no significant differences in the level of contraceptive sales by marital status in either the urban or the rural program. Rather, the important variable was the *assistance* of the distributor's spouse in the sale of contraceptives, which was shown to be the single strongest correlate of contraceptive sales among distributors in the rural program. In short, the advantage came *not* from the mere existence of a spouse but rather from the fact that he or she assisted with sales. The findings presented here suggest that this advantage can be further enhanced by training of the spouse in contraceptive distribution.

One question that is not addressed in the current study is the cost effectiveness of this type of training. It would be of interest in future research to determine whether it is more cost effective to train the spouses of existing distributors or to identify, recruit, and train new distributors, thus increasing the total number of distribution posts. In addition, it would be of interest to follow distributor performance after training for a longer period, to determine whether the effects of this training are sustained beyond the 6 months studied in this experiment.

References

1. Foreit, J. R., Gorosh, M. E., Gillespie, D., and Merritt, C. G.: Community-based and commercial contraceptive distribution: an inventory and appraisal. Popul Rep, Series J, No. 19, March 1978.
2. Cuca, R., and Pierce, C. S.: Experiments in family planning. Lessons from the developing world. Johns Hopkins University Press, Baltimore, 1977.
3. Johns Hopkins University: Proceedings of the Conference on Cost Effectiveness in Family Planning Programs, St. Michaels, Md., Aug. 17–21, 1981. In press.
4. Elkins, H., Pineda, M. A., and Cabrera, A.: Evaluación del programa de FECOAR. APROFAM, Guatemala City, 1978. Mimeographed.
5. Bertrand, J. T., Pineda, M. A., Santiso, R., and Hearn, S.: Characteristics of successful distributors in the community based distribution of contraceptives in Guatemala. Stud Fam Plann 11 (9–10): 274–285 (1980).
6. Annis, S.: Improving family planning programs in the highlands of Guatemala. American Public Health Association report for AID/pha/C-1100. Washington, D.C., 1978.
7. Wishik, S. M., and Chen, K.: Couple years of protection: a measure of family planning program output. International Institute for the Study of Human Reproduction, Columbia University, New York, 1973.

An Example of Record Linkage Methods to Monitor Mortality and Cancer Incidence

ALICE D. STARK, DrPH
DWIGHT T. JANERICH, DDS, MPH
SUSAN K. JEREB, BA
MARGARET HOFF, ScD

The authors are with the New York State Department of Health. Dr. Stark is chief of the Chemical Information Unit, Bureau of Toxic Substance Assessment. Dr. Janerich is director, Division of Community Health and Epidemiology. Ms. Jereb and Dr. Hoff are research scientists in the Bureau of Chronic Disease Prevention.

The staffs of the Department of Motor Vehicles, Bureau of Health Statistics, and the Cancer Registry of New York State assisted in the study by supplying data.

The work was supported in part by grant No. 5 RO1 CA 19564–03 from the National Cancer Institute, Public Health Service.

Tearsheet requests to Alice D. Stark, DrPH, Bureau of Toxic Substance Assessment, New York State Department of Health, Rockefeller Plaza, Tower Bldg., Rm. 359, Albany, N.Y. 12237.

SYNOPSIS

Linkage of New York State record systems was the key strategy in a restrospective cohort study with a 24–34 year followup interval. Parents of children with anencephaly or spina bifida and matched control parents were traced to determine the parents' cancer and death experience. Birth certificates for Upstate New York for 1945–55 were the source of the study groups. This report describes the methodology employed. The New York State Health Department's Cancer Registry and vital records, the State motor vehicle license files, and city and phone directories were searched for the most recent record indicating residence in Upstate New York, cancer incidence, or death.

Among the parents of the 1,152 index children were 18,571 person-years of followup for mothers and 21,675 person-years for fathers. Among the 1,152 controls, there were 19,682 person-years of followup for mothers and 22,596 person-years for fathers. Although losses were larger than the optimal, a large proportion of the maximum possible person-years were obtained, regardless of the birth year of the index child. Patterns of loss to followup were similar for cases and controls.

