The California Automated Mortality Linkage System (CAMLIS)

MAX G. ARELLANO, MA, MS, GERALD R. PETERSEN, PHD,
DIANA B. PETITTI, MD, MPH, AND ROGER E. SMITH

Abstract: The California Automated Mortality Linkage System (CAMLIS), established in 1981 to facilitate the conduct of follow-up studies in the State of California, employs a combination of deterministic and probabilistic linkage decision criteria to perform the death clearance function. The system was evaluated against four traditional death clearance procedures and the performance of each procedure measured in terms of measures of sensitivity and specificity.

Only one procedure was associated with a specificity lower than 0.99; for that one, the specificity was 0.93. There was much greater fluctuation in the observed sensitivity levels. In one of the procedures, CAMLIS demonstrated a sensitivity of 0.97 versus 0.79 for the Social Security Administration. A comparison against the National Death Index (NDI) produced sensitivities of 0.89 for CAMLIS and 0.94 for the NDI. An assessment of manual search procedures using a file of Japanese names produced a CAMLIS sensitivity measure of 0.92 compared with 0.93 for the manual search. Another manual search procedure using microfiche copies of the state death index produced a CAMLIS sensitivity of 0.97; in this evaluation, the sensitivity of the manual search was defined as 1.0.

Another measure of performance of a death clearance procedure is its predictive value in identifying a person who has died; CAMLIS generated predictive values in these evaluations that ranged from 0.93 through 0.99, contrasted with the NDI value of 0.59. (*Am J Public Health* 1984; 74:1324–1330.)

Introduction

Many scientific investigations require periodic information on the vital status of members of a study population. Mortality files generated from the death certification process are commonly used to make the necessary ascertainment. The traditional search procedure has been to manually compare lists of names of members of a study population against death indices provided by registrars of vital statistics. For many obvious reasons, this approach may be inappropriate on the grounds of economy, efficiency, and, most importantly, accuracy.

Several reports have appeared recently which describe implementations of computerized techniques of vital status ascertainment.[1–6] Scientists in Canada have developed a generalized linkage system based on the concepts of probability and the use of "weighted" record comparisons.[3] Another innovative approach, which also uses probabilistic linkage criteria, has been proposed by Mi, *et al.*[4] The United States National Death Index (NDI), which was established to assist in the conduct of follow-up studies for the entire country, has adopted a deterministic approach to the problem of identifying matched records.[5,6]

The California Automated Mortality Linkage System (CAMLIS) has been developed to assist researchers in the conduct of epidemiologic studies in the State of California. The system uses a combination of probabilistic and deterministic decision criteria to establish the necessary associations between study populations file records and state mortality file records.

Development of Linkage Decision Criteria

Computerized approaches to the record linkage problem generally take one of two forms—deterministic or probabilistic. A deterministic approach requires the formulation of a "match key"* to establish the relationship between the files to be matched. The availability of the social security number has led to its widespread use for deterministic linkage purposes. There are several inherent pitfalls to its use, however: it is not always available either on death certificates or study file records,** there are no safeguards to prevent its use by persons other than the individual to whom it was originally assigned (it is not uncommon to find the husband's social security number recorded on the wife's death certificate or on study file records), and there is no provision for ensuring its correct reporting.

Probabilistic linkage procedures are characterized by a much higher degree of complexity than deterministic linkage procedures with respect to both the decision criteria and the software required to implement them. This complexity translates into a significantly higher operating cost. Probabilistic linkage moreover is a "bootstrap" procedure which requires estimates of parameters which can properly only be derived from the set of matched records. Heretofore these considerations have severely limited the applications of probabilistic record linkage in this country. Probabilistic linkage procedures, however, have many desirable features: they are able to correctly link records despite discrepancies which may exist between identifying variables, techniques are available which permit the assignment of weights on the basis of partial agreements, and they can discriminate between rare and common events. Our analysis of the situation led to the conclusion that probabilistic record linkage procedures were a necessary component of a comprehensive record linkage system.

Address reprint requests to Max G. Arellano, Division of Family and Community Medicine, University of California, San Francisco, CA 94143. Dr. Petersen is with the Hanford Environmental Health Foundation, Richland, WA; Dr. Petitti is with the Permanente Medical Group, Oakland, CA; Mr. Smith is with the Vital Statistics Branch of the Department of Health Services, Sacramento, CA. This paper, submitted to the Journal March 5, 1984, was revised and accepted for publication July 24, 1984.
Editor's Note: See also related editorial p 1302 this issue.

