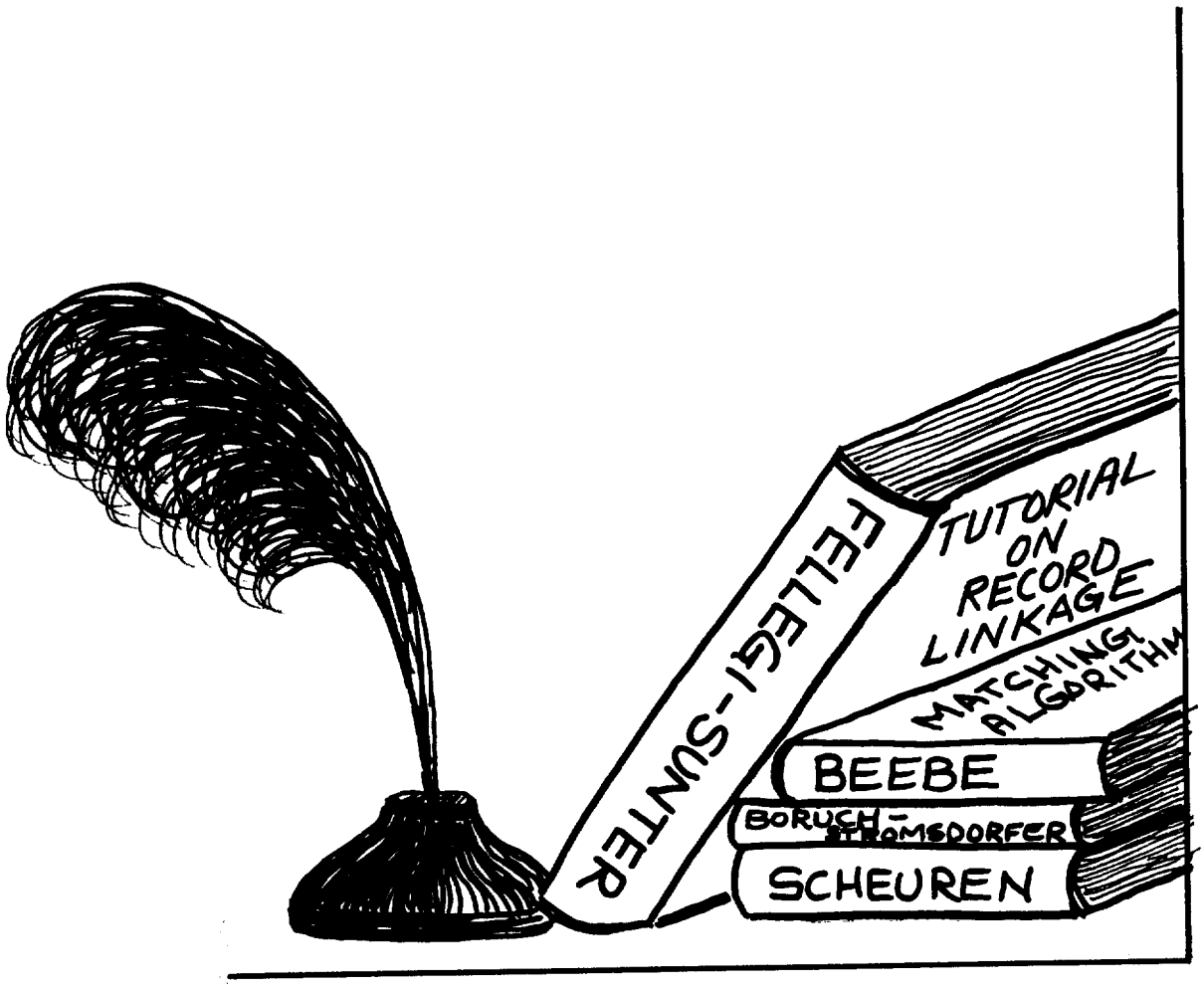
Section II:
Overview of Applications and Introduction to Theory



TUTORIAL ON THE FELLEGI-SUNTER MODEL FOR RECORD LINKAGE

Ivan P. Fellegi, Statistics Canada

EDITORS' NOTE

The following exhibits, numbered 1 to 22, were used at the Workshop on Exact Matching Methodologies (in the form of transparencies) as the basis for a presentation of the essential features and some of the consequences of the Fellegi-Sunter model and theory for record linkage. Many Workshop participants commented favorably on the exhibits and requested copies. The exhibits are presented here, without additional commentary, for the benefit of those who would like to have a convenient summary of the main points. The following chart shows the relationship between groups of exhibits and specific sections of the article, "A Theory for Record Linkage," which can be found on pages 51-78 of this volume.

Figure 1.--Exhibits for Fellegi-Sunter Article

Exhibit Numbers	Topic	Section of Article	Pages
1 to 6, 7a	Basic model and theory	2	52-57
7b, 8 to 10	Method of constructing an optimum linkage rule; consequences	2.1	54-57
11 to 14	Assumptions used in estimating weights	3.2	57-59
15 to 17	Calculation of weights, Method I	3.3.1	60-62
18	Calculation of weights, Method II	3.3.2	62-63
19, 20	Blocking	3.4	64-65
21	Choice of comparison space	3.6	66-67
22	Calculation of threshold values	3.7	67-68

Exhibit 1

Two sets of units: $A = \{a\}$, $B = \{b\}$

Vector of characteristics $\alpha(a)$, $\beta(b)$ associated with units.

$L_A = \{\alpha(a);\ a \in A\}$, $L_B = \{\beta(b);\ b \in A\}$ (lists)

$L_A \times L_B = M + U$

where $M = \{[\alpha(a), \beta(b)];\ a = b,\ a \in A,\ b \in B\}$

$U = \{[\alpha(a), \beta(b)];\ a \neq b,\ a \in A,\ b \in B\}$

$L_A \times L_B$ unmanageable.

Exhibit 2

Code results of comparing $\alpha(a)$, $\beta(b)$: $\gamma(a, b)$

$\gamma[\alpha(a), \beta(b)] = \gamma(a, b) = (\gamma^1, \gamma^2, \ldots, \gamma^k)(a, b)$

Examples: $\gamma_i = 0$ if sex is same

1 if sex is different

Exhibit 3

γ_j = 0 if name is same and is Brown

1 if name is same and is Smith

2 if name is same and is Jones

3 if name is same and not Brown, Smith, Jones

4 if name is different

5 if name is missing on either record

$\Gamma = \{\gamma(a, b)\}$: comparison space.

Exhibit 4

Linkage rule: decision regarding match status of (a, b) based on $\gamma(a, b)$

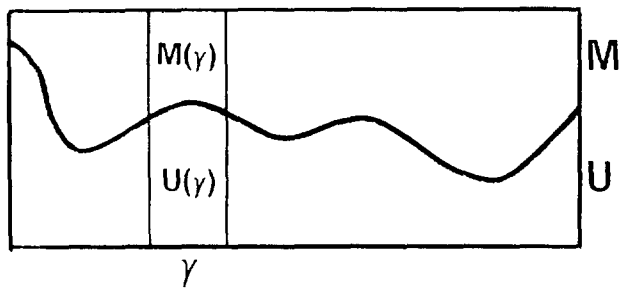
$d(\gamma) = A_1$: link (inference is "match")

$d(\gamma) = A_2$: possible link ("don't know")

$d(\gamma) = A_3$: non-link (inference is "unmatched")

Exhibit 5

$\gamma(a, b) = \gamma_0$ is a subset of $L_A \times L_B$

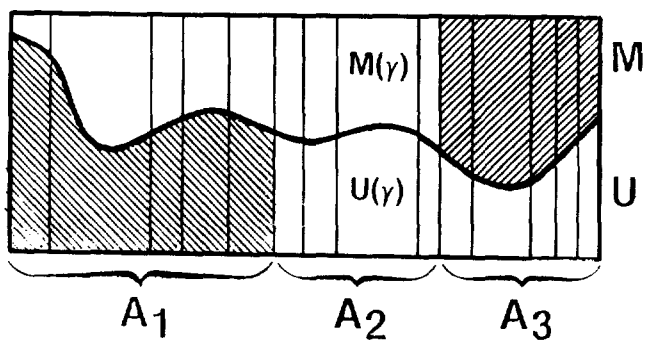


$$m(\gamma) = P\{\gamma(a, b) \mid (a, b) \in M\} = \frac{||\, M(\gamma)\,||}{||\, M\,||}$$

$$u(\gamma) = P\{\gamma(a, b) \mid (a, b) \in U\} = \frac{||\, U(\gamma)\,||}{||\, U\,||}$$

Exhibit 6

A linkage rule partitions $L_A \times L_B$:



For any $\gamma \in A_1$ all record pairs in $U(\gamma)$ are linked in error.

$$\mu = P(A_1 \mid U) = \sum_{\gamma \in A_1} u(\gamma)$$ proportion of linked record pairs in U

$$\lambda = P(A_3 \mid M) = \sum_{\gamma \in A_3} m(\gamma)$$ proportion of unlinked record pairs in M

Exhibit 7

a) **Definition:** Consider all linkage rules R on Γ with error levels μ_0, λ_0. Then R^1 is optimal if $P(A_2 \mid R^1) \leqq P(A_2 \mid R)$ for all R.

b) **Heuristic:** arrange $L_A \times L_B$ so that $m(\gamma)$ monotone decreases and $u(\gamma)$ increases. Choose A_1, A_3 to correspond to desired μ, λ. Then this linkage rule is optimal.

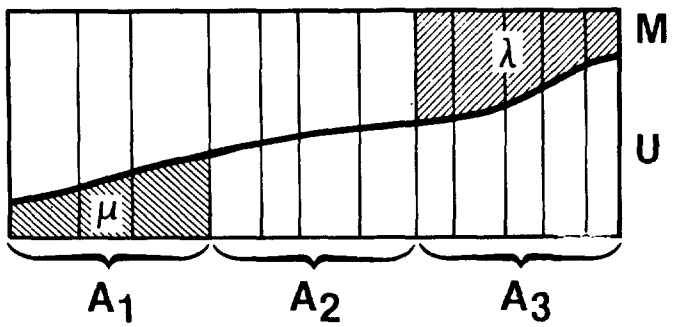


Exhibit 8

Optimal rule: order γ by decreasing values of $m(\gamma)/u(\gamma)$.

A_1 if $T_\mu \leqq m(\gamma)/u(\gamma)$

A_2 if $T_\lambda < m(\gamma)/u(\gamma) < T_\mu$

A_3 if $m(\gamma)/u(\gamma) \leqq T_\lambda$

T_μ chosen so that $\mu = \mu_0$, T_λ so that $\lambda = \lambda_0$

Likelihood ratio tests: A_1 at level μ, A_3 at level λ.

Uniformly most powerful.

Tepping's test (JASA, 1968) functionally equivalent.

Exhibit 9

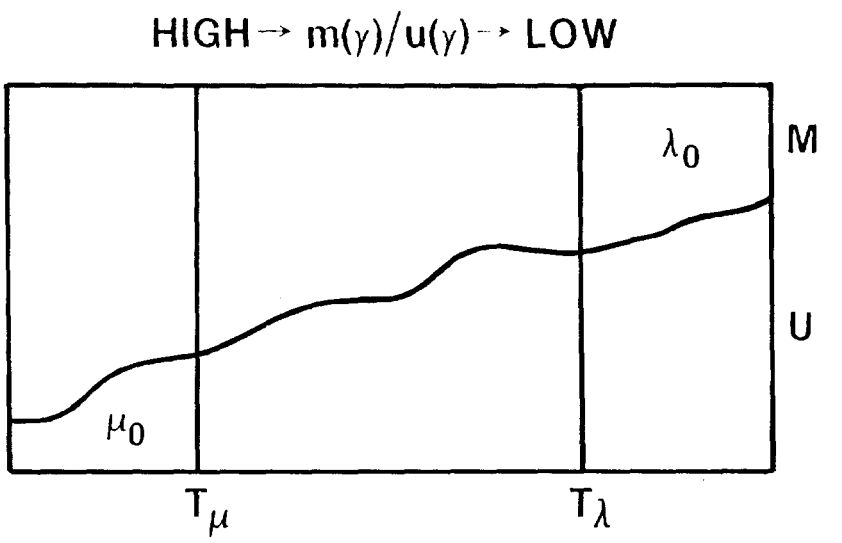


Exhibit 10

1. Trade-off between decreasing μ_0, λ_0 or A_2
2. A_2 can be eliminated if $T_\mu = T_\lambda$
3. Typically $\mu_0 << \lambda_0$ should hold. If N is the number of matched record pairs, $(N_A N_B - N)$ the number of unmatched record pairs, then condition for number of linked record pairs to be N is

$$N(1 - \lambda_0) + (N_A N_B - N)\mu_0 = N.$$

$$\text{True if } \mu_0 = \frac{N}{N_A N_B - N}\lambda_0$$

4. Randomized decision may be needed to achieve $\mu = \mu_0$, $\lambda = \lambda_0$ exactly.

Exhibit 11

Estimating m/u

If $\gamma = (\gamma^1, \gamma^2, \ldots, \gamma^K)$

γ^k has n_k values

then γ has $n_1 \cdot n_2 \ldots n_K$ values.

Simplifying assumption:

$$m(\gamma) = m(\gamma^1) \cdot m(\gamma^2) \ldots m(\gamma^K)$$

$$u(\gamma) = u(\gamma^1) \cdot u(\gamma^2) \ldots u(\gamma^K)$$

Components of γ are conditionally independent w.r. to m and u.

