Computer Record Linkage on a Survey Population

NEDRA B. BELLOC, MA, and MAX G. ARELLANO, MA

Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

WITH the preceding poetic words, Halbert L. Dunn began a discussion of the general subject of record linkage in a paper published in 1946 (*1*).

Since that time the theory of record linkage has been well documented (*2–11*), but only a few reports (*12–18*) give the practical procedures and the results of an application of a computer record linkage operation. None of the authors cited attempted to measure the accuracy of the process on other than a sample basis.

For long-term studies of important problems, public health workers have available a vast resource in the certificates of births, marriages, divorces, and deaths which are registered and filed routinely and which can be utilized by application of electronic computer techniques (*19*). Identified populations which have been studied for other purposes can become the basis for mortality studies through a search of death records, if this procedure can be carried out efficiently and inexpensively. An example of the way in which a population of 8,000 adults, surveyed in 1965, was followed by a search of the death records for the succeeding 5½ years is described in this paper. A comparison of the results is made with the expected number of deaths calculated from age-specific death rates.

In 1965 the Human Population Laboratory completed a survey of health and ways of living in Alameda County, Calif. (*20*). The survey population consisted of all adults (age 20 or over or under 20 if ever married) living in a probability sample area of 4,735 housing units in the county. Persons living in 97 percent of the households were enumerated, and 86 percent of those enumerated filled out questionnaires (respondents).

Mrs. Belloc is a research specialist in the Human Population Laboratory and Mr. Arellano is a biostatistician with the California Tumor Registry, California State Department of Public Health, Berkeley. *This study was supported by Public Health Service grant No. HS 00368, National Center for Health Services Research and Development, and contract No. NIH-69-5-Biometry Branch, Etiology, National Cancer Institute.* *Tearsheet requests to Nedra B. Belloc, Human Population Laboratory, 2151 Berkeley Way, Berkeley, Calif. 94704.*

The nonrespondents included proportionately more older persons, males, single or widowed persons, and whites than the respondents. For most purposes of the survey, however, these differences had negligible effects on population estimates, and the respondents were an adequate representative sample of the adult noninstitutional population of the county.

Six years after the completion of the survey, the death records in the State office of vital statistics registration were searched by two independent computer matching programs. This search was made to determine which of the 6,928 persons who filled out questionnaires in the survey (respondents) and the 1,146 persons in the enumerated sample who did not respond had died in California in the intervening period. It was undertaken with the knowledge that records of some of those who died were likely to be missed; that is, those who had moved from the State, women who had married, and others for whom name or other identifying information was erroneous either in the survey or in the death records.

Procedures

A decision had to be made between two record linkage systems which were available for use. One, developed initially for the periodic updating of master files, in which decisions are based on a complex mathematical model and the other, an empirical system, developed for death clearance purposes, which utilizes a simple scoring system to produce pairs of possible matching records for subsequent visual inspection. Time and cost considerations dictated a decision in favor of the empirical system; it was felt that the effort required to adapt the system based on the mathematical model to the requirements of this study and to derive estimates of the relevant parameters would not be offset by significantly improved results.

The initial matching operation was done on an RCA Spectra/45 computer with core capacity 131K. Primary matching items were the first four characters of the surname, sex, and color (white or nonwhite). The survey file and the California State Master Death Index file for each year, 1965–70 (approximately 160,000 records per year) were ordered to correspond. Those records on the two files which matched on the three primary matching items were then compared on the following secondary matching items:

1. Second four characters of the surname
2. Next three characters of the surname (surname limited to 11 letters)
3. Initial letter of first name
4. Next four characters of first name
5. Remaining three characters of first name (limited to eight letters)
6. Middle initial
7. Month of birth
8. Day of birth
9. Year of birth (±5 years)
10. Initial letter of birthplace

These items were selected because they were common to both files. Marital status was the only other item which might have been used. It was eliminated because it is subject to change. The four-character divisions of the name were chosen because of the requirements of our computer.