*The match key may consist of any conceivable combination of last name, first name, sex, social security number, birth date (or portions thereof), or any other set of items available on the file.

**In California, the proportion of death certificates with social security numbers has steadily increased from 57 per cent for 1962 to 94 per cent for 1982. However, only one-third of the user files submitted to CAMLIS to date have contained any information whatsoever on the social security number.

Fellegi and Sunter[7] have proposed a theory which provides a conceptually sound probabilistic approach to the problem of record linkage. Their approach requires data on: 1) relative frequencies of variables used for record linkage purposes in mortality files; 2) relative frequencies of variables used for record linkage purposes in mortality files; 3) relative frequencies of variables used for record linkage purposes among correctly matched records; and 4) various estimates of the probability that variables used for record linkage purposes are incorrectly recorded (recording error probabilities). The first two categories of information can be obtained by direct enumeration from the user file and the mortality file. However, the last two categories involve information which can only be obtained from the set of correctly matched user and mortality file records. If that information were available, however, there would be no need to carry out the record linkage.

This dilemma was resolved by recognizing the complementary nature of deterministic and probabilistic linkage procedures. Deterministic procedures are capable of identifying matching records which might escape detection by probability linkage. Conversely, probability linkage is able to identify instances in which the match key variable is either incomplete or incorrectly recorded. Further, by carrying out deterministic linkage first, it is possible to extract estimates of recording error probabilities for most identifying variables from the "Matched Set" defined by deterministic linkage.

Deterministic Linkage

The social security number was an obvious choice for one of the match keys. However, since not all user files could be expected to contain a social security number, it was necessary to generate an alternative match key from the remaining variables. Since a phonetic coding of the last name was to form comparison groups for probability linkage, the last name was excluded from consideration. The middle name also was excluded because it was not consistently recorded. The second or alternative match key was therefore formed by combining the complete birth date, the first two digits of the first name phonetic code, and the sex code. The deterministic linkage criteria are described in further detail in the Appendix.

Probability Linkage

The CAMLIS implementation of the Fellegi-Sunter model[7] required an extensive application of the probability calculus; a general summary of the probabilistic linkage decision algorithms that were employed is presented in the Appendix.

The variables used to define the range of the search had to be present on all records and relatively permanent. Only two of the generally available variables—last name and sex—met these requirements. For variations in last name spelling, a modified version of the New York State Intelligence and Identification System (NYSIIS)[8] was adopted. In order to identify presumed errors in the coding of sex, a first name/sex code edit was developed. Record comparisons were therefore restricted to comparison groups which shared the same last name phonetic code and sex code.

Major Elements of CAMLIS

There are five steps in conducting a CAMLIS linkage:

1. *Standardization*—In the course of standardization, user files are transformed into a single standardized format.
2. *Deterministic Linkage*—If user file records contain a social security number, a social security number match is performed at this point. The alternative match key is used deterministically to match the two files if the user record does not contain a social security number.
3. *Weight Computation*—Weight tables are generated for each variable used by probability linkage. Weights used for interfile probability linkage are functions of mortality file relative frequencies and estimates of recording error probabilities. Each table entry has two parts: 1) a configuration part, such as "Smith", for the last name table; and 2) a weight part consisting of a real number which reflects the weight assigned to the associated configuration.
4. *Probability Linkage*—In the course of probability linkage, all possible pairwise comparisons are carried out between records in corresponding comparison groups of the user and mortality files. A probability linkage weight, computed for each pairwise record comparison, is used to classify each comparison as a high likelihood nonlinkage, low likelihood linkage, or high likelihood linkage; information is retained on all low likelihood and high likelihood comparisons. For high likelihood linkages, the mortality record that achieves the highest probability linkage weight in comparisons against a specific user record is classified as the definitive match.
5. *Cross-reference File Generation*—The final processing step is the comparison of the results of deterministic and probability linkage and the production of a cross-reference file which describes the relationship CAMLIS was able to establish between the user file and the California state mortality file. The cross-reference file output is ranked in decreasing order of certainty of match as measured by the number of matching procedures that produced the match and the probability linkage weight, to facilitate the detection of false positive linkages. No attempt is made to differentiate among high likelihood linkages produced only by deterministic linkage.

CAMLIS procedures permit a distinction between high likelihood and low likelihood linkages. Record-pairs that have been linked by probability linkage are classified on the basis of an upper threshold value; CAMLIS deterministic linkage procedures make the distinction on the basis of the degree of concordance among the identifiers that are not incorporated into the match key. All references in this paper to the CAMLIS decisions are made in terms of the high likelihood linkages.