Exhibit 12

Matched records: Without errors, all γ^k should show "agreement". Hence independence $\longrightarrow$ errors in different ident. variables of a and b are independent.

Unmatched records: accidental agreement on one variable (e.g. name) is independent of accidental agreement on another (e.g. address).

Estimands: $m(\gamma^1), m(\gamma^2), \ldots m(\gamma^K)$ -- $n_1 + n_2 + \ldots + n_K$

(also for u).

Exhibit 13

Need care in defining γ :

$$\gamma^1 - \begin{cases} \text{agreement on female given name} \\ \text{agreement on male given name} \\ \text{disagreement on given name} \\ \text{given name missing on either record} \end{cases}$$

$$\gamma^2 - \begin{cases} \text{agreement on sex} \\ \text{disagreement on sex} \\ \text{sex missing on either record} \end{cases}$$

Accidental agreement on $\gamma^1 \rightarrow$ agreement on γ^2.
Independence might hold if first two codes of γ^1 combined.

Exhibit 14

Prefer to use log (m/u) - monotone incr. function of (m/u).

$\log (m/u) = w^1 + w^2 + \ldots + w^k$ where

$$w^k = \log [m(\gamma^k)/u(\gamma^k)]$$

We have

$$w^k \gtrless 0 \quad \text{if} \quad m(\gamma^k) \gtrless u(\gamma^k)$$

(intuitively appealing).

Similar to Newcombe-Kennedy (Communications of ACM, 1962).

Exhibit 15

METHOD 1 FOR WEIGHT CALCULATION (ILLUSTRATION)

Weights for "name" component.

Let proportions of different names in A, B and $A \cap B$ be

$p_A(i)$, $p_B(i)$, $p(i)$ $(\Sigma\, p=1)$. For simplicity:

$$p_A(i) = p_A(i) = p(i)$$

e_A, e_B: prob. of misreporting name in A, B respectively

p observable, e separately to be estimated.

Exhibit 16

$$w \text{ (agreement on jth name)} \approx \log (1/p_j)$$

- Positive
- The smaller p(j), the larger w
- I.e. large positive weight for agreement on rare characteristic

$$w(\text{agreement}) \approx \log (1/p) \quad \text{where} \quad p = \sum_j p_j^2$$

- Large for uniformly well discriminating variable
- p decreases fast if common outcomes are separated.

Exhibit 17

$$w \text{ (disagreement)} = \log \frac{e_A + e_B}{1-p}$$

- Typically negative
- The smaller the error, the larger the negative weight
- I.e. disagreement on well reported variable $\longrightarrow$ large negative weight
- E.g.: sex. Don't restrict linkage variables to high discrimination.

w (name missing on either file) = 0

- neutral contribution.

Exhibit 18. SECOND METHOD (ILLUSTRATION)

Assume only three components; each coded to two states: "agreement", "disagreement".

Conditional probabilities of "agreement" are m_h, u_h.

$$N_A N_B U_h = N m_h + (N_A N_B - N) u_h \qquad h = 1,2,3$$

where U_h: proportion of record pairs with "agreement" in h-th component.

U_h, N_A, N_B observable; N, m_h, u_h unknown.

Above 3 equations can be supplemented by other 4; all involve observable quantities + 7 unknown variables.

Solvable; generalizable; heavy dependence on independence.

Exhibit 19

Blocking

Objective: reduce number of comparisons.

Implicit assumption: comparisons <u>not</u> made are non-linked (A_3).

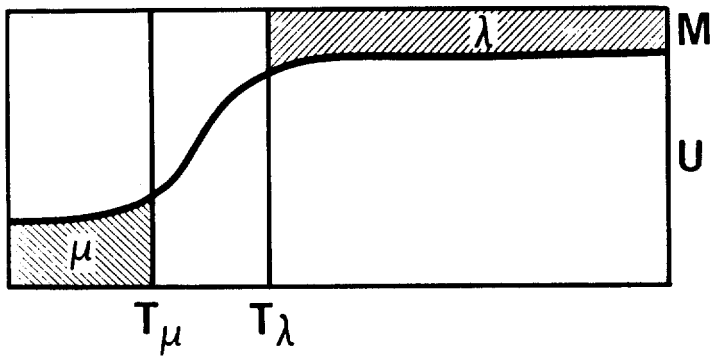


Exhibit 20. <u>IDEAL BLOCKING VARIABLE</u>

1. If a variable is such that disagreement results in very large negative weight -- corresponding e_A, e_B very small. Does not increase λ.

2. High discrimation results in maximum file blocking (comparisons restricted to records which agree on the blocking variable).

Frequent compromise: <u>coded</u> name where code is designed to reduce impact of misspellings.

<u>Additional</u> use of any well reported variable, even of low discrimination (e.g. sex), is net bonus.

Exhibit 21. CHOICE OF COMPARISON SPACE

1. How many separate values to recognize for agreement?

   Trade-off between complexity and reduction in Σp_j^2

2. How many of the variables common to both files should we use?

   Generally: the more the better.

3. w is positive for agreement, negative for disagreement almost certainly.

4. If $e_A + e_B < \frac{1}{2} < 1-p$, then each additional variable increases total weight for matched records, decreases total weight for unmatched records -- both with probability $> \frac{1}{2}$.

Exhibit 22. ESTIMATING THRESHOLDS

1. Select at random one value of <u>each</u> γ^k. Higher probabilities for high $|w|$;

2. Combine into γ; compute corresponding weight (w);

3. Repeat n times;

4. Arrange γ by decreasing w;

5. Set T_μ , T_λ as in Γ, but counting each γ with inverse of probability of selection.

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.

EXACT MATCHING OF MICRO DATA SETS IN SOCIAL RESEARCH: BENEFITS AND PROBLEMS

Robert Boruch, Northwestern University
Ernst Stromsdorfer, Washington State University

1. INTRODUCTION

The first objective here is to review some applied social research projects that have benefited from exact matching. The examples are merely illustrative but stem from a variety of disciplines.

The second objective is to discuss the negative aspects of matching. In particular, our argument is that, by espousing the opportunity to match too ardently, we may constrain or misdirect our ability to respond to other research issues and problems. An issue of special interest here is obtaining unbiased estimates of the effects of manpower projects.

The idea of matching records in the interest of science has a long pedigree. For instance, R.A. Fisher lectured at a Zurich public health congress in 1929, arguing the usefulness of public records supplemented by (and presumably linked with) family data, in human genetics research (Box, 1978, p. 237). Earlier, Alexander Graham Bell exploited geneological records, administrative records on marriages, census results and others, apparently linking some sources, to sustain his familial studies of deafness (Bruce, 1973; Bell, 1906).

2. HOW AND WHY HAS MATCHING BEEN HELPFUL

The fundamental reasons that matching has been useful do not differ appreciably from those implied by the above examples. Nor do the reasons differ much across the social and behavioral sciences. The following illustrations are taken from Boruch and Cecil (1979); unless otherwise noted, specific references are given there.

2.1 Matching to Understand Phenomena and Avoid Egregious Error

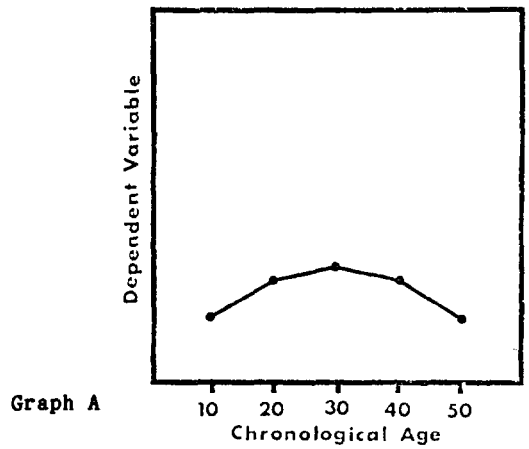
In psychology, for example, graphs of the sort used in Figure 1A were commonly used during the 1940's and 50's to describe the gradual increase in IQ with age, an IQ plateau and gradual decrease in IQ with age. The data are based on cross-sectional surveys.

The ability to match, as in linking individuals' records obtained at one point in time to those collected at another to generate longitudinal files, yielded an entirely different picture of behavior. This, given in Figure 1B, tells us that earlier declines in IQ are an artifact of cross-sectional studies and that cohort differences are important and account for the misleading interpretations of the earlier data.

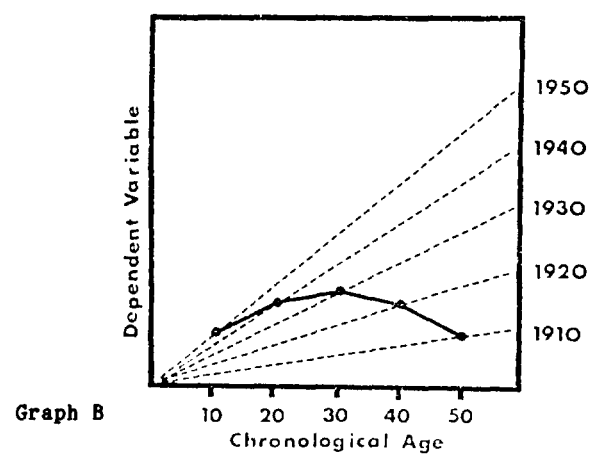
Lest you think the example confined to a quantitatively naive discipline, consider an economic example. Table 1, based on simple cross-sectional surveys, suggests that a graph similar to Type A is appropriate for earnings data as well as IQ data. Such earnings data were commonly used during the 60's to describe increases, plateau, and gradual decline in income. Table 2 gives cohort earnings obtained in longitudinal surveys, matching on individuals. It shows a different picture, one that is less dramatic and more similar to the Type B figure.

Studies that try to separate genetic and environmental influences in schizophrenia are bound to be more controversial. But they are important and worth pursuing... So, for example,

Figure 1. Confounding of Age and Cohort Differences in Cross-sectional Research.



Graph A



Graph B

From: Boruch, R.F., and Cecil, J.S. Assuring the Confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press, 1979.

Table 1.--Estimates of Mean Annual Income in Dollars for Men Aged 25-64
(Data is based on independent samples taken in 1947, 1948, and 1949.)

Year	Age			
	25-34	35-44	45-54	55-64
1947	2,704	3,344	3,329	2,795
1948	2,898	3,508	3,378	2,946
1949	2,842	3,281	3,331	2,777

From: Boruch, R.F., and Cecil, J.S. Assuring the Confidentiality of Social Research Data Philadelphia: University of Pennsylvania Press, 1979.

Table 2.--Estimates of Mean Annual Income in Dollars Over Ten-Year Intervals for Six Cohorts

Year	Ages		
	25-34	35-44	45-54
1. 1947	2,704 (1947)	5,300 (1957)	8,342 (1967)
2. 1948	2,898 (1948)	5,433 (1958)	8,967 (1968)
3. 1949	2,842 (1949)	5,926 (1959)	9,873 (1969)
Year	Ages		
	35-44	45-54	55-64
4. 1947	3,344 (1947)	5,227 (1957)	7,004 (1967)
5. 1948	3,508 (1948)	5,345 (1958)	7,828 (1968)
6. 1949	3,281 (1949)	5,587 (1959)	8,405 (1969)

Note: Each cohort was surveyed every ten years. The first cohort, for example, contains individuals who were 25-34 years of age in 1947 and had an average income of $2704; in 1967, when they were 45-54 years of age, their mean income was $8342.

From: Boruch, R.F., and Cecil, J.S. Assuring the Confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press, 1979.

Danish-U.S. collaboration supported by the National Institute of Mental Health (NIMH) has involved intensive record matching to determine how children born of schizophrenic parents fare when they are adopted and reared by non-schizophrenic, foster parents. Matching among records of hospitals, surveys, and psychiatric systems was required to execute the research. The work appears to confirm a genetic component in that incidence of schizophrenia among such children is higher than its incidence among adopted children born of nonschizophrenic parents, including children adopted by schizophrenic parents.

That use of matched records can improve scientific analysis seems clear from studies of the economic impact of education. Paul Samuelson, for example, has argued that returns on higher education are substantial. Christopher Jencks has analyzed various survey data sets to argue that the returns are marginal. Fagerlind used Swedish data that were better than data available to either Samuelson or Jencks: matching individual records from military screening; birth registries, tax registries on earnings of the respondent, census records on occupational mobility. These analyses favor Samuelson's theory.

Neither the schizophrenic study nor the Samuelson-Jencks-Fagerlind work is unambiguous, of course. There has been considerable debate about the models exploited in each. The main point is that improvements in data, notably through linkage of records from a variety of sources, can enhance the analyst's ability to explore ideas and test hypotheses. The "sources" may be additional survey panels in a longitudinal design. Or they may be administrative records that are at least as good as survey data.

2.2 Matching to Avoid Aggregation Error and Ecological Fallacy

We often compute correlations between X and Y based on aggregate data, being cautious, of course, in generalizing to the individual level. The opportunity to match individual records often gives us the opportunity to entirely avoid the problems and caution engendered by aggregation.

One of the oldest illustrations is still the most dramatic. At a particular point in time, the correlation between literacy rate and color (black vs. white) computed on the basis of nine census regions in the United States was .95. When the data are aggregated by State instead of region, the correlation becomes .77. Finally, access to individual records led to a correlation of .20.