In the matching process, each of the 10 secondary matching items was weighted equally. The scoring system assigned a value of 1 to each positive match, a value of 0 to each nonmatch, and a

Table 1. Number of computer linked records (name-match method) with number and percent of verified deaths, by point score

Point score [1]	Survey respondents				Survey nonrespondents			
	Linked records	Searched further	Death verified		Linked records	Searched further	Death verified	
			Number	Percent			Number	Percent
10.0 (max.)	240	240	239	99.6	...	...	...	...
9.5	10	10	9	90.0	1	1	0	...
9.0	88	80	69	78.4	3	3	1	33.3
8.5	18	11	3	16.7	155	143	70	45.2
8.0	372	79	24	6.5	97	41	5	5.2
7.5	95	15	0	...	845	249	12	1.4
7.0	2,262	76	9	.4	348	34	0	...
Total	3,085	511	353	11.4	1,449	471	88	6.1

[1] Number of matched items with failure to match because of unknown equals ½ point.

value of ½ to each comparison in which one or both items were unknown. The program printed out all pairs of records for which a score equal to or greater than a specified threshold value was attained. These printed linked records were then inspected visually by the senior author (N.B.) and, if judged to be possible matches, the survey data were compared with the death certificate for ultimate verification.

In this application, a low threshold value which would minimize the proportion of false negatives was deliberately chosen, even though it meant that the proportion of false positives would be high. The rationale behind this decision was that most of the false positive matches could be eliminated quickly by visual inspection. We used the threshold of 7.0 with the results shown in table 1.

Lowering the threshold from 7.0 to 6.5 would have increased the number of linked records by 75 percent; further lowering it to 6.0 would have resulted in an increase of more than threefold. Since few death records had scores of 7.0 and 7.5, we believed that the yield with 6.0 and 6.5 would have been insufficient to justify the additional work of a further search.

The computer produced a printout on which the specified items from the two record sources were listed on parallel lines, together with the death certificate file number and the county of residence of the decedent. Inspection of these linked records eliminated a large proportion. For example, although other items matched, the year of birth might differ by 30 years, or persons with similar names and identical birth dates might appear on the list with birthplaces in California or Colorado (Cal and Colo), which counted as a match on the first letter.

About one pair in five was selected for comparison with the death certificate in the Bureau of Vital Statistics Registration. The certificates were pulled, and the following items were compared:

Survey questionnaire	*Death certificate*
Address at time of survey in 1965	Address at time of death
Marital status at time of survey	Marital status at time of death
Name of spouse	Name of spouse
Occupation	Occupation
Name of person to contact or relative in household	Name and address of informant

If the address was the same on both records, obviously no further confirmation was required. If, however, the addresses differed, the name of spouse or the occupation often provided verification. Deaths of widows whose occupation was keeping house were the most difficult to confirm but, for some of these, the informant on the death certificate was the same as a member of the family named in the questionnaires. In only six instances did the death certificate fail to yield some item which verified the match previously made by name, birth date, and place of birth. These were verified if they were unusual names and if there were no items which proved that they were not the same persons.

Those who filled out questionnaires gave, in addition to the items used in the computer linkage and those listed above, the birthplaces of the parents. In linked records in which there was some discrepancy in name or age, all these items could be used as clues for the verification or rejection of the match. For nonrespondents, the enumerators obtained name and where possible age, occupation, and marital status (table 2). For this

Table 2. Items which provided verification of death on records previously linked by computer, by year

Item	Total deaths		1965	1966	1967	1968	1969	1970
	Number	Percent						
Survey respondents	371	100.0	21	72	60	83	66	69
Address	288	77.6	18	63	46	62	51	48
Name of spouse	34	9.2	....	5	3	9	10	7
Occupation	27	7.3	....	3	7	7	3	7
Informant	12	3.2	1	....	2	2	2	5
Other	10	2.7	2	1	2	3	....	2
Survey nonrespondents	88	100.0	21	22	12	9	13	11
Address	65	73.9	20	15	8	7	10	5
Name of spouse	10	11.4	....	3	2	1	2	2
Occupation	7	7.9	1	2	1	....	....	3
Informant	5	5.7	....	2	1	1	1	....
Other	1	1.1	....	....	....	....	....	1

group, then, the verification of linked records could not be carried as far, and for that reason, the matching may be less complete than it was for the survey respondents.

Because the first method would miss any decedents whose names differed in the two records, through misspelling, marriage, or the use of an alias, a second program was developed which matched on month and day of birth, first two letters of birthplace, sex, and first letter of first name. This program could be used only with records for which all of the items were known. This method excluded 307 of the respondents and all the enumerated persons for whom there were no questionnaires (nonrespondents). To reduce the possible matches by eliminating those with gross differences in age, the year of birth was required to be within 10 years. Secondary matching items were not used in this program.

The yield in verified deaths by the two methods is shown in table 3. The name-match system produced more linked records which were finally verified than did the birth date system, but the birth date system did add 18 deaths, about 5 percent, to the total.