Record linkage techniques are especially applicable in followup studies if the risk factor is identifiable from routinely collected information (for example, congenital neural tube defects listed on birth certificates) and the outcome is also identifiable from such records (for example, cancer registry certificate or death certificate). If the outcome is definitive, reported routinely and comprehensively, and stored on a machine-readable medium, use of a computerized record linkage design is very efficient. A major advantage of the design is that cases and controls are treated equally with respect to outcome ascertainment and followup, so that some potential biases are eliminated. Finally, the method is nonintrusive; the subjects are never contacted or interviewed. Strictly maintained confidentiality is, of course, required.

IN THE STUDY OF CHRONIC DISEASES, epidemiologists may need to examine putative risk factors that precede the onset of disease by many years. The choice of a method of study is determined by the nature of the disease and the risk factors as well as the characteristics of the population from which the study sample may be selected. A basic problem that must be confronted in any cohort study is incompleteness in the ascertainment of outcome. This difficulty may be attributed to many factors, but one of the most common is emigration of the study population. As a rule of thumb, it has been suggested that the degree of followup should be as close to 95 percent as possible (*1*) and that the same degree of completeness of ascertainment of outcome be obtained for all categories of exposure (*2*).

Perhaps more important in a cohort study is avoidance of selection bias in determining exposure and comparability of the classification of outcome (*3*).

In this report we describe the methodology used in a retrospective cohort study with a followup interval of 24 to 34 years; linkage of New York State record systems was the basic strategy employed. We also assess the magnitude and possible effects of losses to followup.

These methods were employed to investigate the possibility of an unusual familial association between anencephaly and spina bifida in children and the subsequent development of cancer in their parents. We were also interested in unusual distributions of age at death and causes of death in these parents. Evidence of an association would enhance our understanding of the etiology of cancer and could have immediate value in devising preventive measures to reduce cancer mortality through early detection of persons at high risk. The underlying biological basis for this study is the fact that carcinogenesis and teratogenesis can sometimes be caused by the same agents (for example, radiation) as well as the fact that carcinogens can sometimes act through teratogenic-like mechanisms (for example, diethylstilbestrol). Indeed, Elwood and Elwood (*4*) suggest that use of vital statistics systems, especially if mortality data are linked to other information, may clarify how risks related to anencephaly and spina bifida are modified.

Methods

The State health department's bureau of cancer control maintains all the New York State Cancer Registry data that have been collected since 1940. We were also able to use New York State vital records and to obtain extensive data from the State's department of motor vehicles. The ready availability of all these record systems, as well as the relative ease of identification of both the exposure variables (congenital neural tube defects in offspring) and the outcome variable (cancer or death in parents) led us to the use of a record linkage study design.

We used birth record data to identify parents who had had a child born with either anencephaly (N = 493) or spina bifida (N = 659) during the years 1945–55 in Upstate New York. Birth certificate data are relatively complete for these conditions because they are easily recognized at birth and diagnostic criteria have been stable for a long time (*5*). Both live-born and stillborn infants were used for case ascertainment. Only marital pairs were included in the study, and the mothers of these infants were at least 30 years old at the time of the birth. These criteria assured that, with a followup period of 24 to 34 years, the parents would be in an age range in which the incidence rates of cancer would be high enough to provide sufficient tumors for analytic purposes. Control parents were selected from adjacent birth certificates of normal infants (birth date ± 3 months of the case baby). Parent sets were matched on the following variables: mother's age ± 1 year at the time of the child's birth, parents' race (white only), and county of residence at the child's birth. The study cohort of 4,608 persons was composed

of 1,152 sets of case parents matched with an equal number of control parents. Followup was from the date of birth of the child to December 31, 1979. Cancer and death records were searched beginning with the year of the child's birth and continuing until an event was located or until December 31, 1979. Drivers' license files contained records from 1972 through 1979. We were thus able to confirm residence in New York State by this method only for those 8 years. City and telephone directories were searched beginning in 1979 and working backward until the parent was located or until the year of the birth of the child.

