Twin Research (1998) 1, 150–153
© 1998 Stockton Press All rights reserved 1369–0523/98 $12.00
http://www.stockton-press.co.uk/tr

Similarity of social security numbers among twins: data from the Virginia Twin Registry

William F Page[1] and Linda Corey[2]

[1]Medical Follow-up Agency, Institute of Medicine, National Academy of Sciences, Washington, DC, USA
[2]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA

At least two twin registries in the United States have been or are being assembled using the similarity of Social Security Numbers in computerized records to help identify possible twin pairs. While the success of such enterprises depends directly on a high probability of twinness given Social Security Numbers, there are theoretical and practical reasons to study the probability of Social Security Number similarity given twinness. For example, the number of twin pairs with similar Social Security Numbers obviously determines the maximum number of twin pairs that can be discovered by similarity algorithms. To study this issue, we examined the similarity of known Social Security Numbers in twin pairs from the Virginia Twin Registry by age, sex, race, and zygosity of the pair. We found that similarity between the Social Security Numbers of twin pairs varies markedly by age, and MZ twin pairs have significantly more similar Social Security Numbers than DZ pairs at all ages. Among older twins, there are also significant differences by sex and race. For younger twins, algorithms that identify putative twin pairs on the basis of the similarity of their Social Security Numbers hold the promise of being able to identify a large proportion of all true twin pairs. Such algorithms will be substantially less successful, however, in identifying a large proportion of older twin pairs.

Keywords: twins, epidemiologic methods, ascertainment bias, medical records linkage, United States Social Security Administration

Introduction

A recent methodological article on identifying elderly twins in the United States[1] examined the use of record linkage algorithms to identify putative twin pairs from Medicare records. Mailings were then made to the members of these putative pairs to confirm whether the subjects were indeed twins. The record linkage algorithms relied heavily on similarity of Social Security Numbers (SSNs), and the investigators found that the more similar the SSNs for a given pair, the more likely the pair members were to be twins. At least one other twin registry, the Vietnam Veteran Twin Registry, was assembled along similar lines.[2,3]

In this paper, we examine this issue from the opposite perspective by investigating the similarity of SSNs among known true pairs. This issue is of practical import because, obviously, the number of twin pairs with similar SSNs determines the maximum number of twin pairs that can be discovered by SSN similarity. If, for example, it turns out that the number of twin pairs identifiable by SSN similarity is small relative to the total number of twin pairs, then a twin panel assembled on the basis of SSN similarity is clearly at risk of being unrepresentative of the true twin population. In the course of our investigations, we also examined how SSN similarity among twin pairs varies by sex and zygosity, by race, and by age.

Materials and methods

The Virginia Twin Registry, developed at the Medical College of Virginia by Drs Linda Corey and Walter E Nance in 1978,[4] is a population-based panel that now includes pairs of twins who were born in the Commonwealth of Virginia between 1915 and 1979. Of the more than 80 000 twins and other multiples born in Virginia during this period, it has been possible to trace a total of 31 537 individuals by using birth record information (twin's name, sex, and date of birth) to match records of the Virginia Department of Motor Vehicles. This match provided SSNs for a total of 7092 complete twin pairs. All twins and other multiples for whom a current address was available were contacted and asked to participate in the twin registry by completing a one-page zygosity determination questionnaire. Data analyses for this report are limited to the 4908

Correspondence: Dr William Page, Medical Follow-up Agency, National Academy of Sciences, 2101 Constitution Avenue, NW, Washington, DC, 20418, USA. Tel: 202 334 2828; Fax: 202 334 2685; E-mail: wpage@nas.edu
Received 16 February 1998; accepted 2 April 1998

https://doi.org/10.1375/twin.1.3.150 Published online by Cambridge University Press

Table 1 Percentage of twin pairs with similar[a] Social Security Numbers (number), by race, age, and pair sex-zygosity[b] (number of pairs shown in parentheses)

	Race and age category					
Pair's sex-zygosity	White 30 & under	Black 30 & under	White 31–45	Black 31–45	White 46 & over	Black 46 & over
MMMZ	90.7 (354)	95.4 (43)	78.2 (537)	66.2 (77)	59.9 (314)	70.0 (30)
MMDZ	86.6 (194)	79.3 (29)	75.6 (390)	56.0 (75)	41.4 (350)	48.9 (45)
FFMZ	91.5 (307)	88.5 (52)	75.4 (142)	67.7 (34)	69.4 (49)	–
FFDZ	90.3 (196)	85.1 (47)	67.1 (85)	55.6 (36)	72.4 (29)	–
MFDZ	86.6 (592)	80.7 (140)	62.6 (404)	47.7 (128)	24.7 (178)	21.1 (19)