2.3 Matching Records in Randomized Tests of Social and Education Programs

In Middlestart education programs at Oberlin College, for instance, a series of experiments was undertaken to understand whether precollege programs worked for promising but poor adolescents. The evaluators relied on randomization to assure statistically unbiased estimates of long-run program effect. They relied on records matched among surveys, high school records, and standardized precollege records to avoid the problem of low validity in student reports of grades, and to enhance the statistical power of the tests.

Randomized field experiments, designed to understand how one can increase compliance with food stamp registration rules, have been mounted by the Office of Analysis and Evaluation of the U.S. Department of Agriculture's Food and Nutrition Service (1984). These tests depend on matches of records among participant reports and records of State Employment Security agencies and the Food Stamp Agency. Results show remarkable decreases in food stamp costs and employment benefits for certain innovative approaches to compliance assurance.

Police research is relevant, too, of course. In the Minneapolis Domestic Violence Experiments,

the object was to understand how to handle domestic violence effectively, for example, immediate arrest versus referral to social services, within limits. Undertaken by the Police Foundation, the experiment involved matching among police patrolman records, precinct arrest records, and the experimenters' records. Arrest, incidentally, seems to work in the sense of reducing subsequent incidence of domestic violence (Sherman and Berk, 1984).

Motor vehicle research is pertinent to matching, too. Work done some years ago by the Insurance Institute for Highway Safety, for example, involved linking an experimenter's observations on vehicle registration, the drivers' seat belt use, and advertisements on the topic, to motor vehicle records that contained data on the drivers' residence area. The residence area match with the other information made it possible to determine how effective alternative TV commercials, directed to different areas, were in encouraging seat belt use.

Program Implementation and Validity of Reporting

The New Jersey Negative Income Tax Experiments attended to the potential problem of overpaying welfare recipients. This set a standard for validity studies in later experiments. Overpayment of benefits in such experiments was critical insofar as (a) other sources of assistance were available to participants in the experiment, and (b) they might receive such assistance illegitimately through error (welfare rules are complicated) or deceit (crime is still a bastion of the free enterprise system). All participants reported their income based on recall. Matching these reports with administrative records helped to assure reasonable implementation of the program and to assess quality of reporting.

For example, welfare audits were created to reduce or prevent the problems: these depended heavily on the experimenters' ability to match research records with records of welfare boards. Internal Revenue Service (IRS) W-2 forms were required of families and permitted comparisons between IRS-reported income and income reported to the experiment. (Underreports of income to the experiment relative to IRS appear to have been less than 15 per cent). The Social Security Administration (SSA) cooperated by taking the experimental data, matching to its own records on individuals, and providing aggregate earnings data (not individual records) to permit estimates of underreporting of earnings in the experiment (Kershaw and Fair, 1979). (The SSA comparison suggests that about 80% of families underreport to researchers by 15% or less even when they have incentives to misreport.)

In the Seattle and Denver Income Maintenance Experiments (SIME/DIME), research records were matched to public agency records on food stamp purchase, rent subsidy, and wages. The experiment produced some small surprises through evidence that public records on rent support and food stamps were less accurate than respondents' reports in the experiments, evidence that was later strengthened by independent investigation. Underreporting of wages appeared in the expected direction based on matches with IRS records (Halsey, 1980).

In the New Jersey Negative Income Tax Experiments, Mercer County Welfare Board records were used in a pilot test to determine composition, work history, and residential mobility of families that attrited from the experiment and could not be interviewed without great difficulty. More generally, the attrited families in five cities were traced through post office change-of-address cards, motor vehicle registration agencies, welfare boards, prisons, and community groups. Apparently, face-to-face interviews with former neighbors were most productive (Kershaw and Fair, 1979).

The use of administrative records to trace attriters and assess misreporting in all the income maintenance experiments is an important but underexamined topic. The experiments themselves were well run, relative to any pragmatic standard. They cover a sufficient number of sites to tantalize any scholar with an interest in regional differences in record accuracy, misreporting models and so on. Sample sizes for validity studies were small, however. This may account partly for the disinterest of scholars. Still, it is a bit distressing to some that otherwise thoughtful commentators such as Hausman and Wise (1985) fail to recognize the policy import of misreporting and the methodological contributions of randomized tests of economic programs to this area.

2.4 Matching and Testing New Ways to Elicit Information

Innovative ways to elicit information, such as randomized response, need to be tested despite their cleverness. We are unaware of any individual match studies in this arena. But studies that compare marginals or point estimates for individuals on whom both responses and archival records are available are done.

So, for example, Bradburn, Locander and Sudman found that a randomized response method worked at times to reduce response distortion on sensitive topics such as drunk driving arrests. The basis for comparison was administrative records on the same individuals, e.g., arrest records. Individual records were not matched; comparisons are based on marginal counts or averages. But matching in this and related research is possible in principle. And it may be useful insofar as it helps us to understand how response distortion varies with sensitivity of the traits that are being examined and characteristics of individual respondents.

A fascinating example of a near match study on reporting energy use to the Census Bureau was given by Tippett (1984) in recent 1984 Proceedings of the ASA. Her experiment involved encouraging utility companies to send a randomly

assigned group of individuals a statement of the year's utility bills. A randomly assigned comparison group was not sent the statement. The statements were sent prior to the 1980 census to understand whether providing such records could enhance quality of respondents' reports of utility costs to Census. Both groups overstated costs; the "primed" group overstated costs appreciably less than the control group. Again, matching could be helpful in understanding how degree of reporting error varies with the true state of the individual.

2.5 Matching Records to Understand Validity of Response and Inferential Errors

We know that error in measurement of a response variable degrades statistical power. More important, it can lead to invidious biases in covariance analyses based on fallibly measured covariates. That is, the analyses can make programs look useless when their effects are in fact slightly positive, and can make programs look harmful when indeed they are merely useless (Riecken et al., 1974). The recent work by Andersen, Kasper, Frankel and their colleagues (1979) on total survey error clarifies the effect of imperfections in observational studies generally.

The point is that understanding validity of the measures is important in applied social research, especially policy research, as well as in basic work. Matching studies undertaken in education and supported by the National Institute of Education and the National Center for Education Statistics, for instance, show that females are appreciably more accurate than males in responding to questions about their own grades and coursework, and more accurate in reporting on income and education levels of parents. There are race differences as well as gender differences in respondents' ability and willingness to furnish information. Failure to recognize these differential validities can lead to errors in understanding which programs work and for whom. Matching helps us to avoid those errors merely by showing which subgroup differences in reporting quality may account for differences in performance.

Imperfect measures of employment and occupation can produce similar biases in explanatory models of income gain and other response variables. Matching studies of the sort undertaken by Mathiowetz and Duncan (1984) in which private employer records are linked to survey records of the Panel Study on Income Dynamics are not common. But they have potential for revising ideas about error structure. Errors in retrospective reporting on employment and occupation seem to depend less on time or recency than on salience of events in a particular month (e.g., a raise) and task difficulty (e.g., a single unemployment spell vs. multiple spells). Gender and race differences in reporting error are reduced when these variables are taken into account.

3. WHEN BENEFITS OF MATCHING ARE NEGATIVE OR AT LEAST NOT SO CLEAR

Having the option to capitalize on existing records and to match so as to obtain a better file is important because the idea and the relevant technology have been so useful. For instance, the 1984 Proceedings of the ASA, Section on Survey Research Methods contains over 30 articles that concern exact matching methods or analysis or depend heavily on matching for conclusions (validation studies, capture-recapture, others). Unlike the 1984 Proceedings, the 1978 Proceedings of the same section contained no sessions on using administrative records in conjunction with surveys or on quality control of statistical systems (partly through linkage).

The Interagency Linkage Study participants --Internal Revenue Service, Census, and Social Security Administration--deserve special credit for advances in this arena. Other agencies have worked at least as vigorously and as often, however, e.g., the National Center for Education Statistics and the National Center for Health Statistics. And a good many research projects undertaken with support of the U.S. Department of Labor's Employment and Training Administration, the National Institute of Justice, the National Center for Health Services Research (and the Department of Health and Human Services more generally) have made use of matching where it has been useful and legally possible to match.

Matching is a seductive option, however. That is, we may capitalize on matching existing records to obtain estimators that are efficient and cheaply produced, but wrong. They are wrong at times partly on account of the administrative system in which matching must take place. They are wrong partly because the matched data (observational data more generally) are inappropriate despite their accessibility and ostensible relevance.

Consider a recent case, one in which the role of matching is important.

3.1 The Case at Hand

Estimating the effect of manpower employment and training programs in this country is a significant policy issue. Since 1965 or so, most estimates have been based on observational data, i.e., sample surveys. Two kinds of observational data are most relevant here--the Continuous Longitudinal Manpower Survey (CLMS) and the Current Population Survey (CPS). Both are based on large, well-designed samples. Both have been augmented by matching respondent records with social security (SSA) earnings records.

The CLMS-SSA match works as follows. The Bureau of the Census, under agreement with the Department of Labor, designs the CLMS probability sample and collects the data. The record on each individual includes identifying information and social security number. A list of respondent SSA numbers is given to the SSA which then searches

SSA files for records on the relevant individuals. The SSA records include the social security number, earnings, birth year, six letters of surname, and other bits of information. These SSA records are then given to Census for matching to the CLMS survey records under an interagency agreement that assures confidentiality of both sets of files. Census matches the records, deletes identifying information and geographic area related characteristics. The geographic data are deleted to prevent deductive disclosure.

Recently, the U.S. Department of Labor contracted for two kinds of analyses bearing on the impact of manpower programs and based on these files. In the first kind, different, well regarded contractors were asked to use such data to estimate the effects of training programs (Westat, 1984; Dickinson, et al., 1984; Bassi, et al., 1984). In the second kind of study, estimates based on observational survey data, similarly constructed, were compared to estimates yielded by randomized field experiments. In particular, the models used on CLMS and CPS data were used to construct quasi-experimental comparison groups. The performance of these comparison groups was compared to randomized control groups generated in the National Supported Work Demonstration (Fraker & Maynard, 1985).

The results of three independent analysts generating models and using them to estimate program effects based on CLMS and CPS data yielded the following results:

(a) Effects of training on earnings are positive and significant, especially for females and all post Comprehensive Employment and Training Act follow-up years (Westat, 1984, p. 61).

(b) Effects on earnings for men are not generally significant; effects on women's earnings are significant (Bassi, et al., 1984, p. xv).

(c) Effects on earnings for men tend to be significant and negative, but effects on women are positive and significant but small (Dickinsen, et al., 1984, p. xiii).

We have oversimplified here, of course. "Significance" is emphasized too much and the statements are misleadingly blunt. But the conclusions are as they appear in the final reports.

Comparing estimates of control group performance similarly constructed to estimates of control group behavior based on randomized experiments had the following results: depending on the particular model and matching strategy used, estimated effects on earnings range from minus 2000% of "true" earnings to plus 50% of "true" earnings, "true" being estimated from the randomized trial.

These results should be a bit disconcerting. They are indeed puzzling and potentially embarrassing. The Labor Department deserves praise for scholarship in disclosing the puzzle and for its political fortitude in willingness to tolerate potential embarrassment.

More to the point, what are the reasons for the discrepancies? Sampling variations may account for some of the differences. But it is not likely to account for all. In the next section, the reasons engendered by another line of argument are discussed, in the interest of understanding the strength and weakness of the argument.

3.2 Line of Argument

The critic can propose that part of the reason for discrepant results lies in relying---

(a) solely on observational data, matched or otherwise, and
(b) on models whose validity is untestable with the data at hand.

Critics who are more blunt may further suggest that the CPS, SSA, and CLMS are used because they are available and seemingly appropriate and not because they are sufficient.

Finally, the administrative system in which matching occurs demands that one give up some opportunities that should not be given up if the object is to produce good estimates of program effects.

To illuminate the contentions, consider SSA earnings matches with observational data from surveys. Problems similar to ones discussed here occur in other contexts. The material that follows is based on thoughtful reports by Bassi, et al. (1984), Dickinson, et al. (1984), and Westat (1984), that is, the producers of the estimates of manpower program effects.

State Identifiers and Areas as Missing Data

Welfare laws differ appreciably among states. These laws determine who gets welfare and how much they get. It makes sense to incorporate such data into any analysis of the way a federal employment program is used by the poor and what the impact of the program is. Local labor market information is also crucial to thoughtful analyses of why people do or do not get jobs as a consequence of programs.

Yet such information is absent from public use microdata files that are released after matching records. The result is that the economist must be content with data that are bound to generate estimates of program effect that are likely to be biased. That is, important major variables are left out of the left hand side of explanatory equations because they are deleted from public use files or remain unmeasurable variables. The incompleteness of the model is responsible for biased estimates of effect.

Why are they left out of such files? Because their inclusion will permit deductive disclosure. That is, it becomes possible to deduce the identity of anonymous respondents if

information about geographic area is supplied. The Census, for example, cannot countenance the possibility of deductive disclosure of information that it has collected, and invokes Title 13 to justify its position. Census perspective on this matter is important not only for this case: The Bureau "performs a major portion of its survey work on a reimbursable basis for other Federal agencies" (Cox, et al., p. 1, 1985). It is important as a survey agency and as a model of virtue in this respect.

Exclusion of relevant data seems to us to be the most serious consequence of our use of Census-SSA in data collection and matching. From such a matching system, we cannot produce credible estimates without the appropriate variables.