Evaluations of CAMLIS Methodology

The CAMLIS system was evaluated against four other mortality search procedures: 1) use of the Master Beneficiary Record (MBR) of the Social Security Administration (SSA); 2) manual search of microfiche copies of the California Death Index; 3) the NDI; and 4) manual search of a printed death certificate listing. In each evaluation, an attempt was made to estimate the comparative sensitivity and specificity of CAMLIS against the other mortality search procedures.*** The most accurate method of ascer-

***Sensitivity is the proportion of true positives, subjects truly dead, who are correctly identified by the search as deceased. Specificity is the proportion of true negatives, persons truly alive, who are correctly identified as alive by the search.

taining the vital status of a member of a study population, which is to establish direct contact with the subject or members of the immediate family, is not always feasible. Exclusive reliance on the comparison of information recorded on study file records and death certificates can be expected to result in the misclassification of some study subjects, especially when the available information is incomplete. Thus, precise measures of sensitivity and specificity cannot be applied to the results of a death search, since the truth is rarely known.

Hanford Study

The Hanford Environmental Health Foundation (HEHF), has ascertained deaths among 56,081 employees of the nuclear energy industry for the period 1966–80; death clearance searches of the MBR by the SSA have been one of the primary mechanisms employed by HEHF to accomplish this purpose. All deaths identified by the SSA are routinely verified by HEHF personnel by comparing other items such as addresses and names of close relatives of the employees.

The 56,081 records were cleared against State of California mortality files for the period 1966–80 and for the SSA through 1981. Since the deaths already known to HEHF had previously been verified for their membership in the study cohort, only deaths newly identified by CAMLIS were verified for purposes of this study. Verification is accomplished by comparing supplementary information on the death certificate with that on the employee file.

To estimate sensitivity, the true number of deaths in California were assumed to equal the number of verified deaths. The sensitivity of CAMLIS was 0.97 and that of the SSA search 0.79 (Table 1). Fifteen verified deaths were missed by CAMLIS. The reasons for these false negatives are given in Table 2. Assuming that all workers not known to have died in California are alive, the specificity of the CAMLIS search was 1.0.

Forty-four of the deaths identified by CAMLIS could not be verified. Had these deaths been accepted without verification, 7 per cent of the CAMLIS deaths would have been false positives. These false positives and their linkage weights are displayed in Table 3, along with the true positives, in the same range of linkage weights, for comparison.

These results indicate that false positives can be almost entirely eliminated by concentrating the review process on the high likelihood linkages which achieve the lowest probability linkage weights.

Kaiser-Permanente Test

Independently of the CAMLIS project, investigators at the Department of Medical Methods Research of the Kaiser-Permanente Medical Care Program have developed a procedure for ascertaining the vital status of subjects in the cohort studies they are conducting. Study subjects who are known to be active members of the Kaiser-Permanente Health Plan are considered to be alive. For study subjects who had discontinued their Health Plan membership, a manual search of microfiche copies of the California Death Index is performed to determine whether a person with a same or similar name and the same or a similar birth date died after the date of last known contact. Additional information on potentially matching death certificates is used to verify that the death was correctly ascertained.

For this evaluation, a CAMLIS search of a file that comprised 999 subjects known to have died during the period 1972–80 based on the manual search and 1,000 subjects who had a health check-up at a Kaiser-Permanente facility in 1981 was carried out. The sensitivity of CAMLIS was 0.97 and the specificity 0.93 (Table 1).

Thirty-nine per cent of the false negatives were the result of discrepancies between the dates of death recorded on the Kaiser-Permanente test file and the California mortality file (Table 2). An additional 30 per cent of the deaths were not identified by CAMLIS because of a deficiency in the system's operation which was subsequently corrected. Several of the false negatives could not be accounted for. In the absence of these errors, the sensitivity would have been 0.99.

Sixty-six records in the Kaiser Test (Table 1) were incorrectly identified by CAMLIS as linkages to mortality records. Eighty per cent of these false positives were produced solely by deterministic linkage. Since the test file did not include the social security number or birth day, a variation of the alternative match key was generated for the purposes of this study which omitted the birth day and included two additional digits of the first name phonetic code. Because of the disproportionate number of false positives which it introduced, and since it was able to produce only one valid match that was not also identified by probability linkage, the use of this variation of the alternative match key was subsequently discontinued. All of the false positive records produced by probability linkage had probability linkage weights in the 3.0–3.9 range. Among the true positive records, only 0.7 per cent had weights in this range.

