# WHY ARE EPIDEMIOLOGISTS INTERESTED IN MATCHING ALGORITHMS?

Gilbert W. Beebe, National Cancer Institute

## INTRODUCTION

Both public and scientific concerns about hazards to health determine the agenda of epidemiology. The more we learn about health hazards the more there is to be learned, it seems, and the more the public comes to recognize health hazards the more it demands risk identification, risk estimates, and control measures. In recent decades new chemicals have been entering the environment at a very rapid pace. Under the Toxic Substances Control Act [1], passed in 1976, the Environmental Protection Agency (EPA) has been receiving over 1,000 pre-manufacture notices annually. There is now a list of about 30 chemicals and industrial processes recognized by the International Agency for Research on Cancer (IARC) as carcinogens for man, and another 61 thought to be probable carcinogens [2]. Another 103 are known to be carcinogenic for experimental animals, but IARC has reviewed only somewhat more than 600 chemicals and industrial processes on which there is adequate published information. I think we must assume that the carcinogens for man are far from identified and that the pace of industrial change exceeds our capacity for refined etiologic studies. We need inexpensive surveillance systems that will tell us where to look for significant hazards to health, and we need alert medical practitioners and industrial physicians to spot the unusual and unexpected [3].

The public is increasingly concerned with risks of a size that would have passed unnoticed in earlier years, risks associated with ionizing radiation, foods, drugs, toxic wastes, non-ionizing radiation, and the quality of our air and water. The MMR vaccine against measles, mumps, and rubella may cause brain damage in only one in a million vaccinees, but this risk is now sufficient to discourage manufacture of the vaccine because of the burden of litigation [4]. To identify small risks requires large samples, which in some instances may not be possible.

Ours has been aptly called an information society. Our capacity for recording, storing, transmitting, and manipulating information has been growing by leaps and bounds under the impetus of the computer revolution. I commend to you the recent (26 April 1985) computer issue of Science. The epidemiologist contributes to our understanding by bringing together for examination facts about individuals derived from different contexts. Increasingly these facts, or leads to them, are to be found in computer files. And since his unit of study is generally the individual, the epidemiologist wants to link files, which means matching, and to transfer data from files other than his own. And when he matches files he wants to be sure he is identifying the same person in each file.

In the U.S. we are experiencing a budgetary crunch. Funds for research are being reduced and staffs are being cut. The use of administrative records in research through record linkage, which means computer matching, is often the most economical way of obtaining information. For reasons of economy alone we should be looking more to record linkage as an adjunct to the more expensive procedures that we may have been following.

## THE SPECTRUM OF EPIDEMIOLOGIC INTERESTS

The following illustrations are drawn from the field of chronic disease epidemiology with which I am more familiar, but record-matching routines are also of interest to epidemiologists working in the infectious diseases.

Etiology. -- (1) The cause of multiple sclerosis remains an enigma but epidemiologists are developing a great deal of information on differentials in risk; and (2) we may be getting closer to an understanding of the role of viruses in human cancer. There are animal cancers of known viral etiology and several human cancers are now being linked to viruses.

Risk Estimation. -- (1) There is a widespread desire to know the carcinogenic risk of exposure to low doses of ionizing radiation; and (2) we are interested in the hazards of certain prescription drugs such as oral contraceptives.

Value of Early Diagnosis. -- A prime example is breast cancer. At issue is the value of a screening regimen that includes mammography.

Prevention of Disease. -- (1) Epidemiologists are involved in intervention trials to prevent coronary heart disease, as illustrated by the Multiple Risk Factor Intervention Trial (MRFIT) program of the National Heart, Lung and Blood Institute; and (2) numerous intervention trials are also being conducted against cancer; for example, the National Cancer Institute (NCI) has trials in righ-risk areas of China where micronutrients, principally vitamins, beta-carotene, and minerals, are being prescribed on a controlled basis.

Treatment. -- Breast cancer is a recent example. At issue are the extent of the surgery and the value of adjuvant drugs and radiation.

Natural History. -- Acquired Immune Deficiency Syndrome, or AIDS, is a current example.

## RECORD LINKAGE

Whether epidemiologists are working retrospectively or prospectively, in case-control or cohort mode, or are testing hypotheses or generating new ones, they are typically trying to link together, within the lives of individuals, events that are displaced in time and independently recorded. This underlies our dependence on record linkage; i.e., on matching and data-transfer. Matching requires rules of agreement, an algorithm, whether it be done manually or electronically.