[a] Social Security Numbers are similar if the first five digits are identical (see text).
[b] pair sex-zygosity codes are as follows: MMMZ=male/male monozygous; MMDZ=male/male dizygous; FFMZ=female/female monozygous; FFDZ=female/female dizygous; MFDZ=male/female dizygous.
–denotes number of pairs less than 10.

complete twin pairs with SSNs for whom it was possible to attribute zygosity. In particular, for same-sex pairs it was necessary to determine zygosity by questionnaire, whereas for different-sex pairs, zygosity is already known. Zygosity determined by questionnaire has been shown by others to be quite accurate.[5,6]

Each pair in the study was categorized as similar in SSN (identical in the first five digits of SSN) or not, and the pair's sex and zygosity was categorized as male–male monozygous, male–male dizygous, female–female monozygous, female–female dizygous, or male–female dizygous. Race was categorized as Black, White, and Other, and age in 1996 as less than or equal to 30 years, 31–45 years, and 46 years and over, the latter categories chosen primarily to yield groups of sufficient size for analysis. There were too few pairs of other race to be analyzed.

Regarding SSN similarity, it should be noted that SSNs consist of three separate data fields,[7] the first three digits (area number) showing the state of issuance and the next two digits the group number within state of issuance. Thus similar SSNs (identical first five digits) share the same state and group number. Until 1972, within a particular area and group number, the last four digits were assigned in simple sequential order; since then, they have been issued in a randomized order, largely to avoid issuing consecutive numbers to persons with the same surname.[7] Also, at one time the area number indicated the Social Security Administration field office from which the SSN was issued, but now it has no significance other than to identify the SSN applicant's state of residence. We did not make use of the ordering of group numbers (they are not ordered in a simple sequential fashion) nor of the last four SSN digits in defining SSN similarity.

Logistic regression analysis was used to quantify the effects of the various factors on SSN similarity. Pair sex and zygosity were used to create three dummy variables: two for sex (male–male pairs versus female–female pairs and same sex versus unlike sex) and one for zygosity (monozygous [MZ] versus dizygous [DZ]). Race and age were coded categorically, as was SSN similarity score (yes/no).

Results

Table 1 shows the percentage of twins with similar SSNs by race, age, and pair sex–zygosity, except that there were too few (less than 10) black female–female twin pairs in the 46 years and older group to provide meaningful data. The table shows that SSN similarity rates decrease with increasing age of the twin pairs, that there is also a general tendency for SSN similarity rates to be higher among MZ than DZ twin pairs of the same race and sex, and that SSN similarity rates for male–female pairs tend to be lower than for same-sex pairs, even in the under age 30 group. Also, because zygosity is known for male–female pairs and thus there was no need to ascertain it by questionnaire, there are relatively large numbers of such pairs.

The logistic model showed significant effects of pair zygosity, age, and race (see Table 2, first column). Specifically, different-sex pairs, black pairs, and DZ pairs had about two-thirds the odds of having identical SSNs in the first five digits as same-sex pairs, white pairs, and MZ pairs, respectively. More strikingly, twin pairs in the two older age groups had odds of SSN similarity of only 0.28 and 0.10, respectively, relative to pairs in the youngest age group. Among same-sex pairs, there were no differences in SSN similarity between male–male and female–female pairs.

Given the fact that SSN similarity scores appeared to be much more uniform in the youngest age group, we fitted separate logistic regression models for each

Table 2 Odds ratios for twin pairs having similar[a] Social Security Numbers, for the group as a whole and separately by age group

Factor	All ages	Age 30 & under	Age 31–45	Age 46 & over
Female vs male	1.030	1.178	0.816	2.367[b]
Male-female vs all same sex	0.637[b]	0.922	0.588[b]	0.426[b]
Black vs white	0.685[b]	0.713	0.526[b]	1.380
DZ vs MZ	0.656[b]	0.688[b]	0.778[b]	0.506[b]
Age 31–45 vs age 30 & under	0.285[b]	–	–	–
Age 46 & over vs age 30 & under	0.103[b]	–	–	–

[a]Social Security Numbers are similar if the first five digits are identical (see text).
[b]P<0.05.
–parameters not estimated.

of the age groups; these results are also shown in Table 2. In brief, only zygosity (ie, MZ versus DZ) is significant (barely, P = 0.0492) in the under age 30 group; zygosity, race, and pair sex (same-sex versus different sex) factors are all significant in the age 31–45 group; and zygosity and pair sex (male versus female and same sex versus different sex) are significant factors in the age 46 and over group.