TABLE 1—Distribution of Members of Each of the Study Populations by True Vital Status, Observed Vital Status, Study and Death Clearance Method

True Vital Status by Observed Vital Status*	Study by Death Clearance Method						
	Hanford Study		Kaiser Test	Kaiser Study		Japanese Study	
	CAMLIS	SSA	CAMLIS	CAMLIS	NDI	CAMLIS	Manual
Verified Deaths	584	584	999	83	84	222	222
Classified as Dead (TP)	569	461	966	74	79	205	210
Classified as Alive (FN)	15	123	33	9	5	17	12
Known to be Alive	55,497	**	1,000	4,613	4,612	3,587	3,587
Classified as Dead (FP)	44	**	66	3	54	3	0
Classified as Alive (TN)	55,453	**	934	4,610	4,558	3,584	3,587

*TP—True Positive; FN—False Negative; FP—False Positive; TN—True Negative.
**Data not obtained for SSA validation.

TABLE 2—Distribution of CAMLIS False Negatives by Type of Misclassification and Study

Type of Misclassification	Number of False Negatives by Study				
	Hanford Study	Kaiser Test	Kaiser Study	Japanese Study	All
I. Search Strategy Deficiency					
a. Invalid Date of Last Contact or Death on User File	2	13	0	0	15
b. Sex Code Discrepancy	1	1	0	0	2
c. Last Name Change and Different or No Match Keys	5	2	1	0	8
II. Match Strategy Deficiency					
a. Classified as a Low Likelihood Linkage	2	0	4	12*	18
b. Not Classified as a Low Likelihood Linkage	0	0	0	1	1
III. System Operation Deficiency					
a. Improper Parameter Specification	2	0	4	0	6
b. System Design Error	0	5	0	0	5
c. Program Bug	0	10	0	2	12
IV. Information Deficiency	3	2	0	2	7
Total	15	33	9	17	74

*The classification as low likelihood linkages was a consequence of improper parameter specification.

Kaiser-Permanente Study

This evaluation was made possible by submitting a file of information on 4,696 subjects in a study of potentially less hazardous cigarettes being conducted by investigators in the Kaiser-Permanente Medical Care Program simultaneously to CAMLIS and NDI, with a request for clearance against 1979 and 1980 mortality files (this is referred to as the Kaiser Study in the tables). Death certificates for deaths identified by either procedure were obtained and verified using other information. The sensitivity of CAMLIS was 0.89 and that of the NDI search 0.94 (Table 1).

Nine deaths were missed by CAMLIS (Table 2). The major source of these false negatives was the failure to provide a value for the probability of a change in the social security number, a system operation deficiency. This resulted in the classification of all comparisons between records with different social security numbers as non-matches. Safeguards have since been incorporated into the parameter specification procedures to prevent the recurrence of this type of error. If allowance is made for the corrections made in the parameter specification procedures, the difference in sensitivity between the NDI and CAMLIS is attributable primarily to a single death which CAMLIS missed because the last name was recorded differently on the user and mortality records and the user record did not contain a social security number. If the CAMLIS low likelihood linkages are included, then CAMLIS identified five deaths that were missed by the NDI. These missed deaths reflect the greater tolerance for error of the CAMLIS probabilistic decision criteria. Three of the five contained discrepancies in either birth month or year; the remaining two had different first name phonetic codes. Including the CAMLIS low likelihood linkages in this analysis, however, would increase the number of CAMLIS false positive to 32. The probability linkage weights of the three false positive CAMLIS records ranged from 3.0 to 3.9.

Japanese Male Cohort

The study population consisted of a cohort of 3,809 Japanese males residing in the greater San Francisco Bay Area who had participated in a study, sponsored by the National Heart and Lung Institute, of risk factors associated with cardiovascular disease in 1969 and 1970. Members of the study population were actively followed for the first four years of the study; thus, for some members of this population, the necessary mortality data were available. A study protocol was developed which called for a simultaneous manual and CAMLIS clearance of the study population and the subsequent comparison of results obtained by the two procedures. A file was created and printed which displayed records of all Japanese males on the California State mortality files between the years 1969 and 1980. The file contained complete name, birth month and birth year; the absence of birth day information again made it necessary to generate a match key which replaced the birth day in the standard alternative match key with the last two digits of the first name phonetic code. To estimate sensitivity and specificity, the mortality information previously recorded on members of the cohort was used in conjunction with a careful review of all instances in which there were discrepancies between the manual death clearance and the CAMLIS search. The sensitivity of CAMLIS was 0.92 and that of the manual death clearance was 0.95 (Table 1).