Epidemiologists create their files from their own observations and from such records as are

available to them. Often they must reach out to administrative record files of large organizations such as medical care providers, insurers, state government agencies, and even the Federal agencies, for some of the facts they need to complete the history of the individual subject. It may even be necessary, for example, to go to the Internal Revenue Service (IRS) to obtain addresses needed to locate subjects for examination or interview.

Agencies with large files tailor their matching algorithms to the identifying information they characteristically deal with and understand. One cannot, for example, go to IRS for an address or to the Social Security Administration (SSA) for a mortality check, without a social security account number. The Health Care Finance Administration (HCFA), on the other hand, can search its files for addresses on the basis of a name and date of birth, after first passing the incoming file through a nominal index file that provides the SSNs essential for the address search of its Medicare file. The Veterans Administration (VA) has a very flexible approach to matching with algorithms that will work on almost any variable or combination of variables the requestor may provide. Epidemiologists often do not have any number other than the date of birth, and lack of a SSN will often keep Federal agency files beyond their reach.

Matching algorithms must depend on the identifiers available but they also reflect the scientific imagination and experience of those responsible for the programming. Newcombe has stressed the importance of experience in the manual matching of representative records as preparation for designing programs for matching by computer. He also emphasizes the value of redundancy in identifying variables when matching is involved. It was his 1959 paper, more than any other single contribution, I believe, that paved the way for technically adequate machine matching in the absence of a central ID number like the SSN [5]. With a number like the SSN it is possible to insist on an exact match. Even though the SSN is not precisely a unique number and lacks a check digit, it is nevertheless a very good number in most situations requiring linkage. If you transpose digits of your SSN in your tax return you will soon receive a query from the IRS. Names may be abbreviated to 4-6 letters of the surname if main reliance is placed on the SSN, but in other contexts the surname may be coded phonetically in New York State Identification and Intelligence System (NYSIIS) or Soundex fashion.

The investigator wants the benefit of a matching algorithm that minimizes both false positive and false negative matches but he may have no idea of the false negative rate in the absence of formal tests such as are being made on the National Death Index of the National Center for Health Statistics (NCHS) [6]. If the false positives are frequent, and in some applications NCHS algorithms have returned two false positives for each true positive match, the consumer may be hard put to evaluate the output without a weighting scheme such as Newcombe has devised.

Record linkage is now often being required on such large files that matching must be performed electronically or not at all. One cannot think of the IRS file of individual taxpayers being searched for addresses in any fashion except electronically. I am told the file contains 155 million records and takes three weeks to run. And if you want to locate a large roster of subjects under age 65 and 20-40 years after some occupational exposure, alternative sources of addresses would probably be expensive and inefficient.

## THE BACKGROUND OF MY OWN INTEREST

From the medical experience of World War II came the suggestion, by Dr. Michael E. DeBakey, the heart surgeon, that a medical research program be established to follow up the injuries and diseases of the war [7]. We both served as staff for a committee of the National Research Council (NRC) that looked into his idea and I wound up in charge of the statistical work of the group known today as the Medical Follow-up Agency of the NRC. Knowing that work with records would be a large part of the effort, one of the first persons I hired was Nona-Murray Lucke. She had been working with Dr. Halbert Dunn, then director of the Vital Statistics Division of the Bureau of the Census and originator of the term "record linkage," on his scheme for matching birth and death records at the state level [8]. Although there were Army punchcard indices to the entire medical experience of the war, the cards contained Army serial numbers but not names. A manual look-up was required to obtain the corresponding names that we could then match to the nominal VA Master Index in order to find VA claim numbers and to locate the offices having custody of the hard-copy VA files. All the linkage was manual, but usually there was enough detail beyond name and Army serial number to rule out misidentification. Identification was a problem in only about 2-4 per cent of the cases and records were unavailable in less than one percent. Starting in 1972 we benefitted from automation of the VA Master Index, now the Beneficiary Identification and Records Locator Subsystem (BIRLS) file, as well as from the automated record systems for hospital discharges and for compensation and pension status. Tape-to-tape matching has long been the rule. But the detailed medical records, not only those of World War II but also those generated today as well, are available only in hard copy.