Reliability of Findings

This method of doing a record linkage on a survey population appears to be effective in producing possible matches which can be easily verified as decedents by the information available. One important question remains. How many decedents are missed? We attempted to answer this question in two ways, by a manual search of part of the names and by a comparison of the number of deaths found with the number estimated by applying age-specific death rates to the survey populatic n.

The senior author (N.B.) and an assistant made a manual search of 1,100 names from the sample population in the 1968 index and 1,100 in the 1969 index. In 1968, no deaths were found which had not been revealed by the computer. In 1969, two possible matches were found among the nonrespondents. These differed on a number of the items, so it was impossible to determine without investigation beyond the vital records whether they were indeed decedents from the sample population. Because the manual method is subject to a rather high error and because it would also be likely to miss persons whose names differed in the first four letters, this check was not pursued further. (In a study in which a death clearance was done on a population in which many were known to have died, clerks did a manual search on the entire file without knowledge of which were known deaths. About 10 percent of the known deaths were missed by the manual search.)

For the estimated death rates for the Alameda County population, the following data were used. Numerators for death rates were secured from the registered deaths of residents of Alameda County for each year from 1965 through 1970. These were sex specific and in 5-year age categories. For denominators, estimates were obtained by interpolation between U.S. Bureau of the Census tabulations of the 1960 and 1970 Alameda County populations, adjusted to the current estimates of the Alameda County intercensal populations by the department of finance. Using life table techniques, the resulting death rates were applied to the enumerated population (*21*).

Before comparing these expected numbers with the numbers found in the record linkage, two adjustments were necessary. First, because the Alameda County survey covered the noninstitutional population and thus excluded deaths of residents of nursing homes and other long-term care facilities, the survey population would not have the same proportion of deaths as the general population of the county.

Alameda County deaths are shown in table 4 by the type of facility in which death occurred.

Table 3. Yield from two methods of locating decedents through computer linking of records, by year

Year	Total	Number verified deaths located through—		
		Name method only [1]	Both methods	Birth-date method only [2]
1965 (6 months)......	21	3	15	3
1966................	72	8	61	3
1967................	60	5	53	2
1968................	83	17	64	2
1969................	66	17	44	5
1970................	69	13	53	3
Total..........	371	63	290	18

[1] Reason for nonlinkage with birth-date method: variation in birth date, 20; birth date unknown, 15; place of birth unknown, 11; place of birth different, 10; unknown or different first name, 5; more than 10 year difference in year of birth, 2.

[2] Reason for nonlinkage with name method: variation in spelling of name, 12; color different, 3; use of alias, 2; change of name by marriage, 1.

During the period covered by the data, the proportion dying in nursing homes increased from 15.6 in 1965 to 21.9 in 1969, probably because of the increased availability of this type of care. During the first 3 years of the period, the proportion dying in mental, military, penal, and veterans hospitals decreased from 13.7 to 9.9 percent, maintaining a total for these two categories of about 29 percent.

Deaths in the health and ways of living survey population were tabulated in the same categories in table 5. In 1965, only six, or 12 percent, of the 42 known deaths were in long-term facilities; compared with the 29 percent of Alameda County deaths in such facilities, 17 percent of the expected deaths could be attributed to the institutional population which was not included in the 1965 survey. By 1966, the proportion of deaths among the survey population in long-term facilities had increased to 21, in 1967 and 1968 it was 26, and by 1969 it had reached 29 percent, or the same as the proportion in the county population. These results are consistent with the authors' idea that in 3 or 4 years enough of a survey cohort will enter long-term care facilities to make the proportion of deaths in such facilities the same as that in the general population.

The first line in table 6 shows the number of deaths which might have been expected year by year in the survey population of 8,074 adults had the age-sex-specific death rates for the Alameda County population applied. The second and third lines give the adjustment for the institutional population, which was found by comparing the known deaths in the survey population with the proportion of deaths in long-term care facilities among Alameda County deaths. Carrington (*22*), in a study of a similar population in Alameda County, noted that 4.5 percent had moved outside the State in a period of 2½ years, a rate of 1.8 percent per year. On one hand, one might argue that the migrant population tends to be younger and thus to have a lower risk of death than the nonmovers. On the other hand, some high risk older persons, after retirement or the death of a spouse,

Table 4. Deaths of Alameda County residents, by type of facility, 1965–69

Type of facility	1965	1966	1967	1968	1969
Total[1]	8,940	9,253	9,026	9,667	9,724
Not in hospital	2,667	2,627	2,570	2,563	2,533
General and maternity hospitals	3,654	4,022	3,831	([2])	([2])
Nursing homes, convalescent hospitals	1,297	1,507	1,619	1,914	2,001
Mental, penal, military, and veterans hospitals	526	510	437	([2])	([2])
Other specialized hospitals[3]	615	512	403	([2])	([2])
All other facilities	181	75	166	([2])	([2])
Number of deaths over age 20 (=100 percent)	8,335	8,674	8,491	9,098	9,127
Percent dying in nursing homes	15.6	17.4	19.1	21.0	21.9
Percent dying in other long-term facilities	13.7	11.8	9.9	([2])	([2])
Total	29.3	29.2	29.0	(29.0)	(29.0)

[1] Excluding out-of-State deaths for 1965, 1966, 1967.