Seventeen verified deaths were missed by CAMLIS (Table 2). Fifty-eight per cent of the deaths with low probability linkage weights were missed because of a reversal in the first and middle names. It is apparently a common practice for Japanese living in America to adopt both Japanese and American first names, each of which is equally valid. The name that appears first in a record is relatively arbitrary. Since the values which were specified for the CAMLIS first name error probabilities assumed the normal "American" naming convention, an unusually high frequency of true matches were classified as "low likelihood linkages."

The misclassification was a consequence of the inability of CAMLIS to generate adequate estimates of first name error probabilities from deterministic linkage since the first name phonetic code had been incorporated into the match key. These circumstances can be taken into account in future applications of the CAMLIS methodology on Japanese study population files either by introducing records

TABLE 3—Distribution of Probability Linkage Weights for All CAMLIS False Positive Matches and the Corresponding True Positive Matches by Match Weight Group among Hanford Study Subjects

Match Weight Group*	Number of Occurrences	
	False Positives	True Positives
3.0–3.9	30	6
4.0–4.9	11	0
5.0–5.9	0	2
6.0–6.9	1	2
7.0–7.9	0	3
8.0–8.9	2	7
Total	44	20**

*These are base 10 logarithms.
**These represent 3.2% of all verified deaths.

with reversed first and middle names or by adjusting the first name and middle name error probabilities to reflect the higher probability of variation.

The specificity of the manual death clearance is defined as 1.0 by the nature of the evaluation process, since the classification was based on the results of manual review (Table 1).

Two of the false positives were produced by probability linkage; both had a probability linkage weight in the 3.0 to 3.9 range. The third, which was produced by deterministic linkage, agreed in all identifiers except for the last name.

Discussion

Seventy-four deaths were missed in the four evaluations of the CAMLIS methodology that are reported on in this paper (Table 2).

Type I misclassifications are the result of deficiencies in the search strategy. Sixty per cent of the Type I deaths were missed as a result of invalid dates of last contact or death on the user records. Whether these missed deaths can be attributed to the CAMLIS methodology is a moot point. Although the date of last contact could be disregarded, the consequences would be an increase in the number of false positives and a greater processing cost. We have attempted to address this issue by not utilizing information on day of last contact and by advising users that unreliable dates of last contact should either not be employed or that one or two years should be subtracted from the recorded dates of last contact before they are employed for death clearance purposes. The potential for reduction of these Type I misclassifications appears to be very good.

Type II misclassifications are the result of deficiencies in the match strategy. Ninety-five per cent of the Type II deaths were classified as "low likelihood" linkages and most of these were obtained from the Japanese Male Cohort Study. Proper attention to the special processing requirements of Oriental study populations and greater experience in the setting of the estimates of recording error probabilities for the first name would reduce these misclassifications considerably. The potential for reduction of these Type II misclassifications thus appears to be very good.

Type III misclassifications are the result of deficiencies in the operation of CAMLIS. Thirty per cent of the Type III deaths can be attributed either to deficiencies in the original system design or to software "bugs" which have since been corrected. The misspecified parameters were estimates of the probabilities of recording error for variables used for record linkage purposes. As experience is gained in the specification of these parameters, we can expect to see an improvement in this area also. The potential for reduction of Type III misclassifications thus appears to be excellent.

Type IV misclassifications are the result of deficiencies in the information that is available to CAMLIS. Fifty-seven per cent of the Type IV deaths represent unaccountable events in that the corresponding deaths could not be located on the California mortality files. The remaining deaths represent instances in which the state file number could not be located, thus preventing a determination of the factors responsible for the missed death. Our expectation is that the level of Type IV misclassifications will probably remain constant.

The results from the Hanford Study require some caution in drawing blanket conclusions. The period covered in this study, 1966–80, overlapped a critical event at SSA. In November 1977, a change was made that resulted in the loss of an undisclosed number of deaths for 1978 through 1980.[9] Recent information from CAMLIS, for the period 1960 through 1982, and for the SSA through September 1983 indicates that, for each year up to 1977, about 6 per cent of all known California deaths among former Hanford employees were missed by the SSA search. In 1978, this increased to 22 per cent and in 1979 it was almost 70 per cent. Forty per cent of the California deaths identified by CAMLIS were missed by the SSA search for years 1980 and 1981. It remains to be seen whether or not SSA can recover. This makes a strong argument for using linkage systems such as CAMLIS in regional settings and possibly even on a national level with the addition of deaths for the year 1978.