Earnings not Covered by SSA

Many public sector jobs are not covered by SSA reporting. Insofar as the employment and training program leads to jobs that are public sector and not covered, two problems occur. When earnings are a dependent variable, estimates of impact will be understated when the comparison groups jobs are more likely to be SSA covered. When earnings are used as a covariate, e.g., "prior base year," estimates of program impact will be biased because the covariate is fallible.

One way to assess the problem is by looking at interview-based earnings reports and SSA earnings, of course. Dickinson, et al. (1984) did so. They found substantial error in CLMS interview reports, e.g., 33% of CLMS respondents who said they did not work in 1977 had positive SSA earnings reported. The rate for CPS is about 10%. We still have a dilemma: SSA is clearly better than self-reports of earnings, although they are imperfect.

SSA earnings data are also truncated at both ends. For example, the maximum earnings subject to SSA tax is the maximum recorded earnings level. Dickinson, et al. (1984) examined interview earnings and SSA cap earnings to find no appreciable difference between analyses using each. i.e., estimates of program effect are about the same (p. 98).

Updatedness: A Possibly Tractable Problem

As of 1983-84, the period of DOL analyses of interest here, 1979 SSA records merged with CPS and CLMS data are incomplete. That is, not all 1979 SSA earnings for members of these samples were available. A "zero" entry for the missing data means we cannot tell how much missing data there is. Bias cannot be estimated. Still, this problem seems tractable.

Program Participation not Measured: A Possibly Tractable Problem

The CPS does not now measure participation in employment programs. Consequently, a public use file will not permit construction of a comparison group that is "uncontaminated." Among youth in the CPS comparison group, for example, it has been estimated that between 1975-78 30% entered CETA. So the contamination issue seems important. It, too, seems tractable but not without substantial effort.

Alignment Problems

According to Dickinson, et al. (1984), in Westat's analysis of the FY76 cohort, SSA earnings in calendar year 1975 were used to match individuals, despite the fact that calendar year 1975 earnings included up to six months of post-enrollment earnings for some CLMS members, (p. 35). Dickinson, et al., used calendar year cohorts rather than fiscal year cohorts. The disadvantage is in potentially missing the preprogram drop in earnings.

4. RESTATEMENT OF THE PROBLEMS AND POSSIBLE SOLUTIONS

4.1 Core Problems

There are two kinds of problems implicit in the case just presented. The first concerns reliance solely on surveys coupled to administrative records to understand relative effects of programs. Problems engendered by relying on such data affects not only efforts to estimate impact of manpower training programs, of course. They also appear in health services research, psychiatric and mental health services evaluations, assessments of court procedures, tax compliance, and police procedures (Riecken, et al., 1974). We attribute the problems partly to the seductiveness of matching and partly to the more dangerous problem of untestable models.

The second kind of problem stems from our inability to use all the data in ways that permit confidence that the analysis is statistically unbiased. Denial of access to micro-records on account of deductive disclosure affects research by Bureau of Labor Statistics (Plewes, 1985) as well as the DOL Employment and Training Administration, by the National Institute of Justice (e.g., in victimization studies), and others. The issue is also likely to affect newer statistical programs, e.g., the Survey of Income and Program Participation (David, 1984). We attribute this problem to the administrative environment in which matching technology must be exploited.

4.2 Resolving the First Kind of Problem and Exacerbating the Second

A scientifically reasonable solution to the first kind of problem is to actively experiment. That is, we need to run randomized trials of projects, project components, or project variations. The research policy option that seems worth exploring is routinely adjoining randomized experiments to the longitudinal studies and/or record files that are matched. See for instance, the Hollister, et al. (1985) report on evaluating the effectiveness of youth employment programs.

Exercising the option of randomized experiments can exacerbate the second problem, i.e., of deductive disclosure. That is, experiments generally involve a smaller number of individuals than national probability samples and more detailed information on each individual. This makes deductive disclosure easier. It also makes it difficult to adopt sampling rates as a partial index of likelihood of deductive disclosure (Cox, et al., 1985). If an agency with restrictive rules is involved in data collection then no public use tapes with sufficient detail will be released and no sensible competing analyses will be done.

Apart from the information demands of randomized experiments, the demand for microdata is increasing. Cox, et al. (1985) recognize that this increase has strong implications for Census policy on disclosure and they provide a thoughtful analysis.

4.3 Resolving the Second Problem

The possible resolutions to the disclosure problems are of at least three kinds: procedural, statutory, and empirical. The following options illustrate each.

Avoiding Restrictive Agencies

One may stay away from agencies that have data worth matching but that also have restrictive disclosure policies. Indeed, it is not hard to argue that private agencies are as capable of producing good data with equal privacy protection for the respondent and fewer constraints on the research than a government agency. The case is especially arguable for controversial topics of research such as AIDS, but it is also relevant here (Boruch, 1984).

Still, doing without micro-records from agencies such as the Census Bureau, Social Security Administration, or others, and doing without their capacity to serve as a broker for linking records from independent sources, is not an attractive prospect. We may gratuitously abandon opportunities to do socially useful and reliable research by foregoing collaboration with such agencies. So it is sensible to consider other options in addition to this one.

Proactive Change in Law and Policy

Alteration of law and more feasibly the interpretation of law is possible and seems desirable. The battles for statistical enclaves suggest, however, that this war will not be won easily, if at all. Still, sensible work has been done and some progress in clarifying issues has been made (Alexander, 1983). Assaults on Census's stewardship of Title 13 seem not to have been productive, for example (Plewes, 1985). Still, working toward legitimate reinterpretation of law seems an effort worth making, especially if more empirical research can be brought to bear on the issue of perceived risks of disclosure to populations. This brings us to the next option.

Empirical Research

Research on the role that privacy and consent have in record matching contexts seems sensible. How much the assurance of confidentiality means to respondents and how it influences the cooperation rate has received some attention from empiricists. For example, randomized field tests have been run under the auspices of the NAS Committee on National Statistics to understand whether people attend to assurances about privacy (Panel on Privacy and Confidentiality, as Factors in Survey Response, 1979). We agree with Thomas Plewes (1985) of the Bureau of Labor Statistics (BLS) in urging that more related work needs to be done.

In particular, obtaining respondent consent to disclose and link records for research purposes is an avenue for resolving deductive disclosure/confidentiality problems at Census, SSA, and elsewhere. We are aware of no good field experiments to determine effective strategies to elicit consent or their consequences. The BLS has been successful, according to Plewes, in eliciting consent for disclosure of its data to the Department of Agriculture, for instance, so that better sampling frames for forms could be developed. But this evidence is anecdotal and few hard data from controlled trials are available.

Both Cox, et al. (1985) at Census and Plewes (1985) at BLS recognize that public perceptions of government agencies are important in this context. That is, public confidence in government affects cooperation in surveys and resultant public data.

This chain of reasoning is plausible. But our agreement is a matter of intuition, not hard evidence. Moreover, the politicians' view of the idea and its implications for a bureaucracy and votes seem important. Neither the Census Bureau nor BLS (nor other agencies) can work on this tangle of issues with impunity, at least not always. Academic researchers have some responsibility to do so if they expect to have access to good data. We know of very few who are involved in such work, e.g., Flaherty, Hanis, and Mitchell (1979) in Canada, Mochmann and Muller (1979) and Damman and Simitis (1977) in Germany.

Research: Analytic

The Department of Labor's support of competing analyses, and of comparisons of the results of randomized tests to the results of nonrandomized assessments, is admirable. Research in the same spirit on matching and disclosure is warranted.

The thoughtful observer ought to admire the work by Nancy Spruill and Joe Gastwirth (1982) on microaggregation and masked data and work by George Duncan and Diane Lambert (1985) on disclosure limited dissemination. Their analysis helps to actualize a balance between privacy needs and the need to assure quality of released data. The thoughtful observer will also recognize, however, that not much work has been

done on the costs, traps, flaws, and benefits of using the suggestions of these analysts. We ought to know more about these issues. And so we ought to invest some resources routinely in the design of side studies to illuminate the limits on the utility of their work.

The importance of this matter stems partly from the fact that the effects of social programs in tax compliance, police, training, and employment effects are usually small. Expecting small effects, we should then be better able to anticipate the effects of micro-aggregation, random perturbation (contamination), random rounding, collapsing, and other strategies used to transform data so as to make it suitable for public use. All such tactics are used by the Census and other agencies to protect individual (and at times institutional) privacy (Cox, et al., 1985). But very little has been published about their implications for the validity of inferences based on analyses of such public use data.

Administrative Procedures

Suppose that we create a matching system under which public use tapes that are first expurgated or "adjusted" to reduce deductive disclosure problems are used for crude analyses. These analyses are eventually verified using the unexpurgated records by the agency that maintains the more detailed micro-records. The procedure achieves a balance between privacy concerns and scientific demands for quality in analysis.

But it demands substantial resources, i.e., a sequential system of crude analyses, based on public use tapes, followed closely by confirmatory analyses, based on within-agency analysis of micro-records. Still, the option seems worth considering especially because the procedure seems generalizable, e.g., to matching economic variables in the Survey of Income and Program Participation (David, 1984).

For example, 1976 Annual Housing Survey data on energy use were matched on geographic area to local utility company data. Census created the file. To protect against deductive disclosure, the Census adjusted the accuracy of energy use data "prior to release to guard against the possibility that the utility companies could uniquely identify individuals on the released file from their reported cost data" (Cox et al., 1985, p. 22). The adjustment involved random perturbation (that can be accommodated up to a point in analyses, given the perturbation parameters) and rounding. We are unaware of any formal benefit-cost analysis of this case. We believe that some sort of evaluation of such cases should be undertaken and published.

5. REPRISE AND CONCLUSION

There is no doubt that matching can be and has been useful in a variety of social research projects. Moreover, the analytic work on the topic by Felligi and Sunter (1969) and others is remarkable for its thoughtfulness. The technology for matching, considered apart from the matching system (organization and data), has stimulated fascinating research by academic and bureaucratic scholars. But solutions to the problem of getting the benefit of matching without reducing interpretability of data are not yet clear.

The ingeniousness of a matching algorithm is one thing. The system in which the algorithm is applied is quite another. It is clear that the administrative environment of the matching system can lead to invidious problems in analysis at the policy level. The problems lie not so much in matching technology as in other elements of the matching system: the data and rules under which it was collected, the institutional vehicle for matching and the rules governing it, and the procedures one uses to understand the errors we make based on analyses of matched data. The problems are severe enough to warrant the serious concern of applied statisticians and social scientists. Unless attention is dedicated to the matter we will do far less than we should for science, society, and the profession.

ACKNOWLEDGMENTS

The background research for this paper was supported by the National Science Foundation during 1984-85 to the Social Science Research Council (Measurement and Methodology Division, NSF) and 1985-86 to Northwestern University (RII-8418179 Ethics and Values in Science and Technology, NSF). We are indebted to Robert Pearson of the Social Science Research Council and Richard Beatty of Westat for background information. Readers should be grateful, as I am, to Theodore Clemence of the Census Bureau for graciously providing references on Alexander Bell.

REFERENCES

Alexander, L. Proposed legislation to improve statistical and research access to federal records. In R.F. Boruch and J.S. Cecil (Eds.), Solutions to ethical and legal problems in social research. New York: Academic Press, 1983, 273-292.

Andersen R., Kasper, J., Frankel, M.R. and Associates. Total survey error. San Francisco: Jossey Bass, 1979.

Bassi, L.J., Simms, M.C., Burbridge, L.C., and Betsey, C.L. Measuring the effect of CETA on youth and the economically disadvantaged. Washington, D.C.: The Urban Institute, April 1984.

Bell, A.G. The deaf. In: U.S. Department of Commerce and Labor, Bureau of the Census. Special Reports: The blind and the deaf, 1900. Washington, D.C.: U.S. Government Printing Office, 1906.

Boruch, R.F., and Cecil, J.S. Assuring the confidentiality of social research data. Philadelphia: University of Pennsylvania Press, 1979.

Boruch, R.F. Should private agencies maintain federal research data? IRB, 1984, 6(6), 8-9.

Box, J.F. R. A. Fisher: The life of a scientist. New York: Wiley, 1978.

Bruce, R.V. Alexander Graham Bell and the conquest of solitude. Boston: Little, Brown, and Company, 1973.

Cox, L.G., Johnson, B., McDonald, S.K., Nelson, D., and Vazquez, V. Confidentiality issues at the Census Bureau. Presented at the first Annual Census Bureau Research Conference. Reston, Virginia, March 20-23, 1985.

Damman, U., and Simitis, S. Bundesdatenschutzgesetz. Baden-Baden: Nomos Verlagsgesellschaft, 1977.

David, M. Discussion. Proceedings of the American Statistical Association: Social Statistics Section. Washington, D.C.: ASA, 1984, pp. 534-536.

Dickinson, K.P., Johnson, T.R., and West, R.W. An analysis of the impact of CETA programs on components of earnings and other outcomes. Menlo Park, CA: SRI International, November 1984.

Duncan, G.T., and Lambert, D. Disclosure limited data dissemination. Journal of the American Statistical Association. 1985, in press.

Fellegi, I.P., and Sunter, A.B. A theory for record linkage. Journal of the American Statistical Association, 1969, 64, 1183-1210.

Flaherty, D.G., Hanis, E.H., and Mitchell, S.P. Privacy and access to government data for research. London: Mansell, 1979.

Fraker, T., and Maynard, R. The use of comparison group designs in evaluations of employment related programs. Princeton, N.J.: Mathematica Policy Research, 1985.