Searches were limited to Upstate New York (New York State exclusive of New York City) because vital records for New York City are not readily accessible in Albany and cancer reporting did not include New York City until 1973. For the years 1945–57, searches for deaths were conducted by hand only because these data have not been computerized. Deaths are listed on microfiche by year according to the Russell–Soundex System of Indexing (a semi-alphabetic code). If there were apparent matches, then the entire death certificate, stored on microfilm, was examined. The criteria used to determine positive and probable matches follow:

1. Positive matches agreed on exact spelling of name, sex, age ± 3 years, race, and any 3 of the additional identifiers (street address, occupation if male, spouse's name, birthplace, maiden name if female);
2. Probable matches agreed on exact spelling of name, sex, age ± 3 years, and any 1 of the additional identifiers cited in 1.

Both types of matches were included in our study; more than 85 percent of matches were positive.

For the years 1958–79 a computerized matching scheme was used to locate deaths. Essentially the same procedure followed for the 1945–57 records was used except that instead of hand searches of microfiche, the computer searched each year's deaths as recorded on vital records tapes and matched on Soundex. A printout with the person's name and death certificate number speeded up the process because names that did not match could be eliminated immediately. The remaining death certificates were then inspected, and the same criteria for positive and probable matches were applied. A total of 3,041,857 death records were searched.

To locate apparent cases of cancer, a computerized searching scheme was used for all years. Again, matching was based on the Soundex code. For the years before 1967, case reports of cancer are not on tape. Consequently, hand searching of records was necessary to verify the match. Criteria for determining matches in the cancer records were as follows:

1. Positive matches agreed on exact spelling of name, age ± 3 years, sex, race, and any 1 of the additional identifiers (street address, marital status, maiden name if female).
2. Probable matches agreed on exact spelling of name, age ± 3 years, sex, and race.

For the years 1967–79, a printout of the cancer registry record was produced for each matched Soundex code. Again, use of the printout greatly improved efficiency because incorrect records could be eliminated immediately. A total of 859,519 cancer records were examined.

Whenever matches—either of deaths or of cases of cancer—were found, the relevant information was abstracted. An operator spent approximately 80 hours keypunching the data. The programmer and computer time used was minimal. However, the handwork required three full-time clerks for a period of 2 years.

Therefore, a considerable effort was devoted to designing a method for determining the number of person-years that were lost to followup in the case and control groups. The name of everyone in the study population was compared via computer matching to applications or renewals for drivers' licenses in the entire State. We began with the 1979 files and worked backward until the individual had been located or the 1972 file, the earliest year for which drivers' license files existed, had been searched.

All holders of a current driver's license are in the computerized files. Licenses are valid for 4 years and, if a license is not renewed, the name remains in the computer for 7 years. The search was conducted in 1979, but it could extend backward only until 1972. A positive match agreed on exact spelling of name, race, year of birth ± 3 years, and street address. We used fewer data points for matching because fewer were available on the license files than on either the death certificates or the cancer reports. Women were searched by their maiden name as well as the married name. If a license was found, the last known residence in New York State was noted as the year of license expiration minus 4 years.

We also searched for everyone in the cohort in the city directories and telephone books, beginning with the city of residence at the time of the child's birth.

We started with the 1979 directories and worked back to the child's birth year. The most recent date and address for the person was recorded. No effort was made to recheck the death and cancer records or the records of adjacent States because of the large effort required for a small expected return. It is possible that some parents moved in and out of the State between the birth of the child and were subsequently located by the methods discussed. We are assuming that the proportion is negligible and similar for cases and controls. Thus, it was possible to compare located and unlocated case and control parents with regard to a number of variables which could affect the study's outcome.