The NDI was established in 1981 by the National Center for Health Statistics (NCHS) in response to the demand for better access to mortality files by investigators engaged in cohort studies in the medical and health research fields. By collecting standardized death certificate information from 54 registration areas, representing 100 per cent of the deaths in the United States, the NDI is able to provide the necessary access to national mortality files. The NDI and CAMLIS represent approaches to death clearance which differ in many significant respects. The salient features of these two systems are compared in Table 4.

TABLE 4—Comparison of the National Death Index (NDI) and the California Automated Mortality Linkage System (CAMLIS)

Process	NDI	CAMLIS
Application	Formal	Formal
Review/Approval	Expert Committee	State Registrar of Vital Statistics
Elapsed Time	4–6 months	2 weeks
File Submission Requirement	Strict	None
Purchase Requirement	None	Yes
Decision Criteria	Deterministic	Combination of Deterministic and Probablistic
Cost as a Function of File Size*		
File Size		
100	$110	$550
1,000	$200	$600
10,000	$1,100	$1,100
25,000	$2,600	$1,900
50,000	$5,100	$3,300

*The cost calculations are for studies conducted in California and assume a search over the time period 1979–82.

Rogot[10,11] and Wentworth[12] have recently reported on their experience in using the NDI. Both investigators found approximately two-thirds false positives among the NDI matches.[10,12] CAMLIS generated predictive values in the evaluations reported on in this paper that ranged from 0.93 to 0.99, equivalent to false positive rates of from 1 to 7 per cent. The predictive value of an NDI match for the Kaiser-Permanente Study evaluation was 0.59; if multiple reports are included, approximately half of the NDI matches were false positives, a finding which is not inconsistent with the conclusions of Rogot and Wentworth.

Beebe has summarized the need for record linkage systems to monitor health and to facilitate large-scale epidemiologic studies of health hazards in the United States.[13] As he points out, this country has historically lagged behind Canada in the development of mortality and morbidity data bases and the means of accessing them. The efforts of Newcombe and his colleagues at the Chalk River Nuclear Laboratories[1-3,14-16] have culminated in the development of powerful probabilistically based systems which are capable of performing the necessary record linkage operations quickly and accurately.

Conclusion

An automated mortality linkage system, CAMLIS, has been successfully implemented at the University of California at San Francisco under the sponsorship of the Department of Epidemiology and International Health. The system now offers a rapid, accurate, and cost-effective means of accessing the California state mortality files.

REFERENCES

1. Smith M, Newcombe H: Accuracy of computer vs manual linkage of routine health records. Methods Inf Med 1979; 18:89–97.
2. Smith M, Newcombe H, Dewar R: Automated nationwide death clearance of provincial cancer register files—the Alberta Cancer Registry study. *In:* Alvey W (ed): Statistics of Income and Related Administrative Record Research, 1983. Washington, DC: US Treasury, IRS (Statistics of Income Division), 1983; 43–52.
3. Howe G, Lindsay J: A generalized iterative record linkage computer system for use in medical follow-up studies. Comput Biomed Res 1981; 14:327–340.
4. Mi M, Kagawa J, Earle M: An operational approach to record linkage. Methods Inf Med 1983; 22:77–82.
5. Patterson J: The establishment of a national death index in the United States. *In:* Cairns J, Lyon J, Skolnick M (eds): Cancer Incidence in Defined Populations, Banbury Report 4. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1980; 443–447.
6. Patterson J: Evaluation of the matching effectiveness of the National Death Index. *In:* Alvey W (ed): Statistics of Income and Related Administrative Record Research, 1983. Washington, DC: US Treasury, IRS (Statistics of Income Division), 1983; 65–74.
7. Fellegi I, Sunter A: A theory for record linkage. Am Stat Assoc J 1969; 64:1183–1210.
8. Taft R: Name Search Techniques. Project Search (System for Electronic Analysis and Retrieval of Criminal Histories) Special Report No. 1, New York State Identification and Intelligence System, Bureau of Systems Development. Albany: NYSIIS, 1970.
9. Delbene L, Aziz F: Further investigation into mortality coverage in Social Security Administration data. Proc Am Stat Assoc, Sec Survey Res Methods 1982; 292–297.
10. Rogot E, Feinleib M, Ockay K, Schwartz S, Bilgrad R, Patterson J: On the feasibility of linking Census samples to the National Death Index for epidemiologic studies: a progress report. Am J Public Health 1983; 73:1265–1269.
11. Rogot E, Schwartz S, OConor K, Olsen C: The use of probabilistic methods in matching Census samples to the National Death Index. *In:* Alvey W (ed): Statistics of Income and Related Administrative Record Research, 1983. Washington, DC: US Treasury, IRS (Statistics of Income Division), 1983; 75–82.
12. Wentworth D, Neaton J, Rasmussen W: An evaluation of the Social Security Administration Master Beneficiary Record file and the National Death Index in the ascertainment of vital status. Am J Public Health 1983; 73:1270–1274.
13. Beebe G: Record linkage systems-Canada vs the United States. Am J Public Health 1980; 70:1246–1248.
14. Newcombe H, Kennedy J, Axford S, James A: Automatic linkage of vital records. Science 1959; 130:954–959.
15. Newcombe H, Kennedy J: Record linkage. Commun Assoc Comput Mach 1962; 5:563–566.
16. Newcombe H: Record linking: the design of efficient systems for linking records into individual and family histories. Am J Hum Genet 1967; 19:335–359.
17. Arellano M, Simborg D: A probabilistic approach to the patient identification problem. Proc Comput Applications Med Care, 1981; 852–856.