[2] Categories not comparable to prior years.

[3] Includes Fairmont Hospital, the county long-term care facility.

Table 5. Deaths in health and ways of living survey population, by type of facility, 1965–70

Type of facility	1965	1966	1967	1968	1969	1970	Total
Total	42	94	72	92	79	80	459
Not in hospital (or DOA)	13	24	14	23	21	24	119
General and maternity hosptials	23	50	39	45	35	33	225
Nursing homes and convalescent hospitals	1	10	14	18	18	21	82
Military, veterans, and Public Health Service hospitals	1	4	2	3	2	1	13
Other specialized hospitals[1]	4	6	3	3	3	1	20
Percent dying in nursing homes	2.4	10.6	19.4	19.6	22.8	26.3	17.9
Percent dying in other long-term facilities	9.5	10.6	6.9	6.5	6.3	2.5	7.0
Total	11.9	21.2	26.3	26.1	29.1	28.8	24.9

[1] All but 1 (in 1965) were in Fairmont Hospital, the county's long-term care facility.

Table 6. Expected deaths in a population of 8,074 adults by life-table method of estimation, with adjustments for institutional population and out-of-State migration

Line item	1965 [1]	1966	1967	1968	1969	1970	Total
1. Expected number of deaths by life-table method	47	100	99	106	107	105	564
2. Adjustment for institutional population, percent	−17	−8	−3	−3	....	....	3.9
3. Adjustment for institutional population, number	8.0	8.0	3.0	3.2	....	....	22.2
4. Adjustment for out-of-State migration	.8	2.6	4.1	5.9	7.4	8.5	29.3
5. Net expected number of deaths in resident noninstitutional population (line 1 minus lines 3 and 4)	38.2	89.4	91.9	96.9	99.6	96.5	512.5
6. Found by record linkage	42	94	72	92	79	80	459
7. Percent difference	+10	+5	−22	−5	−21	−17	−10

[1] Half year.

leave California to be with relatives. We have assumed that the effect of these tendencies would be to make the overall death rate for migrants no different from that of the general population of Alameda County and have applied the crude death rate to an estimated 1.8 percent per year, accumulated. The fourth line shows this adjustment.

The net number of deaths in the fifth line in table 6, after the adjustment for the institutional deaths and for outmigration, can then be compared with the number found by record linkage. In 1965, the number found was four more than the estimated number; in 1966, it exceeded the estimate by five. In 1968, the number found approached the estimate, but in 1967, 1969, and 1970, there were substantial deficits. The mortality rates (line six ÷ population at risk) ranged from .009 to .012. On a sample of 8,000, these rates have sampling errors of .0011 to .0012 or about 10 percent. Ninty-five percent confidence intervals around the deaths shown in line six of table 6 would include the net expected number of deaths (line five) in each year except 1967 and 1969. Overall, the record linkage found 90 percent of the number of deaths estimated for the survey population.

The record linkage was done on the enumerated sample in the survey, but proportionally more deaths occurred among the nonrespondents than would be expected by chance ($X^2 = 9.91$, 1 *df*, $P < .01$), and these deaths were concentrated in the first 1½ years after the survey, as shown in table 2. Although nonrespondents made up 14.2 percent of the enumerated populations, deaths among the nonrespondents accounted for 19.2 percent of the total verified deaths. In 1965, half of the deaths were among nonrespondents; in 1966, the proportion was nearly one-quarter. As noted before, the nonrespondents included a higher proportion of older persons, males, and single or widowed persons than the respondents. For these groups the mortality rates are higher. The nonrespondents probably included some persons who were seriously ill and for whom illness was the reason for nonresponse.

When the projected followup survey of this population is done in 1973, we hope to determine whether some decedents were missed by the record linkage process and, if so, why. If it is determined that few decedents were missed, we may find that institutional deaths accounted for a larger proportion than we estimated here, that migration was a greater factor, or that the population included in the survey was in some ways not representative of the general population of the county. We can also determine whether errors in reporting or recording name, birth date, or birthplace contributed substantially to the failure of the computer to link some records.