Halsey, H.I. Data validation. Chapter 2 of P.K. Robins, R.G. Spiegelman, S. Weiner, and J.G. Bell (Eds.) A guaranteed annual income: evidence from a social experiment. New York: Academic, 1980, pp. 33-55.

Hausman, J.A. and Wise, D.A. (Eds.) Social experimentation. University of Chicago Press, 1985.

Hollister, R. and others (Eds.) Report of the Committee on Youth Employment Programs. Washington, D.C.: National Academy of Sciences, 1985.

Kershaw, D. and Fair, J. The New Jersey income maintenance experiment: Operations surveys, and administration, Volume I, New York: Academic Press, 1979.

Locander, W., Sudman, S. and Bradburn, N.M. An investigation of interview method, threat, and response distortion. Journal of the American Statistical Association, 1976, 71, 269-275.

Mathiowetz, N.A., and Duncan, G.J. Temporal patterns of response errors in retrospective reports of unemployment and occupation. Proceedings of the American Statistical Association: Section on Survey Research Methods. Washington, D.C.: 1984, 652-657.

Mochmann, E., and Muller, P.J. (Eds.), Data protection and social science research. Frankfurt/New York: Campus Verlag. 1979.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J.Strugnell, A. and Abbatt, J.D. Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. Computers in Biology and Medicine, 1983, 13(3), 157-169.

Panel on Privacy and Confidentiality as Factors in Survey Response. Committee on National Statistics. Privacy and confidentiality as factors in survey response. Washington, D.C.: National Academy of Sciences, 1979.

Plewes, T.J. Confidentiality principles and practice. Paper presented at the first Annual Census Bureau Research Conference. Reston, Virginia, March 20-23, 1985 (available from the author, Bureau of Labor Statistics).

Riecken, H.W., and others. Social experimentation: A method for planning and evaluating social programs. New York: Academic, 1974.

Sherman, L.W., and Berk, R.A. The specific deterrent effects of arrest for domestic assault. American Sociological Review, 1984, 49, 261-272.

Spruill, N.L., and Gastwirth, J. On the estimation of the correlation coefficient with grouped data. Journal of the American Statistical Association, 1982, 77, 614-620.

Tippett, J.A. An experimental project to improve reporting of selected costs in the 1980 census. Proceedings of the American Statistical Association: Survey Research Methods Section. Washington, D.C.: ASA, 1984, 323-328.

U.S. Department of Agriculture, Food and Nutrition, Office of Analysis and Evaluation. Food stamp work registration and job search demonstration. (Contract No. 533198085). Alexandria, VA: DOA, 1984.

Westat, Inc. Continuous Longitudinal Manpower Survey: Net impact results. Rockville, MD: Westat, April 1984.

METHODOLOGIC ISSUES IN LINKAGE OF MULTIPLE DATA BASES

Fritz Scheuren *

Data linkage offers several obvious benefits in studying the dynamics of aging. Retrospective and prospective approaches are possible. Many ad hoc epidemiological studies could serve as examples here (e.g., Beebe, 1985). Perhaps of even more importance are broad-based statistical samples composed of linked administrative records, either used alone or in conjunction with survey data (e.g., Kilss and Scheuren, 1980; Scheuren, 1983).

In general, linked administrative records, when structured longitudinally (e.g., Buckler and Smith, 1980), can be very effective in tracing changes with age in income and family relationships--including the onset of some forms of morbidity (e.g., Klein and Kasprzyk, 1983); and, with the advent of the National Death Index, mortality as well (e.g., Patterson and Bilgrad, 1985).

Survey data can be used, among other things, to explore the underlying causal mechanisms for these administratively recorded outcomes. The design challenge, of course, is how to build a data collection process which exploits the comparative advantages of both administrative and survey information.

The present paper examines settings where linkages of U.S. federal government records for individuals are feasible and of interest in the study of the dynamics of aging. Both administrative and survey records will be considered. Our focus will be on the barriers to and benefits from data linkages, with examples drawn from studies conducted using records from the Social Security Administration (SSA), the Health Care Financing Administration (HCFA), the National Center for Health Statistics (NCHS), the Bureau of the Census and, of course, the Internal Revenue Service (IRS).

Organizationally, the paper has been divided into three main sections. Structural questions (e.g., legal and procedural) in the development of a data linkage system are taken up first (Section 1). Technical issues in the matching process itself are discussed next (Section 2). The paper concludes (in Section 3) with some recommendations on areas for future study. An extensive set of references is also provided, along with some additional bibliographical citations (See Appendix A).

1. STRUCTURAL DESIGN CONSIDERATIONS

During the last several decades numerous data systems have been built by linkage techniques in an attempt, among other objectives, to study various aspects of the aged population. Some of these, like the Continuous Work History Sample, remain enormously valuable (e.g., Kestenbaum, 1985) but are no longer fully exploited because of access problems and severe resource constraints (e.g., Cartwright, 1978). Others, notably the Retirement History Survey (Irelan and Finegar, 1978), have not been continued. Many studies had an ad hoc character to begin with. While successful, they have not been repeated (e.g., The 1973 Exact Match Study, Kilss and Scheuren, 1978; the Survey of Low Income Aged and Disabled, Barron, 1978). Still other studies originally envisioned as stand-alone survey systems have not exploited available data linkage opportunities to extend their useful life beyond the point at which interviewing has stopped (e.g., the National Longitudinal Survey, Parnes, et al., 1979). What can we learn from these experiences and others that are similar--

- First, agency support for the activity has to be very strong and continuing. Social Security, which supported most of the projects listed above, has moved away from such general research efforts and shifted towards examining improvements in program operations (Storey, 1985). A sustained long-run commitment to basic research simply may not be possible in what is inherently a policy-oriented environment (President's Reorganization Project for the Federal Statistical System, 1981).

- Second, strong user support is essential. The products must have high, perceived public value, be delivered in a timely manner and with sufficient regularity to sustain continued interest. Start-up problems with the Retirement History Survey caused it some major difficulties from which it may never have been able to fully recover (Maddox, Fillenbaum, and George, 1978). The Continuous Work History Sample has, especially in recent years, been unable to sustain user interest outside of Social Security because of access issues raised by the 1976 Tax Reform Act. Also, the emphasis on employee-employer relationships, long a main feature of the Continuous Work History Sample, may not have been seen to be as important as the resource commitment required to maintain it.

- Third, start-up costs may be high for data linkage systems, especially if based in part on survey data. Linkage systems tend to be easily maintained at low cost unless

*Prepared for the Panel on Statistics for an Aging Population and presented September 13, 1985. Reprinted with permission from the National Academy of Sciences, Committee on National Statistics (to appear in their forthcoming report).

continued surveying is done; however, certain data problems, due to insufficient attention in obtaining good matching information, can cause continuing expense and difficulty at the analysis stage. Obviously also, as turned out to be the case with the Continuous Work History Sample, data quality limitations in the administrative records may necessitate considerable additional expense.

- Fourth, data linkage systems employ methods that may not be seen as entirely ethical (e.g., Gastwirth, 1986) or that have confidentiality constraints that make the systems hard to maintain as with the Retirement History Survey or hard to use as with the Continuous Work History Sample (e.g., Alexander, 1983). These controversial elements in data linkage techniques, it may be speculated, could be one of the reasons linkages to the National Longitudinal Survey (NLS) have never been attempted (despite the collection of social security numbers in the NLS).

It is only with the last of these points that we touch on risks that data linkage systems encounter, which are not also encountered to some degree in more conventional data-capture approaches. The force of these concerns will be discussed below.

Confidentiality and Disclosure Concerns

Data linkage operations bring us face-to-face with a "dense thicket" of laws, regulations and various ad hoc practices justified on heuristic grounds. There are statutory considerations which apply either to the particular statistical agencies involved or to the federal government, as a whole. These include the Privacy Act; the Freedom of Information Act; special legislative protections afforded to statistical data, for example, at the Census Bureau and the National Center for Health Statistics; and, of course, legislative protections afforded to administrative data, notably the 1976 Tax Reform Act. The paper by Wilson and Smith (1983) gives a good summary of the legal protections afforded tax data. For a more general treatment of legal issues and one which advocates change, see Clark and Coffey (1983); also see Alexander and Jabine (1978).

The regulations and practices of each federal statistical agency differ too, not only because of the different legislative statutes under which they operate, but also because of the varying approaches that they have taken in the accomplishment of their missions. Indeed, interagency data sharing arrangements almost defy description; they vary, among other reasons, depending on which agencies are sharing whose data and for what purpose. One excellent, albeit incomplete, taxonomy of current practice is found in the work of Crane and Kleweno (1985).

Despite the complexity of this topic, several general trends emerge that are worth noting:

- First, the American People are at best ambivalent about letting their government conduct linkages across data systems, specifically between different agencies and for purposes not obviously central to the missions of both agencies. For example, in a recent survey, questions were asked about the sharing of tax records with the Census Bureau, something which is a longstanding practice specifically permitted by law. Three-fourths of those surveyed did not support this use of administrative records even though an attempt was made to put the matter in a very favorable light, arguing for it on efficiency grounds. (Gonzalez and Scheuren, 1985; see also Appendix B for exact question wording).

- Second, bureaucratic practices which do not respect this general unease about linkage may need to be reexamined (e.g., Gastwirth, 1986). It is the duty, after all, of government statisticians to uphold both the letter and the spirit of the law. The whole tenor of the post-Watergate, Privacy Act and Tax Reform Act era has been to limit administrative initiatives (both big and little "a") and only to permit the expansion of access after the enactment of positive law. The failed initiative regarding Statistical Enclaves illustrates this point quite nicely. The Enclave proposal (Clark and Coffey, 1983) sought what many regarded as a degree of reasonable discretion on data linkage and data access; however, the authority requested was too broad for the current political climate. The arguments put forward in the proposed legislation's defense, for example, that it would increase efficiency and bring order to a patchwork of disparate practices, simply did not carry the day. In summary, we do not seem to be even close to a general solution on access to data for statistical purposes.

- Third, absent new legislation, many statistical agencies have begun to reexamine their traditional access arrangements and tighten still further their practices (e.g., Cox et al., 1985). For example, the use of special Census agents to facilitate linkages or to improve their subsequent analysis has been drastically curtailed resulting in a clear short-run loss in the utility to outsiders of linkage methods at the Census Bureau. On the other hand, new linkage practices have emerged from such reviews which may be superior to what otherwise might have been done. The linkage between the Current Population Survey and the National Death Index is an excellent example (Rogot, et al.,1983). Neither the Census Bureau nor the National Center for Health Statistics felt it could give up access of its data to the other agency; however, a compromise was worked out where joint access was maintained during the linkage operation and this has proved satisfactory. In fact, similar arrangements have been made successfully between the Center and the Internal Revenue Service as part of a study of occupational mortality (Smith and Scheuren, 1985b).

- Fourth, the extent to which public use files can be made available from linked data sets has been greatly curtailed because of new concerns about what is called the "reidentification" problem (Jabine and Scheuren, 1985). Simply put, this means that if enough linked data are provided in an otherwise unidentifiable (public-use) form, then each contributing agency could reidentify at least some of the linked units, almost no matter what efforts at disguise are attempted (Smith and Scheuren, 1985b). The only major exception occurs when the data made public from the contributing agencies are extremely limited (Oh and Scheuren, 1984; Paass, 1985); but then, usually, the incentives for cooperation on the part of the contributing agencies are limited as well. In practice, of course, there is almost no incentive for the contributing agencies to reidentify; thus, legally binding contractual obligations might be entered into that could stipulate that there was no such interest. Contractual guarantees, however, may not satisfy all parties to the linkage, because of the public perception issues mentioned earlier. It is conceivable, moreover, that no degree of legal or contractual reassurance would be adequate at the present time to permit the release of certain public use linked data sets--for example, those involving Census surveys linked to Internal Revenue Service information. Historically it was only the impossibility of reidentification which made the release of matched CPS-IRS-SSA public use files possible (Kilss and Scheuren, 1978).

It goes almost without saying that confidentiality and disclosure concerns pose the greatest barriers to the development of data linkage systems for studying aging. We will, however, defer to Section 3 a discussion of what might be done to deal with such issues and go on to explore the technical side of matching.

2. MATCHING DESIGN CONSIDERATIONS

This section is intended to provide a brief discussion of matching design questions that must be looked at in developing data linkage systems. We begin with some historical background and then focus specifically on "person" matches, where the social security number is a possible linking variable. Linkage systems based in part on survey information are emphasized. Analysis problems also are covered, particularly ways of estimating and adjusting for errors arising from erroneous links or nonlinks.

Historical Observations

The main theoretical underpinnings for computer-oriented matching methods were firmly established by the late nineteen sixties with the papers of Tepping (1968) and especially Fellegi and Sunter (1969). Sound practice dates back even earlier, at least to the nineteen fifties and the work of Newcombe and his collaborators (e.g., Newcombe, et al., 1959).

The Fellegi-Sunter approach is basically a direct extension of the classical theory of hypothesis testing to the problem of record linkage. A mathematical model is developed for recognizing records in two files which represent identical units (said to be matched). As part of the process there is a comparison between all possible pairs of records (one from each file) and a decision made as to whether or not the members of the comparison-pair represent the same unit, or whether there is insufficient evidence to justify either of these decisions. These three decisions can be referred to as a "link," "non- link" or "potential link."

In point of fact, Fellegi and Sunter contributed the underlying theory to the methods already being used by Newcombe and showed how to develop and optimally employ probability weights to the results of the comparisons made. They also dealt with the implications of restricting the comparison pairs to be looked at, that is of "blocking" the files, something that generally has to be done when linking files that are at all large.