ACKNOWLEDGMENTS

Initial software development support was provided by the Northern California Cancer Program (NCCP) under National Cancer Institute Grant No. 263-MD-014565. Dr. Andrew Moss, NCCP, was instrumental in arranging the cooperative agreement for use of State of California mortality files between the Vital Statistics Branch of the State of California's Department of Health Services and the Department of Epidemiology and International Health of the University of California at San Francisco. Initial mortality file conversion support was provided by the Hanford Environmental Health Foundation under US DOE Contract No. DE-AC06-76RLO 1837. Dr. Peggy Reynolds provided the data for the Japanese cohort.

APPENDIX

I. Deterministic Decision Criteria

The social security number (SSN) is the match key of choice. If the user record does not contain a SSN, an attempt is made to generate an alternative match key from the birth date, first two digits of the first name phonetic code and sex code.

A. Social Security Number

Classification as a high likelihood match requires agreement on the SSN and four out of the following six items: Last name phonetic code, first two digits of the first name phonetic code, middle name initial, birth month, birth date, and birth year. All other SSN matches are classified as low likelihood matches.

B. Alternative Match Key

User file records for which the month and year of last contact are subsequent to the month and year of death on the California State mortality file record are rejected from further processing.

1. *Males*—Classification as a high likelihood match requires agreement on the alternative match keys, first name phonetic codes, middle name initials, and birthplace codes.
2. *Females*—Classification as a high likelihood match requires agreement on the alternative match keys and two out of the following three items: first name phonetic code, middle name initial, and birthplace code.

II. Probability Linkage Decision Criteria

Weight computations are based on a modified Fellegi-Sunter[7] model. Procedures have been developed for the computation of probability linkage weights for last name, first name, middle name initial, social security number, birth month, birth day, birth year, birthplace and race. Since different weighting schemes are required for most of these variables, only a general summary of the weight computation process will be presented. For a more comprehensive presentation of probability linkage weight computation techniques, see Fellegi and Sunter[7] and Arellano[17].

A. Definition of Terms

Matched Set: The population of individuals represented by the truly matched records (true positives).

p_j^k = P (Occurrence of the jth value of variable k|A value was recorded)

e^k = P (The kth variable is recorded incorrectly among records associated with the matched set|A value was recorded)

e_T^k = P (The error-free forms of the values of the kth variable in a record comparison are different|The record comparison is associated with the matched set)

e_o^k = P (A value for the kth variable was not recorded among records associated with the matched set)

γ_j^k = The jth agreement configuration of the coded comparison of values for the kth component

B. Estimation of Probabilities

1. Estimates of the p_j^k probabilities have been derived from the California mortality files at five-year intervals. The most appropriate estimates are applied to each linkage situation.
2. Estimates of e^k and e_T^k, the "recording error probabilities", are for the most part derived from the high likelihood matches produced by deterministic linkage. A pool of estimates derived from previous experience is available to provide estimates of probabilities which cannot be derived from deterministic linkage. An estimate of e_o^k is not essential for purposes of probability linkage weight computation.