One of the matching efforts I personally directed was a test of the completeness of VA information on the mortality of war veterans, matching known deaths obtained from NCHS against the military files in St. Louis to determine veteran status, and then submitting the resulting file intermingled with living veterans to the VA for a blind search [9]. We learned that the VA had about 95 percent of the mortality information on WW II veterans.

At the Atomic Bomb Casualty Commission (ABCC) in Japan, where I directed the epidemiologic and statistical work for some years, we followed two main samples of 55,000 and 110,000 for mortality, using the Japanese family registration system devised in 1871 [10]. Each Japanese citizen has a place of family residence (his honseki), and the city office for that place keeps a running family record, the koseki, that shows vital events for all the family members, no matter where in Japan

these events take place or where the individuals live. The koseki tells where any death certificate is retained and for the cause of death one must go there. To enter the system both the name and the honseki must be known. There is very little slippage in this system, but it is manually operated. At ABCC mortality was checked every three years on a rotational scheme that levelled out the workload.

An interesting matching problem arose in the late 1950's when I first went to Japan. The U.S.-Japan Joint Commission had created a file of about 14,000 records of its medical investigations in 1945 that were stored at the Armed Forces Institute of Pathology (AFIP) in Washington. To recapture the 1945 observations for the ABCC files we obtained blow-ups of microfilm copies retained at AFIP. For the Hiroshima portion of the sample, names were written in the Romanized fashion, not in the Japanese ideographs, or kanji. Location at the time of the bomb was given in terms of a numbered radial zone and the direction from the hypocenter, not in terms of a postal address, and age was usually given in the Japanese style which is equivalent to the western style plus one year. That is, in Japan, children are one year old at birth. Under Seymour Jablon's supervision this file was later matched to the ABCC records so that the 1945 data could be added to the ABCC files that represented largely individuals alive in 1950. About 42 percent could be matched, largely because of the considerable ancillary detail on both record sources. The false negatives could not be assessed but tests showed that the false positives probably numbered no more than 5 percent. The matching rate in Nagasaki, for which the records did contain the name in kanji and the postal address, was higher, 60 percent.

At the National Institutes of Health I have also been very much concerned with record linkage, trying to make it easier to link some of the large files of Federal agencies in the furtherance of medical research [11]. We need to restore access to the IRS address file for a broader class of investigators than just National Institute for Occupational Safety and Health (NIOSH) investigators who are concerned with occupational health, and Federal investigators studying the occupational hazards of military service, these being the privileged classes under current law. We also need to restore the kind of freedom we had before the Tax Reform Act of 1976, when SSA was willing to define industrial employment cohorts and determine their mortality. With Dr. Scheuren's help I have been trying to learn how to strengthen the Continuous Work History Sample of SSA so that it might provide some national mortality data by both industry and occupation. In addition, I'm engaged in a research project that has involved extensive matching to the files of the VA, IRS, and HCFA.

## POSSIBLE LIMITATIONS OF COMPUTER-LINKED DATA

If the only observations available to the epidemiologist derive from the linkage of administrative files, his study may be useful for screening a large experience or for developing working hypotheses, but it will probably not illuminate the meaningful aspects of exposure or define end-points precisely. If we link files as part of a larger process, e.g., to obtain addresses so that we can examine or interview subjects, or to learn that deaths have occurred and where we can find the death certificates, such limitations do not apply. Even as an index to hard-copy records, however, a large computer file may prove disappointing: recently I found that a VA diagnostic index I must depend on contains so much coding error for the cancer I am investigating that I will have to review the underlying hard-copy records for validity of diagnosis.

## LANDMARK STUDIES BASED ON MATCHING RECORDS

Any list of landmark studies is bound to be very selective and the following is further limited by my own reading and knowledge of the field:

- Framingham Heart Study [12];
- Follow-up Studies of War Injuries and Diseases, and Registry of Veteran Twin Pairs, NRC Follow-up Agency [7];
- Mancuso's Studies of Occupational Risks Based on Industrial Employment Rosters of the SSA [13];
- Studies of A-bomb Survivors in Japan [10];
- Court-Brown and Doll's Study of Ankylosing Spondylitis Patients Treated by X Ray [14];
- Dorn's Study of the Health Effects of Smoking, WW I Veterans [15];
- Oxford Record Linkage Project [16];
- Selikoff's Study of Asbestos Workers [17];
- The Mayo Clinic Studies of Olmstead County, Minnesota [18];
- The Canadian Studies of Newcombe, Statistics Canada, and the National Cancer Institute of Canada [19]; and
- The British Office of Population Surveys and Statistics Longitudinal Study [20].