REFERENCES

(*1*) Dunn, H. L.: Record linkage. Am J Public Health 36: 1412–1416, December 1946.

(*2*) Deming, W. E., and Glasser, G. J.: On the problem of matching lists by samples. J Am Stat Assoc 54: 403–415, June 1959.

(*3*) Raj, D.: On matching lists by samples. J Am Stat Assoc 56: 151–155, March 1961.

(*4*) Kennedy, J. M.: Linkage of birth and marriage records using a digital computer. Doc. No. AEC L-1258. Atomic Energy Commission of Canada, Ltd., Chalk River, Ontario, 1961.

(*5*) Shapiro, S., and Densen, P. M.: Research needs for record matching. American Statistical Association Proceedings, Social Statistics Sec, Sept. 4–7, 1963, Cleveland, Ohio, pp. 20–24.

(*6*) Du Bois, M. S. D'A., Jr.: A document linkage program for digital computers. Behav Sci 10: 3, 312–319, July 1965.

(*7*) Nathan, G.: Outcome probabilities for a record matching process with complete invariant information. J Am Stat Assoc 62: 454–469, June 1967.

(8) Tepping, B. J.: A model for optimum linkage of records. J Am Stat Assoc 63: 1321–1332, December 1968.

(9) Acheson, E. D.: Record linkage in medicine. E. & S. Livingstone, Edinburgh, London, 1968.

(10) Du Bois, M. S. D'A., Jr.: A solution to the problem of linking multivariate documents. J Am Stat Assoc 64: 163–174, March 1969.

(11) Fellegi, I. P., and Sunter, A. B.: A theory for record linkage. J Am Stat Assoc 64: 1183–1210, December 1969.

(12) Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P.: Automatic linkage of vital records. Science 130: 954–958, October 1959.

(13) Newcombe, H. B., and Kennedy, J. M.: Record linkage making maximum use of the discriminatory power of identifying information. Communication Association for Computing Machinery 5: 563–566, November 1962.

(14) Newcombe, H. B., and Rhynas, P. O. W.: Family linkage of population records. The uses of vital and health statistics for genetic and radiation studies. United Nations, N.Y., 1962.

(15) Phillips, W., Jr., and Bahn, A. K.: Experience with computer matching of names. American Statistical Association Proceedings, Social Statistics Sec, Sept. 4–7, 1963, Cleveland, Ohio, pp. 26–38.

(16) Nitzber, D. M., and Sardy, H.: The methodology of computer linkage of health and vital records. American Statistical Association Proceedings, Social Statistics Sec., Sept. 8–11, 1965. Philadelphia, pp. 100–106.

(17) Phillips, W., Jr.: Record linkage for a chronic disease register. Presented at the International Symposium on Medical Record Linkage, Oxford, England, July 1967.

(18) Richards, I. D. G., and Nicholson, M. F.: The Glasgow linked system of child health records. Dev Med Child Neurol 12: 357–367, June 1970.

(19) Moriyama, I. M.: Uses of vital records for epidemiological research. J Chronic Dis 17: 889–897, October 1964.

(20) Hochstim, J. R.: Health and ways of living. *In* The community as an epidemiological laboratory, edited by I. J. Kessler and M. L. Levin. Johns Hopkins Press, Baltimore, 1970, pp. 149–170.

(21) Chiang, C. L.: Introduction to stochastic processes in biostatistics, Ch. 9. John Wiley & Sons, Inc., New York, 1968.

(22) Carrington, R. A.: Analysis of mobility and change in a longitudinal sample. Public Health Rep 85: 49–58, January 1970.

BELLOC, NEDRA B. (California State Department of Public Health, Berkeley) and ARELLANO, MAX G.: *Computer record linkage on a survey population. Health Services Reports, Vol. 88, April 1973, pp. 344–350.*

A technique for following a survey population by computer, and matching items from the survey with the State file of death records is described. Two computer programs were used—one in which the primary match was on name, sex, and color, the other in which month and day of birth, birthplace, sex, and first letter of first name were the only matching variables.

Final verification of computer linked records was made by a comparison of items not coded from the death certificates, such as address, name of spouse, occupation, and name of relative or persons to contact. The name-match method produced more verified deaths than the birth-date method, but the birth-date method added 5 percent to the total.

An estimate was made of the number of deaths to be expected in the survey population. Following adjustment for deaths in long-term care facilities (not included in the survey) and for persons moving from the State, it was found that the computer record linkage had located records of 90 percent of these deaths.