Discussion

The results of these analyses show that several factors are related to SSN similarity among twin pairs. Most important is the effect of birth cohort, here measured using current age, with the youngest cohort showing the greatest similarity and the oldest the least.

These large differences in SSN similarity by birth cohort are the backdrop against which the effects of other factors must be viewed. The logistic regression analysis, for example, found significant the effects of same-sex pairs, race, zygosity, and age on SSN similarity, but only one of these factors, zygosity, was significant among the cohort of twins now age 30 and under. In contrast, in the older groups, three of four factors (age now excluded) had significant effects on SSN similarity.

The practical effects of these age differences are substantial. Because Virginia twin pairs now under age 30 have similar SSNs close to 90% of the time, the use of an SSN algorithm to identify putative twin pairs would exclude very few true twin pairs. Presumably, this would be the case elsewhere as well, although the rate of SSN similarity could be lower in larger states. Among those twins now age 46 and over, however, it was sometimes the case that fewer than half the twin pair members have similar SSNs (for example, the white male–male DZ and male–female DZ pairs of both races). Use of five-digit SSN similarity scores would thus identify at most a minority of all such twin pairs, a finding mirrored in the low yield of elderly twin pairs seen by Goldberg et al.[1]

In a like manner, only around 75% of male–male twin pairs in the 31–45 year old group had similar SSNs. Thus, it is not surprising that when Goldberg et al compared the list of all Connecticut-born male–male twin pairs born in the years 1939–1955 who served in the military during the Vietnam era with the list of twin pairs identified from computerized military records files using an SSN-similarity algorithm, they found that only 46.7% of the known Connecticut twin pairs were identified from the computer files.[3]

We believe that the usefulness of SSN similarity in identifying twins is explained by the following hypothesis: equality of the first five digits of an SSN pair – for SSNs issued before 1972, one could meaningfully compare all nine digits of the SSN – means that both members of the pair made application for their SSN (or application was made for them) at roughly the same place (ie, state) and time. If so, then SSN similarity, we further assert, raises the probability that the members of a pair are family members, for we believe that twin pairs in the same household (usually the case for twins) have been more likely to apply for SSNs at roughly the same place and time than unrelated individuals who merely share the same last name and birthday.

Using data provided by the Social Security Administration (Bert Kestenbaum, Office of the Chief Actuary, Social Security Administration; personal communication, 1997), we tested the above hypothesis by looking at year of issue by area and group number for a 5% sample of SSN pairs in the study. We found that dissimilarity of SSN reflected dissimilarity of year of issue, especially for older pairs, and that similarity of SSN pairs means that they were usually issued contemporaneously. Here it is useful to remember how the issuance of SSNs has changed over time: early on, SSNs were issued to persons who were working, later to persons who were about to begin work, and later to persons for non-working purposes. Finally, because issuance of SSNs is a national process, we have no reason to believe that Virginia's experience is different from any other US state.

Bayes Theorem is perhaps the best way to sort out the relationships between various mathematical quantities of interest here. Let P(T) denote the probability that a pair of individuals are twins, $P(T^{o})$ denote the probability that a pair of individuals are not twins, P(S) denote the probability that a pair of individuals have similar SSNs, and P(A| B) denote the conditional probability of A, given B. Then,

https://doi.org/10.1375/twin.1.3.150 Published online by Cambridge University Press

$$\begin{aligned} P(T|S) = {} & P(T)^{*}P(S|T)/ \\ & P(T)^{*}P(S|T) + P(T^{o})^{*}P(S|T^{o}) \end{aligned}$$

Thus, $P(T|S)$ depends on both $P(T)^{*}P(S|T)$ and $P(T^{o})^{*}P(S|T^{o})$, and for example, the larger $P(T^{o*}P(S|T^{o})$ becomes, the further $P(T|S)$ departs from 1.0. Although our study design offers no opportunity for direct observations of $P(S|T^{o})$, we nevertheless speculate that in the future, as SSNs are issued at birth or soon thereafter (as is now frequently the case), $P(S|T^{o})$ will increase. Thus, if we further assume that $P(S|T)$ remains unchanged, the hypothesized increase in $P(S|T^{o})$ will mean a decrease in $P(T|S)$ and a lower yield of twins from SSN similarity algorithms.