## SOME OF THE LARGER COMPUTER FILES OF INTEREST TO THE EPIDEMIOLOGIST

It would be fruitless to enumerate all the files used by epidemiologists but generated independently of their own efforts. They cover a wide range of classes: employment, medical care, vital records, finance, life insurance, disability, city directories, licensing, etc. But some examples follow in Table 1.

Table 1. Some Large Files Used by Epidemiologists

| Name of File | Millions of Records |
|---|---|
| IRS, File of Individual Taxpayers | 155 |
| SSA, Master Beneficiary Record (MBR File) | 35-40 |
| HCFA, Medicare Beneficiaries | 30 |
| VA, BIRLS | 35 |
| National Archives Records Agency, "Registry" File of Military Records in National Personnel Records Center, St. Louis | 30 |
| NCHS, National Death Index | 10 |
| SSA, File of Deceased | 30 |
| California Automated Mortality Linkage System (CAMLIS) | 3.6 |
| Army WW II Hospital Diagnosis Index | 12 |

## SOME CURRENT EPIDEMIOLOGIC STUDIES TAPPING LARGE COMPUTER FILES

Apart from current studies that are already represented on our program today, some that I am particularly familiar with include:

The Johns Hopkins Study of Nuclear Shipyard Workers. -- The investigators are sampling the 700,000 nuclear shipyard worker population, stratifying on radiation dose, and seeking to relate cause of death to radiation dose, demographic characteristics, occupation, and other specific risk factors. External linkage has been established with the VA BIRLS file, the SSA MBR file, state death files, the NDI file of NCHS, and OPM files. In addition there is considerable internal file linkage to unduplicate the eight yards and to update study files with radiation dose, job classification, and the like. About 90,000 deaths have been ascertained.

Study of X-Ray Technologists. -- The NCI Radiation Epidemiology Branch has initiated a study, together with NIOSH investigators and epidemiologists of the University of Minnesota, of about 160,000 x-ray technologists in the U.S. whose exposure has long been monitored by radiation badges. Investigative interest centers not only on the carcinogenic effect of low doses of radiation, but also on the highly fractionated character of their exposure. Linkage will involve the SSA MBR file, the NDI file of the NCHS, the HCFA Medicare file, the IRS address file, and possibly other files.

Hepatitis B Virus and Primary Liver Cancer. -- In the NCI Clinical Epidemiology Branch I am doing a study with 6 VA hospitals and the Medical Follow-up Agency of the National Research Council to learn whether the contaminated yellow fever vaccine that led to 50,000 cases of acute hepatitis in the Army in 1942 has also produced excess liver cancer among the vaccinees. Record linkage has involved the Army World War II diagnostic index, the National Archives "Registry" file in St. Louis, the VA BIRLS file, the IRS address file, and the HCFA Medicare file. About 60,000 men are under study.

Study of Atomic Veterans. -- The NRC Medical Follow-up Agency is completing a study of 50,000 "atomic veterans" exposed in weapons tests in the Pacific and at the Nevada Test Site. Rosters of exposed individuals assembled by the Department of Defense were linked with the VA BIRLS file, the VA Master Index (a microfilm file), the NDI file of NCHS, and various military service files. This is another low-dose study, stimulated by the earlier finding of some excess leukemia among men exposed to the Smoky shot.

Study of Cancer from Fallout from the Weapons Tests. -- Epidemiologists at the University of Utah, under a contract with the NCI, are studying leukemia and thyroid cancer among Utah residents downwind from the Nevada Test Site, trying to establish whether fallout from the atmospheric tests of the 1950's caused excess cancer. Linkage involves two files of the Church of Jesus Christ of Latter-Day Saints (Mormons), one of about two million members registered in church censuses, the other of 400,000 deceased members. Matching also extends to the state mortality files and to the population-based cancer registry in the state of Utah.

Health Effects of Agent Orange and Service in Vietnam. -- The Centers for Disease Control have under way a complex investigation of the effect of the exposure of servicemen to Agent Orange in the Vietnam War. A sample of about 30,000 men is under study and record linkage procedures involve the IRS address file, the SSA MBR file, the VA BIRLS file, and the NCHS NDI file.