Results

Location methods. The data shown in table 1 indicate that 73.2 percent of case mothers and 74.3 percent of control mothers were located, as were 84.4 percent of case fathers and 85.9 percent of control fathers. The greater ability to locate fathers resulted from the higher proportion of men having a driver's license in Upstate New York during the study years and the telephone directories' practice of listing only the male head of household during the study years. From these results we concluded that the cases and controls were comparable with respect to their being located or not located.

Table 2 shows the distribution of case-control pairs of mothers and fathers according to whether both, neither, and one or the other was located. These data may be useful if one wished to carry out matched pair analysis. The largest percentages of case and control mothers and fathers were located by driver's license or city or phone directory or both (table 3).

Age distribution. We examined the ages of cases and controls at the birth of the index child because age could have an effect on migration. Mothers were matched by age at the birth. The age distribution for both cases and controls follows: ages 30–34 (59 percent), ages 35–39 (33 percent), ages 40–44 (8 percent), and age 45 (.001 percent).

The distribution of fathers' ages at the child's birth was very similar for cases and controls (table 4), although only the mothers were matched. The unlocated case and control mothers were both slightly younger than the located mothers, and the unlocated case and control fathers were slightly older than the located fathers (table 5). Since the differences are not all statistically significant ($P > .05$), we concluded that age differences did not have an important effect on migration or comparability of outcome between cases and controls.

Occupation. Since occupation also contributes to mobility, we examined usual occupation of father as recorded on the child's birth certificate. Occupations were coded and categorized by use of the U.S. Department of Commerce Classified Index of Occupations, 1980, for indications that occupation could be a factor in the failure to locate individuals. It has been shown that birth certificates have omissions and inaccuracies, but there is no reason to believe that any consistent bias should exist for case or control parents' occupations as recorded on their children's birth certificates. Since all 2,304 mothers' occupations were unknown or had been recorded as housewife, we had to limit our examination to fathers.

The distribution of occupations of case and control fathers as shown in table 6 appears to be comparable. By occupational category, however, the proportion located varies substantially with a greater effect among case fathers. The range (excluding the unknown category) was 73.7 to 93 percent for cases and 81.8 to 91.6 percent for controls; except for farmers, the percent located was similar for cases and controls. Whether this variability resulted in a

Table 1. Number and percent of persons located by case and control status and sex, Upstate New York, 1945–79

	Cases		Controls	
Parent's status	Number	Percent	Number	Percent
Mothers	1,152	100.0	1,152	100.0
Located	843	73.2	856	74.3
Not located	309	26.8	296	25.7
Fathers	1,152	100.0	1,152	100.0
Located	972	84.4	990	85.9
Not located	180	15.6	162	14.1

Table 2. Number and percent of case-control pairs located by at least one method, Upstate New York, 1945–79

		Mothers		Fathers	
Case status	Control status	Number	Percent	Number	Percent
Located	located	640	55.6	846	73.4
Located	not located	202	17.5	126	10.9
Not located	located	216	18.8	145	12.6
Not located	not located	94	8.1	35	3.0
Total		1,152	100.0	1,152	100.0

Table 3. Methods of locating cases and controls, by sex, Upstate New York, 1945–79

Method and location	Mothers				Fathers			
	Cases		Controls		Cases		Controls	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Driver's license only	181	15.7	217	18.8	107	9.3	132	11.8
City or phone directory only	268	23.3	241	20.9	203	17.6	203	17.6
Death certificate only	55	4.8	47	4.1	169	14.7	156	13.5
Death certificate of spouse only[1]	45	3.9	34	3.0	7	0.6	5	0.4
Driver's license and directory	206	17.9	218	18.9	401	34.8	414	35.9
Death certificate and cancer registry	34	2.9	41	3.6	51	4.4	41	3.6
Other (includes all categories with less than 15 in each of the 4 groups)[2]	53	4.6	58	5.0	34	3.0	34	3.0
Not found by any method	310	26.9	296	25.7	180	15.6	168	14.2
Total	1,152	100.0	1,152	100.0	1,152	100.0	1,152	100.0