Despite the early seminal work of Newcombe, Fellegi and others, ad hoc heuristic methods abound. There are many reasons for this state of affairs:

- First, until recently (and maybe even now) there have been only a handful of people whose main professional interest is data linkage. This means, among other things, that most of the applied work done in this field has been carried out by individuals who may be solving matching problems for the first time. Because the basic principles of matching are deceptively simple, ad hoc solutions have been encouraged that could be far from optimal.

- Second, statisticians typically get involved very late in the matching step, often after the files to be matched have already been created. Even when this is not the case, little emphasis may be placed on the data structures needed for linkage because of other higher priorities. Design opportunities have, therefore, been generally limited to what steps to take given files which were produced largely for other purposes.

- Third, until the late nineteen seventies good, portable, general-purpose matching software had not been widely available (e.g., Howe and Lindsay, 1981), despite some important early attempts (e.g., Jaro, 1972). Even in the presence of general-purpose software, the uniqueness of each matching environment may lead practitioners to write complex customized programs, thereby absorbing resources that might have been better spent elsewhere.

- Fourth, especially for matches to administrative records, barriers to the introduction of improved methods have existed

because cruder methods were thought to be more than adequate for administrative purposes.

- Fifth, the analysis of linked data sets, with due consideration to matching errors, is still in its infancy (Smith and Scheuren, 1985a). Qualitative statements about such limitations typically have been all that practitioners have attempted.

More will be said below concerning these issues in the context of computerized person matching.

Person Matching

Typically in a computerized matching process there are a number of distinct decision points:

- First, design decisions have to be made about the linking variables that are to be used, including the extent to which resources are expended to make their reporting both accurate and complete. (This step may be the most important but it is likely also to be the one over which statisticians have the least control, especially when matching to administrative records.)

- Second, decisions have to be made about what preprocessing will be conducted prior to linkage. Some of the things done might include correcting common spelling errors, calculating SOUNDEX or NYSIIS Codes, etc. (Winkler, 1985). Decisions about how to sort and block the files also fall here (Kelley, 1985).

- Third, decisions about the match rule itself come next. If a probabilistic approach is taken, as advocated by Fellegi and Sunter (1969), then we have to estimate a set of weights that represent the extent to which agreement on any particular variable provides evidence that the records correspond to the same person (and conversely, the extent to which disagreements are evidence to the contrary).

- Fourth, invariably there are cases where status is indeterminate regardless of the approach taken and a decision has to be made about excluding them from the analysis, going back for more information, etc.

To give some realism and specificity to our discussion, let us consider potential linkage settings in which we could bring together two files based on common identifying information: name, social security number, sex, date of birth, and address. As appropriate we will contrast the linkage as taking place either entirely in an administrative context or between survey and administrative data.

Linking Variables--The social security number (SSN) is the most important linking variable that we in the United States have for person matching purposes. SSNs were first issued so that the earnings of persons in employment covered by the social security program could be reported for eventual use in determining benefits. SSNs were also used as identifiers in state-operated unemployment insurance programs but no other major uses developed until 1961 when the Internal Revenue Service decided to use the SSN as the taxpayer identification number for individuals. Other uses by federal and state governments followed rapidly and now the social security number is a nearly universal identifier. The Privacy Act of 1974 placed restrictions on the use of SSNs but exempted those formally established prior to 1975. So far these restrictions have had only a minor impact on the widespread use of the social security number by governments and private organizations (Jabine, 1985).

The social security number is nearly a unique identifier all by itself and extremely well reported, even in survey settings, as well as on records such as death certificates (e.g., Cobleigh and Alvey, 1974; Alvey and Aziz, 1979). In survey contexts, error rates may run to 2 or 3 percent; but this depends greatly on the extent to which respondents are required to make use of records in order to provide the requested information. Typically, driver's licenses, pay stubs, and the like are excellent sources (in addition to the use of the social security card itself).

Both administrative and survey reporting of social security numbers are subject to possible mistakes in processing, but these can be guarded against by using part of the individual's surname as a confirmatory variable. For example, IRS and SSA use this method as one way of spotting keying errors.

A difficulty with current administrative approaches is that name changes (especially for females) may lead to considerable extra effort in confirming (usually through correspondence) that the social security number was indeed correct to begin with. (It is a requirement of the social security system that notification is to be made when name changes occur, but many people fail to do this until the omission is called to their attention.)

One disadvantage of the social security number is the absence of an internal check digit allowing one to spot errors by a simple examination of the number itself. At the time the social security system started in the mid-thirties, the widespread use of the SSN as an identifier was not envisioned. Indeed, there is not a one-to-one correspondence between individuals and the social security numbers they use. In some instances more than one person uses the same social security number. Historically, the most important cases of this type arose because SSN's were used by advertisers in promotional schemes. Perhaps the best known such instance is the number 078-05-1120 (Scheuren and Herriot, 1975). It first appeared on a sample social security number card contained in wallets sold nationwide in 1938. Many people who purchased the wallets assumed the number to be their own. The number was subsequently reported thousands of times by different individuals; 1943 was the high year, with 6,000 or more wage earners reporting the number as their own.

While there have been over 20 different "pocketbook" numbers, like 078-05-1120, they are probably no longer the main cause of multiple use of the same number. Confusion can arise (and go largely undetected) when one member of a family uses the number of another. Also, there are incentives for certain individuals, like illegal aliens, to simply "adopt" the social security number of another person as their own. The extent to which these problems exist is unknown, but they are believed, at least by some authorities, to be less prevalent than the opposite problem--issuances of multiple numbers to the same person (HEW Secretary's Advisory Committee, 1973).

Until 1972, applicants for SSNs were not asked if they had already been issued numbers, nor was proof of identity sought. This led to perhaps as many as 6 million or more individuals having two or more social security numbers (Scheuren and Herriot, 1975). A substantial fraction of the multiple issuances have been cross-referenced so that multiple reports for the same individual can be brought together if desired. Based on work done as part of the 1973 Exact Match Study, it appears that, despite the frequency of the problem, multiple issuances can largely be ignored unless one is looking at longitudinal information stretching back to the early days of the social security program. (In other words, people tend consistently to use only one of the numbers they have been issued.)

While the social security number is nearly ideal as a linking variable it is not always available. For example, in the Current Population Survey for adults the number is missing between 20 and 30 percent of the time (Scheuren, 1983). Evidence exists, however, from work done in connection with the Survey of Income and Program Participation, suggesting that with a modest effort the SSN missed rate can be lowered significantly, to less than 10% in Census surveys (Kasprzyk, 1983). Recent experience with death certificates shows a missed rate of about 6% for adults (Patterson and Bilgrad, 1985).

What, then, do we do when the SSN is missing or proves unusable? We are obviously forced either to seek more information or to try to make a match using the other linking variables. Now, as a rule, none of these other linking variables is unique alone and all of them, of course, are subject in varying degrees to reporting problems of their own. Some examples of the problems typically encountered are--

- Surname--As already mentioned, name changes due to marriage or divorce are, perhaps, the main difficulty. For some ethnic groups, there can be many last names and the order of their use may vary.

- Given Name--The chief problem here is the widespread use of nicknames. Some are readily identifiable ("Fritz" for "Frederick") but others are not (like "Stony" for "Paul").

- Middle Initial--People may have many middle names (including their maiden name) and the middle name they employ may vary from occasion to occasion. Often, too, this variable may be missing (Patterson and Bilgrad, 1985).

- Sex--This is generally well reported and, except for processing errors, can be relied upon. The main difficulty with this variable is that it is not always available in administrative records. (IRS does not have this variable except through the recoding of first names which simply cannot be done with complete accuracy.)

- Date of Birth--Day and month are generally well reported even by proxy respondents. Year can be used with a tolerance to good effect as a matching variable. Again, as with "sex," this item is not available on all the administrative files we are considering.

- Address--This is an excellent variable for confirming otherwise questionable links. Disagreements are hard to interpret, however, because of address changes; address variations (e.g., 21st and Pennsylvania Avenue for 2122 Pennsylvania Avenue); and, of course, differences between mailing addresses (usually all that is available in administrative files) and physical addresses (generally all that is obtained in a household survey). Recent research on this variable has been done by Childers and Hogan (1984).

Still other linkage variables could have been discussed, for example, race and telephone number. Race is a variable that is similar to sex except not nearly as well reported (unless it is recoded as black, nonblack (e.g., U.S. Bureau of the Census, 1973). Telephone numbers have problems similar to addresses and, while potentially of enormous value eventually, are not now widely available in administrative files.

Preprocessing Steps--In general, any method of standardization of identifier labels, such as names and addresses, will improve the chances of linking two records that should be linked during the actual matching process; however, it will also, to an unknown degree, result in some distortion and loss of information in the identifying data and may even increase the likelihood of designating some pairs of records as a positive link when, in fact, the pair is not a match.

Typically, for person matches to SSA or IRS information, two preprocessing steps have been undertaken: (1) to validate reported social security numbers; and (2), if missing or unusable, to search for SSNs using surname and other secondary linking variables. Both of these steps have had to be conducted largely within the existing administrative arrangements. The cost of mounting a wholly separate effort has been judged to be prohibitive. (The data sets involved are simply enormous: Social Security has roughly 300 million SSNs now issued. In recent years IRS has been processing about 100 million individual income tax returns annually, containing well over 150 million taxpayer social security account numbers.)

The "Validation Step" itself consists of two parts: first, a simple match on SSN alone is attempted; and, if an SSN is found, then secondary information from Social Security or Internal Revenue records is made available on the output computer file. Further processing then takes place so that the confirmatory matching information (names, etc.) can be examined and coded as to the extent of agreement. It is possible that this part of the current administrative procedure can be readily modified to accord with modern matching ideas. What is needed is to institute probability-based weights for the agreements (disagreements) found. At present administrators and statisticians alike simply employ a series of ad hoc rules to separate what will be considered a link from cases that have questionable SSNs (e.g., Scheuren and Oh, 1975; Jabine, 1985).

The "Search Step" is an elaborate and fairly sophisticated computerized procedure (which differs in detail at SSA and IRS). The files used are in sort; and, for the most part, the only possible links that can be looked at are cases that agree on surname. Since other blocking variables are used as well, the current administrative methods tend to be very sensitive to small reporting errors. This is believed to be true despite the fact that the computer linkage procedures go to great lengths to protect against more common reporting errors (such as those mentioned above). At Social Security they do this by systematically varying the linking information on the record for which an SSN is being searched. An extensive set of manual procedures also exists for cases where computer methods prove unsuccessful.

Unlike the "Validation Step," it may not be possible to bring the "Search Step" into full accord with modern practice. First of all, we would need to reexamine the decisions about what blocking variables to use (Kelley, 1985). Ideally we want variables that are without error themselves, or nearly so, in both sources (Fellegi, 1985) and that divide the files into blocks or "packets" of reasonably small size, within which we can look at all possible linkage combinations (e.g., Smith, 1982). Research is now underway in both agencies to find ways of improving the blocking variables, but it is unlikely that the current deterministic methods will ever be replaced by probability-based ones and for good reason. Linkage techniques for administrative purposes must be employed with high frequency in a great variety of situations and hence be extremely efficient in the use of computer time since the basic files involved are so large.

A compromise that naturally arises within the world of large computer files is to employ some form of multiple, albeit still deterministic, scheme. This is the approach taken with the National Death Index. The NDI currently employs over a dozen different combinations of matching variables. Some give a primary role to the social security number, some to the surname; still others place primary emphasis on the given name or on date of birth (Patterson and Bilgrad, 1985). Adopting the NDI approach at SSA or IRS, if feasible, might be one way to make a real advance.

Match Rules--Usually the computerized matching phase in a data linkage system consists of three steps: (1) comparisons between the linkage variables on the files being matched; (2) generation of codes which indicate the extent to which agreements exist or disagreements are present; and (3) decisions regarding the status of each comparison pair. This structure is the same, whether probability-based methods are being implemented (e.g., Howe and Lindsay, 1981) or heuristic approaches are taken (e.g., Scheuren and Oh, 1975).

- Comparison Step--In a sense, we have already discussed this step earlier. It depends heavily on what linkage variables are present; the reformatting, etc., done of those variables to facilitate comparisons; and the degree to which blocking is required because of resource or other considerations. What is desired here conceptually is to compare every record on each file with every record on the other. Blocking, of course, limits (sometimes severely) the extent to which such comparisons can be carried out. Any recoding of the linkage variables (say SOUNDEX for surname) may possibly, as we have noted, reduce the utility of this step. Generally, if resources permit, all the linking variables should be used in the computer comparisons. When this is not possible, they can still be employed later in manually settling cases where the outcome might otherwise be indeterminate. However, it almost goes without saying that manual intervention needs to be carefully limited and closely controlled. Manual matching is extremely costly and, while individual manual decisions can sometimes be better than with computer matching, usually humans lack consistency of judgment and can be distracted by extraneous information, such that they act more decisively than the facts would warrant.
- Coding Step--As a result of the comparison step, a series of codes can be generated indicating the degree of agreement which has been achieved. These agreement outcomes may be defined quite specifically, e.g., "Agrees on Surname and the value is GILFORD." They might be defined more generally: agree, disagree or unknown (the last arising because of missing information, perhaps).