## OUTLOOK FOR THE FUTURE

I think we can expect the computer to play an ever larger role in future epidemiologic studies through record linkage. There will be no let-up in the demand of society to know its risks and to learn how to control them, and no let-up in the forward march of computer science. We can expect to find more and more data in computer files, with less dependence on them as mere indexes to hard-copy records. And matching algorithms will provide the key to the record linkage. But there are obstacles and there will be missed opportunities. Files that might have been useful for epidemiologic research may not be so because insufficient identifying information will have been collected. For the epidemiologist a critical item is often the social security number but SSA policy seems to be against its widespread use as concern for privacy and confidentiality has led to restraints on access to data that have been placed without regard for the special needs for epidemiologic information on health risks. These restraints are made doubly difficult to deal with by the fractionation of Federal statistical programs and responsibilities, each agency collecting its own statistics in support of its own narrow mission and having laws to limit access to its data. We might wish for a Statistics USA akin to Statistics Canada, but I doubt that day will ever come.

The concern for privacy stems in part from a public fear of "data banks" on the ground that they could too easily be misused. But record linkage need not imply the necessity for huge data banks. It requires only that communication be permitted between files on an ad hoc basis under restrictions that reflect the public interest in both privacy and adequacy of information.

## REFERENCES


[1] PL 94-469, Oct. 11, 1976.

[2] Tomatis, L., "Exposure Associated with Cancer in Humans," J. Cancer Res. Clin. Oncol. 108:6-10, 1984.

[3] Miller, R.W., "The Alert Practitioner As a Cancer Etiologist," Cancer Bull. 29:183-185, 1977.

[4] Medical News, "AMA Offers Recommendations for Vaccine Injury Compensation," J. Am. Med. Assn. 252:2937-2946, 1984.

[5] Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P., "Automatic Linkage of Vital Records," Science 130:954-959, 1959.

[6] Wentworth, D.N., Neaton, J.D. and Rasmussen, W.L., "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 73:1270-1274, 1983.

[7] DeBakey, M.E. and Beebe, G.W., "Medical Follow-up Studies on Veterans," J. Am. Med. Assn. 182:1103-1109, 1962.

[8] Dunn, H.L., "Record Linkage," Am. J. Public Health 36:1412-1416, 1946.

[9] Beebe, G.W. and Simon, A.H., "Ascertainment of Mortality in the U.S. Veteran Population," Am. J. Epidemiol. 89:636-643, 1969.

[10] Beebe, G.W., "Reflections on the Work of the Atomic Bomb Casualty Commission in Japan," Epidemiol. Rev. 1:184-210, 1979.

[11] Beebe, G.W., "Record Linkage and Needed Improvements in Existing Data Resources," Banbury Report 9, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory, 1981, pp. 661-673.

[12] Dawber, T.R., Kannel, W.B. and Lyell, L.P., "An Approach to Longitudinal Studies in a Community: The Framingham Study," Ann. N.Y. Acad. Sci, 107:539-556, 1963.

[13] Mancuso, T.F. and Coulter, E.J., "Methods of Studying the Relation of Employment and Long-term Illness--Cohort Analysis," Am. J. Public Health 49:1525-1536, 1959.

[14] Court-Brown, W.M. and Doll, R., "Mortality from Cancer and Other Causes After Radiotherapy for Ankylosing Spondylitis," Brit. Med. J. 2: 1327-1332, 1965.

[15] Dorn, H.F., "The Mortality of Smokers and Nonsmokers," Proc. Soc. Statist. Sec. Am. Statist. Assoc., 1958, pp. 34-71.

[16] Acheson, E.D., "Medical Record Linkage," London, Oxford Univ. Press, 1967.

[17] Selikoff, I.J., "Cancer Risk of Asbestos Exposure," In Origins of Human Cancer (Hiatt, H.H., Watson, J.D. and Winsten, J.A., eds.), Cold Spring Harbor, New York, Cold Spring Harbor Laboratory, 1977, pp.1765-1784.

[18] Kurland, L.T. and Molgaard, C.A., "The Patient Record in Epidemiology," Sci. Am. 245:54-63, 1981.

[19] Howe, G.R. and Lindsay, J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies," Comput. Biomed. Res. 14:327-340, 1981.

[20] Office of Population Censuses and Surveys, "Cohort Studies: New Developments," Studies in Medical and Population Subjects No. 25, London, Her Majesty's Stationery Office, 1973.