[1] Secondary method derived after the searching began. [2] Includes cancer registry only category as well as other combinations of search methods

Table 4. Paternal age at birth of index child, cases and controls, Upstate New York, 1945–55

Age group (years)	Cases		Controls	
	Number	Percent	Number	Percent
20–29	71	6.2	64	5.6
30–39	722	62.7	766	66.5
40–49	313	27.2	288	25.0
50–59	42	3.7	30	2.6
60–69	4	(1)	4	(1)
Total	1,152	100.0	1,152	100.0

[1] < 0.01

Table 5. Mean age and standard deviation for case and control parents at birth of the index child, by location status and sex, Upstate New York, 1945–55

Sex and location status	Number	Mean age (years)	Standard deviation
Mothers			
Cases:			
Located	843	34.2	3.4
Not located	309	33.8	3.3
Controls:			
Located	856	34.1	3.4
Not located	296	34.0	3.3
Fathers			
Cases:			
Located	972	37.2	5.6
Not located	180	37.7	7.4
Controls:			
Located	990	36.9	5.7
Not located	162	37.3	6.5

biased ascertainment of the outcome variables is not known.

Population density. An additional factor that may influence the mobility of people is whether they live in a rural or an urban area. Therefore, we examined distribution of our study population by place of residence at the birth of the index child. We used the method developed by Nasca and co-workers (*6*); they divided the State of New York into quintiles based on population density (table 7).

Nasca and co-workers (*6*) found that there was a direct association between population density and incidence of cancer for a number of anatomic sites. A statistically significant linear trend of increasing incidence with increasing population density was observed among males and females for cancers of the buccal cavity and pharynx, esophagus, bronchus and lung, stomach, and colon. For carcinomas of the liver, gallbladder, pancreas, bladder, larynx, and rectum, this association was observed only among males, while for malignant neoplasms of the brain and nervous system, only females demonstrated a statistically significant relationship between the two variables. The geographic distribution of incidence rates for the remaining sites appeared not to be related to population density (*6*). For both cases and controls, a larger proportion of those residing in densely populated areas were not located compared with those in less densely populated areas (table 7). This factor may have some influence on the number and distribution of cancers and other causes of mortality found in this study. Unequal losses to followup based on population density could

Table 6. Fathers' occupation recorded on birth certificate of the index child and proportion of cases and controls located, Upstate New York, 1945–55

Occupational category	Cases			Controls		
	Number	Percent	Percent located	Number	Percent	Percent located
Professional-technical	135	11.7	81.5	158	13.7	86.0
Managers and administrators	155	13.5	83.9	174	15.1	83.3
Sales	62	5.4	83.7	62	5.4	88.7
Clerical	57	4.9	93.0	88	7.6	89.8
Craftsmen	263	22.8	87.1	227	19.7	91.6
Operatives	161	14.0	88.2	146	12.7	87.7
Transport equipment operators	56	4.9	80.3	60	5.2	85.0
Laborers	113	9.8	78.7	77	6.7	81.8
Farmers	57	4.9	73.7	70	6.1	91.4
Service workers	52	4.5	84.6	61	5.3	82.0
Unknown	41	3.6	71.8	29	2.5	72.4
Total	1,152	100.0	...	1,152	100.0	...