A few words about the general applicability of these results are in order. The Virginia Twin Registry subfile we analyzed includes twins born between 1915 and 1979 who could be traced using the records of the Virginia Department of Motor Vehicles. Because female twins were traced under their maiden names, and these records list only those names under which an individual has had a driver's license, female twins who obtained their first driver's license after marriage would not be traceable. This has led to under-representation of female twin pairs in the older age groups and an increased proportion of male–male twin pairs. Indeed, less than half of Virginia-born twins ascertained from birth records were traceable, which suggests that more than half of those twins may have migrated from the Commonwealth, not applied for a driver's license, or married and changed their name.

However, these overall figures are somewhat misleading, for traceability rates range from less than 4% for female–female twin pairs born prior to 1935 to greater than 65% for male–male twin pairs born after 1955. Although there are conceivable associations between traceability and SSN similarity – for example, older twin pairs with SSNs issued later in life (which are presumably more apt to be dissimilar) might have been more likely to migrate before they were traced – it is unclear how strong these associations may be or whether they would affect the generalizability of our results, particularly in the youngest age group, where tracing is more complete.

Although our analysis was not constructed to allow a direct comparison of the characteristics of traceable with untraceable subjects, we note that the issue of ascertainment bias has been studied in the VETS (Vietnam Era Twin) Registry.[3] Comparing Connecticut-born veteran twins who were also in the VETS registry with those who were not, the investigators found significant differences between the two groups in year of discharge from military service, total length of military service, branch of service, and foreign service. However, there was no consistent pattern of differences related to physical or psychosocial health.

In summary, similarity between the SSNs of twin pairs varies markedly by age, and MZ twin pairs have significantly more similar SSNs than DZ pairs at all ages. Among older twins, there are also significant differences by sex and race. For younger twins, algorithms that identify putative twin pairs on the basis of the similarity of their SSNs hold the promise of being able to identify a large proportion of all true twin pairs. Such algorithms will be substantially less successful, however, in identifying a large proportion of older twin pairs.

Acknowledgements

The authors wish to thank Bert Kestenbaum, Office of the Chief Actuary of the Social Security Administration, for providing data on the years of issue of SSNs in Virginia by area and group number and for his insightful comments on early drafts of this report. Dr Corey's work was supported by a grant from the National Institute of Neurological Disease and Stroke (R01 NS 31564). The opinions expressed in this report should not be construed as representing the official views of the Institute of Medicine or the National Academy of Sciences.

References

1 Goldberg J, Miles TP, Furner S, Meyer JM, Hinds A, Ramakrishnan V. Identification of a cohort of male and female twins aged 65 years or more in the United States. Am J Epidemiol 1997; 145: 175–183.

2 Eisen S, True W, Goldberg J, Henderson W, Robinette CD. The Vietnam Era Twin (VET) Registry: method of construction. Acta Genet Med Gemellol (Roma) 1987; 36: 61–66.

3 Goldberg J, True W, Eisen S, Henderson W, Robinette CD. The Vietnam Era Twin (VET) Registry: ascertainment bias. Acta Genet Med Gemellol (Roma) 1987; 36: 67–78.

4 Corey LA, Berg K, Pellock JM, Solaas MN, Nance WE, DeLorenzo RJ. The occurrence of epilepsy and febrile seizures in Virginian and Norwegian twins. Neurology 1991; 41: 1433–1436.

5 Jablon S, Neel JV, Gershowitz H, Atkinson GF. The NAS-NRC Twin Panel: Methods of construction of the panel, zygosity diagnosis, and proposed use. Am J Hum Genet 1967; 19(2): 133–161.

6 Magnus P, Berg K, Nance WE. Predicting zygosity in Norwegian twin pairs born 1915–1960. Clin Genet 1983; 24: 102–112.

7 Jabine TB. Properties of the Social Security Number relevant to its use in records linkage. In: Record Linkage Techniques – 1985. Washington, DC: US Internal Revenue Service, 1985, Publication 1209, 2–86.