 1 in 100 sample and the other not, only slightly more than half the yield of the probabilistic linkage methodology. On the other hand, although less powerful than probabilistic linkage, exact linkage offers a much simpler alternative for searching for twins, which appears capable of identifying close to half of twin pairs who have been issued account numbers. In a country as populous as the United States, this is a large number of twins.

The distribution by sex and race of the 324 twin pairs assembled by the probabilistic linkage was unexpected, but sampling error is substantial in a sample this small. The split by sex was: 94 male–male, 144 female–female, and 86 male–female. Since at birth the three-way split is roughly equal, the unlike-sex group, which has the lowest infant mortality, would be expected to have the greatest number of survivors.

As for race, 15 percent of the twin pairs, or 49 pairs, were Black. Taking the higher twinning rates of Blacks relative to Whites (Heuser, 1967) into account, it is reasonable that about 15% of twins born in 1919 were Black; given the much higher mortality of Blacks, however, we expected that Blacks would comprise much less than 15% of the pairs who survived to adulthood.

This paper reported on the potential of the NUMIDENT, the closest thing in the United States to a population register, for supporting twin research, and described probabilistic linkage methodology. We expect to do additional research on twins using the unique data resources of the Social Security Administration; for example, we have begun an investigation of the similarity of earnings histories of twins.

Author Note

Based on a paper presented at the Eighth International Congress of Twin Studies in Richmond, Virginia, USA, May 1995.

Opinions expressed in this paper are those of the author, and no official endorsement by the Social Security Administration should be inferred.

References

Aging twins offer clues to late onset diseases. (1993). *Science, 259*, 1826–1827.

Botting, B. J., Macdonald Davies, I., & MacFarlane, A. J. (1987). Recent trends in the incidence of multiple births and associated mortality. *Archives of Diseases in Childhood, 62*, 941–950.

Bulmer, M. G. (1970). *The biology of twinning in man*. Oxford: Clarendon.

https://doi.org/10.1375/twin.7.2.134 Published online by Cambridge University Press

Elderly twin registry in the works. (1993). *Science, 260*, 1239.

Fowler, M. G., Kleinman, J. C., & Kiely, J. L. (1991). Double jeopardy: Twin infant mortality in the United States, 1983 and 1984. *American Journal of Obstetrics and Gynecology, 165*, 15–22.

Gittelsohn, A. M., & Milham, S., Jr. (1964). Statistical study of twins. *American Journal of Public Health, 54*, 286–294.

Gittelsohn, A. M., & Milham, S., Jr. (1965). Observations on twinning in New York state. *British Journal of Preventive and Social Medicine, 19*, 8–17.

Heuser, R. L. (1967). *Multiple births: United States — 1964* (Vital and health statistics series 21, No. 14). Washington: Public Health Service.

Hougaard, P., Harvald, B., & Holm, N. V. (1992). Measuring the similarities between the lifetimes of adult Danish twins born between 1881–1930. *Journal of the American Statistical Association, 87*, 17–24.

Jablon, S., Neel, J. V., Gershowitz, H., & Atkinson, G. F. (1967). The NAS-NRC twin panel: Methods of construction of the panel, zygosity diagnosis, and proposed use. *American Journal of Human Genetics, 19*, 133–161.

Kaprio, J., Koskenvuo, M., Seppo, S., & Rantasalo, I. (1981). The Finnish Twin Registry: A preliminary report. In S. A. Mednick, A. E. Baert, & B. P. Bachmann (Eds.), *Prospective longitudinal research: An empirical basis for the primary prevention of psychosocial disorders*. New York: Oxford.

Kestenbaum, B. (1988). *The 1919 birth cohort*. Paper presented at the Annual Meeting of the American Statistical Association, New Orleans.

Kleinman, J. C., Fowler, M. G., & Kessel, S. S. (1991). Comparison of infant mortality among twins and singletons: United States, 1960 and 1983. *American Journal of Epidemiology, 133*, 133–143.

National Center for Health Statistics. (1994). *Vital statistics of the United States, 1990, Vol. 1 — Natality.* Washington: Public Health Service.

Newcombe, H. B., Fair, M. E., & Lalonde, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association, 87*, 1193–1203.

Newcombe, H. B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. New York: Oxford Medical Publications.

Rutstein, S. O. (1984). Infant and child mortality: Levels, trends, and demographic differentials (Rev. ed., WFS Comparison Studies No. 43). Voorburg, Netherlands: International Statistical Institute.

Social Security Administration. (1992). *Life tables for the United States social security area 1900–2080. Actuarial study No. 107*. Baltimore: Social Security Administration.

https://doi.org/10.1375/twin.7.2.134 Published online by Cambridge University Press