Table 7. Population density of cases' and controls' place of residence at birth of index child, Upstate New York, 1945–55

Population density (persons per square mile)	Found		Not found	
	Number	Percent	Number	Percent
Cases, total	1,813	100.0	491	100.0
1–193	528	29.1	117	24.0
194–1,238	462	25.5	106	21.7
1,239–3,541	241	13.3	67	13.6
3,542–6,389	210	11.6	65	13.2
6,390–16,925	372	20.5	136	27.6
Controls, total	1,843	100.0	461	100.0
1–193	542	29.4	104	22.6
194–1,238	481	26.1	94	20.3
1,239–3,541	275	14.9	65	14.2
3,542–6,389	237	12.9	56	12.2
6,390–16,925	308	16.7	142	30.7

mask significant differences that may exist between cases and controls for selected cancer sites.

Followup. The maximum possible years of followup varied with year of birth; the range was 24–34 years. The most recent year searched was 1979. Thus, for a child born in 1945, the maximum followup was 34 years. The chart shows the ratio of the observed person-years of followup to the maximum possible person-years of followup, plotted against the year of the child's birth. For example, for the 251 babies born in 1945, the maximum followup is $34 \times 251 = 8,534$ years. The ratio plotted is the observed number of followup years $\div$ 8,534 $\times$ 100.

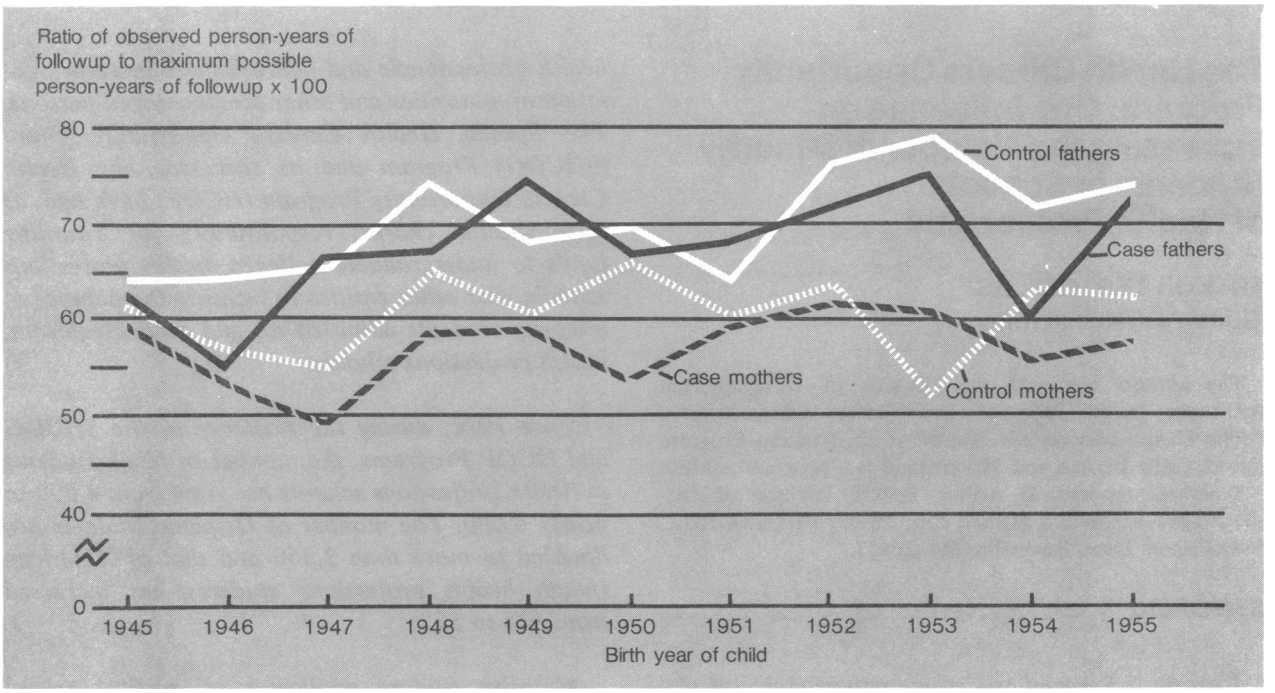
For 1,152 case mothers, we had 18,571 total person-years of followup (median, 21.9 years). Similarly for the 1,152 control mothers, we had 19,682 person-years of followup (median, 22.9 years); for the 1,152 case fathers, 21,675 person-years of followup (median, 23.2); for the 1,152 control fathers, 22,596 person-years of followup (median, 23.5 years).