C. Weight Computation Formulas

1. *Agreement configurations:*
   $w(\gamma_j^k)$ = Probability linkage weight for the jth agreement configuration of the kth component
   $= -\log(p_j^k)+2\log(1-e^k)+\log(1-e_T^k)$ (log to the base 10)
2. *Disagreement configuration:*
   w(disagreement configuration for the kth component)

$$= \log[1-(1-e^k)^2(1-e_T^k)]-\log[1-\sum_{j=1}^{n}(p_j^k)^2]$$

3. *Missing value(s) configuration:*
   w(missing value(s) configuration for the kth component)
   $= \log[(1-(1-e_o^k)^2)/(1-(1-e_o^k)^2)]$
   $= 0.0$

D. Decision Criteria

A total probability linkage weight, obtained by summing over the component weights for each record comparison, is the basis of the probability linkage decisions. These weights determine the assignment of the corresponding comparison pairs to high likelihood nonlinkage, low likelihood linkage, or high likelihood linkage categories; previously specified lower and upper threshold values are the boundary points between these categories. Values of 3.0 for the upper threshold value (corresponding to a likelihood of 1000 to 1 that the match is a true positive) and 0.1 for the lower threshold value (corresponding to a likelihood of 1.25 to 1 that the match is a true positive) are generally used; in our experience, these values have provided a very good basis for making the necessary discriminations.

In the strict statistical sense, the summation of component weights requires the assumption that the components are independent. In practice, we have not observed any serious consequences of the violation of this assumption.

U.S. Postal Service
STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
Required by 39 U.S.C. 3685)

1A. TITLE OF PUBLICATION	1B. PUBLICATION NO.	2. DATE OF FILING
American Journal of Public Health		10/10/84
3. FREQUENCY OF ISSUE	3A. NO. OF ISSUES PUBLISHED ANNUALLY	3B. ANNUAL SUBSCRIPTION PRICE
monthly		

4. COMPLETE MAILING ADDRESS OF KNOWN OFFICE OF PUBLICATION *(Street, City, County, State and ZIP+4 Code) (Not printers)*
1015 Fifteenth Street, NW, Washington, DC 20005

5. COMPLETE MAILING ADDRESS OF THE HEADQUARTERS OF GENERAL BUSINESS OFFICES OF THE PUBLISHER *(Not printer)*
same as above

6. FULL NAMES AND COMPLETE MAILING ADDRESS OF PUBLISHER, EDITOR, AND MANAGING EDITOR *(This item MUST NOT be blank)*

PUBLISHER *(Name and Complete Mailing Address)*
American Public Health Association, 1015 15th St., NW, Washington, DC 20005

EDITOR *(Name and Complete Mailing Address)*
Alfred Yankauer, MD, MPH, 1015 15th St., NW, Washington, DC 20005

MANAGING EDITOR *(Name and Complete Mailing Address)*
William H. McBeath, MD, MPH, 1015 15th St., NW, Washington, DC 20005

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given. If the publication is published by a nonprofit organization, its name and address must be stated.) (Item must be completed.)*

FULL NAME	COMPLETE MAILING ADDRESS
American Public Health Association	1015 15th St., NW, Washington, DC 20005

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

FULL NAME	COMPLETE MAILING ADDRESS
no capitol stock	

9. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES *(Section 423.12 DMM only)*
The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes *(Check one)*

(1) ☒ HAS NOT CHANGED DURING PRECEDING 12 MONTHS
(2) ☐ HAS CHANGED DURING PRECEDING 12 MONTHS
(If changed, publisher must submit explanation of change with this statement.)

10. EXTENT AND NATURE OF CIRCULATION *(See instructions on reverse side)*	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NO. COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
A. TOTAL NO. COPIES *(Net Press Run)*	34,500	34,801
B. PAID AND/OR REQUESTED CIRCULATION 1. Sales through dealers and carriers, street vendors and counter sales	0	0
2. Mail Subscription *(Paid and/or requested)*	33,025	33,535
C. TOTAL PAID AND/OR REQUESTED CIRCULATION *(Sum of 10B1 and 10B2)*	33,025	33,535
D. FREE DISTRIBUTION BY MAIL, CARRIER OR OTHER MEANS SAMPLES, COMPLIMENTARY, AND OTHER FREE COPIES	350	252
E. TOTAL DISTRIBUTION *(Sum of C and D)*	33,375	33,787
F. COPIES NOT DISTRIBUTED 1. Office use, left over, unaccounted, spoiled after printing	1,125	1,014
2. Return from News Agents	0	0
G. TOTAL *(Sum of E, F1 and 2—should equal net press run shown in A)*	34,500	34,801

11. I certify that the statements made by me above are correct and complete
SIGNATURE AND TITLE OF EDITOR, PUBLISHER, BUSINESS MANAGER, OR OWNER
Adrienne Ash Director of Publications

PS Form 3526, July 1984