  It becomes very difficult to talk about the coding step without looking ahead to the decision step and the specific approach that will be taken there. Nonetheless, some general observations can be made. Obviously, when we have, in fact, brought together records for the same person, we would like the agreement coding structure not to obscure this point. For example, to protect against trivial spelling errors, we might use the same agreement code even though there are transposition or single-character differences in the name. (The preprocessing of the files should have taken care of some of this but it may, again, be a consideration in the agreement coding itself.)

In most applications of the Fellegi-Sunter approach the assumption is made that agreement (or disagreement) on one linking variable is independent from that on any other, conditional only on whether or not the records brought together are, in fact, for the same person. To aid in making this assumption plausible, special care needs to be taken in structuring agreement codes for such variables as sex and first name, which are inherently related (Fellegi, 1985).

- Decision Step--An assessment can now be made as to the extent to which an agreement on any particular linking variable, or set of variables, constitutes evidence that the records brought together represent the same person. Conversely, an assessment can be made as to the extent to which disagreements are due to processing or reporting errors or are evidence that the records do not represent information for the same person. Typically, the records are divided into those (1) where a positive link is deemed to have been "definitely" established, (2) where a "possible" link may exist but the evidence is inconclusive, and (3) where it can "definitely" be said that no link exists.

In probability-based methods a statistical weight function is calculated to order the comparison pairs. The weights are developed by examining the probability ratio--

$$\frac{\text{Prob (result of comparison, given match)}}{\text{Prob (result of comparison, given nonmatch)}}$$

The numerator represents the probability that comparison of two records for the same person would produce the observed result. The denominator represents the probability that comparison of records for two different persons, selected at random, would produce the observed result. In general, the larger the ratio, the greater our confidence that the two records match, i.e., are for the same person.

Let us consider a particular example in which we are matching on both sex and race; where sex is always represented as either male or female and where race has been recoded black or nonblack. Further suppose the proportion of males and females is each 50% and that blacks constitute 10% of the population and nonblacks 90%. Also suppose that the chances of a reporting error on race are 1/100 and for sex 1/1000. Finally, we will assume that sex and race are independently distributed in the population and that reporting errors are independent as well.

With these stipulations and assumptions, we have the following table of possible probability or "odds" ratios, say for blacks. Usually, given the independence assumption, the probability ratio is broken up into a series of ratios, one for each agreement or disagreement, and logs are taken (to the base 2). One is now working with simple sums, such that the larger (more positive) the total, the more likely that the pair is a match; conversely, the more negative the sum, the greater the likelihood that the two records are not for the same person.

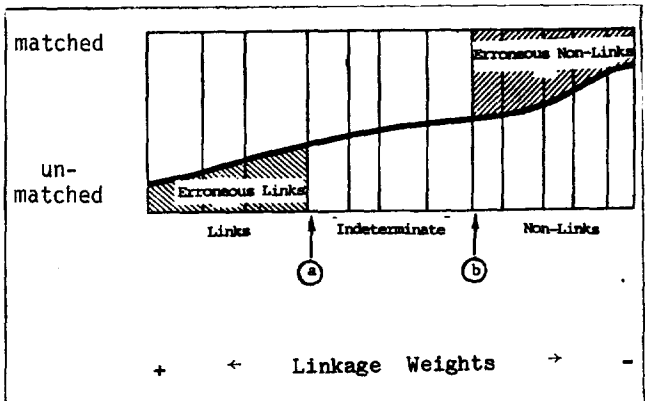
Outcome	Probability Ratio	Base 2 Log of Ratio
Race and sex agree:		
Race is black.....	197.8020	7.6279
Race is nonblack.....	2.4420	1.2881
Race agrees, sex does not:		
Race is black.....	0.1980	-2.3364
Race is nonblack.....	0.0024	-8.7027
Sex agrees, race does not.	0.1110	-3.1714
Neither agree.....	0.0001	-13.2877

See Computational Note at end of paper.

In our particular example it is only when both sex and race agree that the sum of the logs is positive. If the race is black, the log is between +7 and +8, moderately strong evidence in favor of a match. If the race is nonblack, however, the log is only slightly more than +1. As one would expect, the strongest evidence in favor of a nonmatch occurs when both race and sex disagree; for this outcome the log of the probability is about -13. (Parenthetically, it might be noted that this example illustrates nicely the fact that outcomes that are frequent in the population do not add very much to one's ability to decide if the pair should be treated as a link; but if there are disagreements on such variables and reporting is reasonably accurate, then the variable may have a great deal of power in identifying comparison pairs that represent nonlinks.)

Now it can be shown in general, as by Fellegi and Sunter (1969) or by Kirkendall (1985), that we can divide the weight distribution into three parts, as seen in figure A. The points "a" and "b" optimally divide the distribution of weights so that we can simultaneously minimize the error of accepting as a positive link cases that we should not have matched, plus minimize the error of rejecting as nonlinks cases that we should have kept. Assumptions, like independence, must be made, as a rule, and formidable computational problems exist. Nonetheless, the approach is entirely workable, especially since the development of the Generalized Iterative Record

Figure A.--Hypothetical Distribution of Linkage Weights



(adapted from Fellegi, 1985; comparison pairs above the line are matched, those below nonmatched)

Linkage System (GIRLS), which provides a state-of-the-art solution to the major computational problems (Howe and Lindsay, 1981). Other notable approaches in advanced linkage software include the work of Jaro and his collaborators (Jaro, 1985).

Indeterminate Outcomes--Virtually all computerized record linkage schemes may leave at least some cases where the status is indeterminate. Three kinds of indeterminacy might be distinguished:

- Nonlinks--Cases that were "definitely" determined by the method to have no suitable match, given the approach taken, but which might have been matched if another technique had been used (e.g., if we had employed a different set of blocking variables). The difficulty here is that, while all the potential links that get looked at may have proved inadequate, not all possible links are examined and we cannot tell the difference necessarily between a case that should have been a link and one that should not. The only way this issue can be skirted directly is in the implausible situation when the probability of a match between blocks is zero. (An indirect "solution" to this problem can be developed using contingency table ideas as will be discussed below.)

- Multiple Links--These can occur in the Fellegi-Sunter formulation; that is, there may be more than one comparison pair for a unit whose match weight or score exceeded the threshold for acceptance. In some cases, these many-to-one links might be appropriate but, usually, a further step has to be taken to select "the best" one. This problem also can occur with some frequency in administrative contexts and with the National Death Index. Manual resolution is usually the approach taken, especially if further information is going to be sought or is available to help make the selection. Jaro (1985) offers a computerized transportation algorithm to solve multiple linkage problems. His approach is most effective when all the linking information has already been computerized and when there are contention problems in the linkages, that is, "n" records on one file are matching "m" records on another. Smith and Scheuren (1985a) suggest ways of carrying through the statistical analysis using all the links.

- Potential Links--This type may be the largest form of indeterminacy. These are the cases that fall in the middle area in figure A. The usual advice, resources permitting, is to collect more information to resolve the match status. If statistical estimates are to be made, and the resources needed to seek further information are not available, the potential links may be treated as nonlinks and a survey-type nonresponse adjustment may be made (Scheuren, 1980). It is possible, also, to consider keeping some of the potential links and then conducting the analysis, with an adjustment being made for mismatching (Scheuren and Oh, 1975).

Often, the difficulty with indeterminate cases can be traced back to a design flaw in the data linkage system. For example, not enough linking information may have been obtained on one or both files to assure uniqueness. Maybe the degree of redundancy in the identifiers was insufficient to compensate completely for the reporting errors. In an administrative context, the linkage process may be so constrained for operational reasons that, even if there are sufficient linkage items, they cannot be brought fully to bear.

Analysis Issues

Statements about the nature of the matching errors are typically provided in data linkage studies; generally, however, there is no real attempt to quantify the implications of matching errors for the specific inferences being drawn. Data linkage systems, like other survey-based or sample-based techniques, need to be "measurable" and to be structured to be as robust as possible in the face of departures from underlying assumptions. What can be done to achieve this is a separate and sizable subject (Smith and Scheuren, 1985a). For our present purposes it may be enough to sketch some of the issues and indicate general lines of attack.

- Linkage Documentation--Documentation should routinely be provided which tabulates the results of the match effort along dimensions that turned out to be important in the analysis. A distribution of the weights would be one example, perhaps shown for major subgroups. If a public-use file is being created, then the match weight might be placed in the file along with summary agreement codes, so that secondary analysts can "second-guess" some of the decisions made. Providing potential links, at least near the cut-off point, is another example of good practice. Most of the above, by the way, were part of the documentation and computer files made available from the 1973 Exact Match Study (Aziz, et al., 1978).

- Adjusting for Nonlinks--It is generally worthwhile to consider reweighting the linked record pairs actually obtained to adjust for failures to completely link all the proper records to each other (Scheuren, 1980). Conventional nonresponse procedures can be followed (Oh and Scheuren, 1983). Imputation strategies are also possible, but may be less desirable because they tend to disturb the estimated relationships across the two files being brought together (Oh and Scheuren, 1980; Rodgers, 1984). An important problem in this adjustment process, however conducted, is in being able to estimate whether a link should have occurred. Sometimes, by the nature of the problem, we know all the records should have been linked. In other cases (Rogot et al., 1983), one of the key things we are interested in is, in fact, the linkage

rate. Elsewhere (Scheuren, 1983; Smith and Scheuren, 1985a), we have advocated a capture-recapture approach to this estimation problem. Such an approach, in the presence of blocking, will actually allow us to improve the links obtained, as well as make it possible to measure the extent to which our best efforts still lead to erroneous nonlinks. Capture-recapture ideas are well described in the literature (e.g., Bishop et al., 1975; Marks et al., 1974). Here we will only indicate the application.

If we employ more than one set of blocks and keep track for each blocking procedure whether we would have found (and linked) the case in every other blocking scheme, then for any subpopulation of linked records we can construct the usual 2^n table, where we look at the link/nonlink status for each blocking (with "n" being the number of separate blocking schemes). To estimate the number of records not caught by any scheme, three or more sets of blocks are recommended; otherwise, the assumptions made may be unrealistically strong. (The National Death Index, or NDI, already employs many more than this, as we have noted earlier.) For best results the blocks need to be as independent functionally and statistically as is possible, given the linkage information. (Improvements in the current NDI would be recommended here, but these seem to be coming in any case.) Application of these ideas in an IRS or SSA context seems worthy of study (Scheuren, 1983), although the expense of developing such an approach, say at SSA, may never be incurred unless there were a compelling administrative need.

- Adjusting for Mismatches--In most linkage systems practitioners have operated in what they considered to be a conservative manner with regard to the links they would accept. Sometimes this may have meant heavy additional expense in obtaining more information or the risk of seriously biasing results by leaving out a large number of the potential links. In any event, further research is needed on how to apply more complex analytic techniques that take explicit account of the mismatch rate, possibly by use of errors-in-variable approaches where the mismatch rate is estimated, e.g., as in Scheuren and Oh (1975), so that a correction factor can be derived. We must also attempt to find ways of estimating the mismatch rate that make weaker assumptions than those made in most Fellegi-Sunter applications. (Some further ideas on this are found in Smith and Scheuren, 1985a).

In summary, the main issues in the analysis of linked data sets are that, at a minimum, we need to examine the sensitivity of the results to the assumptions made in the linkage process. Where possible, we need to quantify uncertainties in the results; specifically, indeterminacies in the linkages should translate into wider confidence intervals in the estimates. To achieve these goals we need to bring in techniques from other areas of statistics and apply them creatively to linked data sets. Examples here include information theory, error-in-variable approaches and contingency table (capture-recapture) ideas.

3. SOME CONCLUSIONS AND AREAS FOR FUTURE STUDY

In this paper we have dealt with the topic of data linkage in abroad conceptual framework, using examples from recent practice. It is appropriate now to draw out the implications of the point of view expressed for studies of aging and to use that summary as a basis for recommending further research.

Overall Perspective

We have argued elsewhere that the potential for the statistical use of data linkage systems is truly enormous (e.g., Kilss and Scheuren, 1980; Jabine and Scheuren, 1985). The suggestion has even been made that data linkages among administrative records (with some supplementation) might eventually replace conventional censuses in the United States (Alvey and Scheuren, 1982). Such ideas are not new, certainly not to Europeans, where many developed nations have been rapidly moving in this direction (e.g., Pedfern, 1983). Indeed some countries, like Denmark (Jensen, 1983), may have "already arrived."

In the United States there has been some reluctance and resistance to accepting the inevitability of such a future. Grave concerns have been expressed (Butz, 1985) about moving too fast or in the wrong way. After all, while Denmark has succeeded in its efforts, other countries (notably West Germany) have encountered major problems which did grave damage to their statistical programs.

In view of what has happened elsewhere and, especially, given the current state of public opinion, we would caution that any planned use of data linkage systems be grounded firmly in existing practice and not be based on new legislation designed to expand on what it is currently possible to do. On the other hand, it is important to conceptually integrate what is now possible with what might be possible ten or twenty years from now. Some further observations are--

- First, if a data linkage approach is going to be taken, it should be a necessary means, not just a sufficient one, for achieving some required specific purpose. It is simply not enough to argue the need for data linkage on efficiency grounds.

- Second, the linkage should be seen as important by all the cooperating agencies and part of their mission. It is simply not enough that the law can be interpreted to permit such linkages. Positive law, and indeed social custom, must exist which encourages the research, at least in broad outline (Cox and Boruch, 1985).