Discussion

Record linkage techniques are especially applicable to followup studies when the risk factor is identifiable from a routinely collected record (for example, congenital neural tube defects reported on the birth certificate), and outcome is also identifiable from a routinely collected record (cancer registration certificate or death certificate). Indeed, record linkage may be the only feasible method when the risk factor is rare as are those in this study. When, in addition, the outcome is definitive and information about it is collected routinely and comprehensively and stored in a machine-readable medium, the use of a computerized record linkage study design is extremely efficient. Another major advantage of this design is that cases and controls receive equal treatment regarding outcome ascertainment and followup so that some potential observation biases are eliminated. Finally, the method is completely nonintrusive since the subjects are never contacted or interviewed. Strictly maintained confidentiality is, of course, required.

In Britain and Canada (*7–9*) techniques have been developed to facilitate computerized linkage of health-related records. In the United States, however, most systems of health data do not exist in a form

Ratio of observed person-years of followup to the maximum possible years of followup x 100 plotted against the birth year of the child



suitable for computerized linkage, so studies involving comparisons of records from different sources are generally difficult and time consuming. Unique individual identifiers such as social security numbers are not included in most health-related records. Because of this omission, there is a degree of uncertainty in linking records that varies according to the number and accuracy of the data included on each record (*10*). The National Death Index will be extremely useful in future retrospective studies in locating deaths occurring in another State. Perhaps the most troublesome problem associated with record linkage studies, as with all followup studies, is loss to followup. If the risk factor or outcome, or both, influences the rate at which subjects are lost, serious biases will be introduced. If too many subjects are lost, although losses of cases and controls do not differ, the reduction in the power to detect differencies may be so great as to lead to a conclusion of no effect when an effect does, in fact, exist.

As shown in the tables, we were able to achieve comparability in loss to followup of the cases and controls. Also, we were able to show that factors which could influence mobility (age, occupation, and residence) were comparable in those located and those not located. Based on an estimate of the proportion of cancer deaths found in the controls, we expect to be able to detect a difference of 15 percent with a specified $\alpha = 0.05$ and power of 0.95. Consequently, we have concluded that the record linkage method we describe can produce useful, reliable information and may, in fact, sometimes be the method of choice.

References

1. Lilienfeld, A. M.: Foundations of epidemiology. Oxford University Press, New York, 1976.
2. MacMahon, B., and Pugh, T. F.: Epidemiology: principles and methods. Little, Brown and Company, Boston, 1970.
3. Monson, R. R.: Occupational epidemiology. CRC Press, Boca Raton, Fla., 1980.
4. Elwood, J. M., and Elwood, J. H.: Epidemiology of anencephalus and spina bifida. Oxford University Press, Oxford, 1980.
5. Mackeprang, M., Hay, S., and Lunde, A. S.: Completeness and accuracy of reporting of malformations on birth certificates. HSMHA Health Rep 87: 43–49, January 1972.
6. Nasca, P. C., et al.: Population density as an indicator of urban-rural differences in cancer incidence, upstate New York, 1968–1972. Am J Epidem 112: 362–375 (1980).
7. Acheson, E. D.: Medical record linkage. Oxford University Press, London, 1967.
8. Acheson, E. D., editor: Record linkage in medicine. E. and S. Livingston, Edinburgh, 1968.
9. Smith, M. E., and Newcombe, H. B.: Automated follow-up facilities in Canada for monitoring delayed health effects. Am J Public Health 70: 1261–1268, December 1980.
10. Smith, M. E., and Newcombe, H. B.: Accuracies of computer versus manual linkages of routine health records. Methods Inf Med 18: 89–97 (1979).