- Third, strong continuing user support is essential if a long-term basic research effort is to be successful. Program agencies cannot be relied on for really long-run undertakings without this support. Opportunity costs are simply too high. If the linkage system is to be placed in a statistical agency, user involvement is, again, essential (from the outset, if possible). Without strong user involvement, statistical agencies will tend to emphasize continuity of measurement over relevance (while program agencies tend to the reverse).
- Fourth, cost considerations suggest that most data linkage systems be based on, or augment, an existing survey or administrative system. Further, maintenance costs should be low so that in the long run most of the resources can be focussed on exploiting the analytic potential of the system.

- Fifth, access to the results of the linkage system must be basically open not only to the primary user(s), but to secondary users as well. Ways to solve the "reidentification" problem must be built into the undertaking from the beginning and firmly rooted in the best statistical practice.

Still other considerations come to mind, such as adequate physical security during the linkage operation and minimizing the risks by removing identifiers from working files as soon as possible (Kilss and Scheuren, 1978; Steinberg and Pritzker, 1967; Cox and Boruch, 1985; and Flaherty, 1978).

Many ad hoc efforts have succeeded without strictly adhering to one or more of the above; nonetheless, if one is working towards a future which encompasses still more data linkages, it is essential that the strategy taken be absolutely sound and above reasonable reproach.

Potential Data Systems Deserving Further Study

Within the framework just given, there seems to be a clear need to intensively examine the potential of particular data linkage systems to answer certain questions. We will illustrate this point by looking at one of the most pressing areas in the United States where better data are needed -- this is on our rapidly growing aged population. Even if we confine ourselves to this single area, many subsidiary issues must be addressed. For example, where are the greatest gaps: in data on health, general demographic information, financial data, or the extent to which federal programs provide support? In what follows, there has been no attempt to answer this question. To do so, we would go well beyond the scope of the present paper. Instead, there is a discussion of four data linkage environments that, depending on the answer to the question, may warrant further study. Special emphasis has been placed on the limitations of working in each of these settings and of the role that a strong outside user might play in overcoming those limitations.

Social Security and Health Care Financing Administrations -- The Social Security (SSA) and Health Care Financing Administrations (HCFA) are unlikely to take the lead in building and maintaining general purpose statistical data linkage systems, in part because of a reduced emphasis on basic and applied research. Nevertheless, the program-oriented statistical activities of these agencies will continue to give them an important role in data linkage efforts which are consistent with agency missions. The potential at SSA and HCFA for providing improved sources of statistics on the aging population depends on the extent to which they are able to: (1) maintain major in-house data linkage efforts, like the Continuous Work History Sample (e.g., Buckler and Smith, 1980) and the Medicare Statistical System (U.S. Health Care Financing Administration, 1983); (2) continue to sponsor or co-sponsor periodic or ad hoc surveys; and (3) cooperate in linkage studies sponsored elsewhere (for example, in the Survey of Income and Program Participation or in the Health Interview Survey) if they are in support of the agencies' missions.

However, these efforts would need to be coupled with strong outside user support. At SSA and HCFA, there may be a particularly pressing need for outside users to aid in the resumption of some form of public release of subsets, at least, of the administrative samples now being employed almost solely for in-house purposes.

Internal Revenue Service -- It seems pointless to speculate upon the degree to which interagency data linkages can or should take place involving Internal Revenue Service (IRS) data. Formidable statutory barriers narrowly limit access to tax records and, even when the legal requirements can be met, many other agencies, notably the Census Bureau, feel they would be unable to engage in a cooperative study because of concerns about public perception. American social customs, particularly concerns about "Big Brother," stand as nearly insurmountable obstacles in the short run.

It is possible, though, to use IRS records essentially all by themselves as a basis for studying the aged population. This may seem surprising because the statistical program of the Internal Revenue Service is not looked at typically as a source of such information. Certainly the Statistics of Income publication series has focused very little on the aged, and then mainly through the use of the age exemption to identify taxpayers 65 years or older (e.g., Holik and Kozielec, 1984). Broader-based research has been possible through occasional linkages between the IRS's Individual Income Tax Model File and Social Security information. In a few cases, these linkages have resulted in public-use files (DelBene, 1979). What has not been done is to look at the aging population longitudinally, although this is fairly

straightforward, at least back to 1972. Furthermore, with the recent addition of complete SSA year-of-birth information to IRS files, it will be possible to routinely study age cohorts by means other than the age exemption. It is also noteworthy of mention that linkages between IRS files and the recently instituted National Death Index have just been successfully instituted (Bentz, 1985).

Tax returns probably represent the single best source of financial information and could, therefore, prove of value in studying the aging process. There are, however, three main limitations to their use:

- First, the income data, while of exceedingly high quality (relative to surveys), are incomplete since certain nontaxable incomes have been omitted (e.g., tax-exempt bond interest and welfare payments). Until recently, social security benefits were unavailable but they are now potentially taxable (beginning with 1984).

- Second, the population coverage of income tax returns is incomplete. In fact, only about half the population ages 65 years or older show up as taxpayers on income tax returns. Again, recent changes have a bearing here since information documents, notably Forms 1099 from Social Security, are filed with the Internal Revenue Service for all social security beneficiaries. This change permits an expanded population concept that could be essentially complete for the aged population.

- Third, the tax return is exceedingly awkward as a unit of analysis for some purposes since it does not always conform to conventional family and household concepts (Irwin and Herriot, 1982). It is possible though, using information documents like Forms W-2 (for wages), Forms W-2P (for private pensions), and Forms 1099 (for social security payments, dividend, interest, etc.), to develop approximate financial profiles of virtually all individuals aged 65 or older. (Major gaps would exist, of course, for supplemental security income recipients and recipients of veterans disability benefits.) There does not appear to be much hope in inferring changes in lifestyles directly from the current IRS information, although the proposed addition of dependent social security numbers could lead to real progress (Alvey and Scheuren, 1982).

Depending on its extent, the cost of maintaining an IRS data linkage system to study aging could be quite modest. Public-use files are possible; but, as with the Social Security and Health Care Financing Administrations, strong outside support would be needed.

National Center for Health Statistics -- Recent changes (Sirken and Greenberg, 1983) at the National Center for Health Statistics suggest that the Center may be assuming a leading role in sponsoring data linkage systems. Naturally and appropriately, the focus of these systems will be quite narrow, looking almost solely at health concerns. The National Health Interview Survey (HIS), involving about 40,000 households annually, appears to be the Center's main survey vehicle for the approach it is planning to take. Continued periodic matching to Medicare records seems planned (Cox and Folsom, 1984) and, of course, the National Death Index can be expected to be fully exploited (Patterson and Bilgrad, 1985). Still other linkage efforts are underway (e.g., Johnston, et al., 1984) which, taken together, suggest that the Center is pursuing a coherent, fully integrated approach, both among its surveys and towards needed vital record systems.

When the social security number question was added to the HIS a few years ago, it was largely for matching to the National Death Index. Great care initially was given to securing informed consent from respondents before obtaining the information. This approach proved tedious and expensive. Now the social security number question is simply asked without much explanation; and, only if requested, are reasons given for why the information needs to be obtained (see Appendix C). Response rates are quite high, about 90%, and it appears that the HIS may constitute a major vehicle for a successful data linkage approach to studying aging. Concerns exist about the reidentification problem, but exactly how the Center will deal with this factor is unclear.

Bureau of the Census -- Historically, the Census Bureau has played a major role in federal data linkage systems involving surveys, sometimes as the sole sponsor (e.g., Childers and Hogan, 1984), but often as a partner in conducting a particular study (e.g., as with Social Security, Bixby, 1970). Much of this work has focussed on the Current Population Survey (Kilss and Scheuren, 1978). Of more promise in future studies of aging has been the development of the Survey of Income and Program Participation (SIPP), which has as one of its design elements the notion that data linkages would be attempted, at least to Social Security information (Kasprzyk, 1983). SIPP, which may settle down to a sample size of about 30,000 households annually, is certainly of sufficient size and scope to look at many general demographic, financial and program related questions concerning aging. The SSN reporting rate is on the order of 90%; hence, the needed resources to "perfect" the linkage (and the analysis problems resulting from faulty or incomplete linkage) should be entirely manageable. Oversampling is possible for particular subgroups (e.g., those aged 65 or older); however, unfortunately, SIPP, like the HIS, is confined to the noninstitutional population and for studies of the very old it may not be suitable alone.

Two difficulties exist with SIPP that further research may resolve. First is the extent to which informed consent is being obtained when the social security number is being secured (SIPP's approach is similar to that in the HIS--see Appendix D). Related to this concern, of course, is the extent to which such consent is

felt to be needed. The second issue, and one that seems exceedingly troublesome to the Census Bureau, is the "reidentification" problem. (Briefly stated, the reidentification problem is particularly acute where linkage is concerned, because the cooperating agencies might have enough data on the linked file to reidentify virtually all of the individuals linked.)

The Census Bureau appears to be searching for a solution that involves either simply not releasing public-use files of linked data or releasing public-use files where only very limited linked data have been provided and some kind of masking technique has been employed to prevent reidentification. Given these restrictions, it must be said, there seem to be real difficulties in concluding that there are sufficient benefits to outside users of a SIPP-based data linkage system. Some further comments on this dilemma and ways a general research program could address it are given below.

General Issues Deserving Further Study

Further research is needed on a wide range of data linkage issues, both structural and technical. Four, in particular, stand out from the rest and deserve special attention: ethical and legal concerns, public perception questions, finding solutions to the reidentification problem, and finally, analysis issues in the presence of matching errors.

Ethical concerns such as those raised by Gastwirth (1986) seem to need a more specific answer than they have been given so far (e.g., as by Dalenius, 1983). What might be done is to obtain some data directly bearing on how respondents actually think about data linkage. We could approach this in a way similar to the earlier study by the Committee on National Statistics concerning confidentiality guarantees (Committee on National Statistics, 1979). Within the context of current survey efforts in HIS and SIPP it might be extremely valuable to know how often respondents ask for clarification before providing social security numbers and to code the cases accordingly so we can look at differential refusal rates, for example. Again, exactly what is said (by respondents and interviewers) typically when respondents do ask? Legal and procedural issues abound here, too. For example, how long, even assuming informed consent, can the consent be treated as binding? Social Security practices with outside researchers (when they obtain consent to gain access to individual records) is to treat the consent as binding potentially only once; thus, requests for information on the same subjects may require a renewal of the consent. Signed consent agreements are also required of outside researchers. Such a requirement has never been imposed, say, in Census Bureau surveys, but should it be? If it were, what would be the costs of such a practice in interview time, reduced response, and cooperation generally?

Public perception concerns deserve to be examined in depth. To what extent are we already violating the public's sense of the social customs within which statisticians are supposed to work? The public opinion polling results reported in Gonzalez and Scheuren (1985) need to be followed up. It does not seem defensible simply to speculate about whether this or that approach to data linkage would be acceptable to the public. While we can never use opinion polling to answer all the many specific issues that exist here, much can be done. Of particular interest may be the extent to which the public knows or assumes such linkages take place now and for what purposes; the perceived legitimacy of actual and perceived purposes; whether statutory or contractual prohibitions against efforts at reidentification would be seen to be adequate; and so on.

We do not believe that an entirely satisfactory technical solution to the reidentification problem is possible; but a great deal more can be done to allow for at least limited release of linked information. The work of Paass (1985) and Smith and Scheuren (1985a) is suggestive here. The line of attack that appears most promising is what might be termed a three-step process. First, "slice" the data up into small enough bits so that each of the "bits" can be adequately masked. (The data, for example, might be divided up into disjoint subsets and for each subset of observations, say, only 2 to 4 different items of administrative data would be provided.) Second, if the slices are chosen appropriately, then one can "splice" back together the complete data set using statistical matching; but in a setting where the conventional--and usually false conditional--independence assumption (e.g., Rodgers, 1984) does not have to be made. Finally, the masking step can add "noise" to the data set in such a way that certain analytic results are either invariant under the noise transformation or correction factors can be calculated and readily applied.

There are some serious losses in this approach. For example, the effective sample size of the linked data items may have shrunk considerably. In any case more research on this problem is definitely warranted, (maybe even if contractual and legal solutions turn out to be eventually possible). Either way, public access to the linked data sets must be seen as a key objective when such studies are undertaken and, to the extent possible, release practices should be as open as with any other data set (Committee on National Statistics, 1985).

Finally, a number of analysis issues have been mentioned which deserve further research, especially in measuring matching errors and adjusting the matched results accordingly. In particular, we need to find a way to escape the historical dilemma that the dissemination and growth of sound theory and practice have been retarded by the perceived uniqueness of many linkage problems (and the customized solutions this perception has led to). The profound nature of the common sense principles upon which good practice is based are not widely enough appreciated. Insufficient attention has been paid to the analysis issues in data linkage systems, perhaps because so much creative energy and financial resources typically go into the linkage steps (Smith and Scheuren, 1985a). It may be too optimistic to suppose that things are now changing, but there is some evidence to this

effect in the success of the 1985 Washington Statistical Society Workshop on Exact Matching Methodologies (Kilss and Alvey, 1985). In any case, it is time to stop treating matching as a necessary but dirty business, isolated from other parts of statistical theory and practice.

ACKNOWLEDGMENTS AND AFTERWORDS

The ideas in this paper owe much to my associations with other professionals in the field of matching. Particular thanks are due to Dan Kasprzyk, for his useful remarks, and, especially, Tom Jabine, whose insightful comments were much appreciated, even though I was unable to incorporate them all in the present version. Tom also acted as a discussant when this paper was originally given and, among other things, corrected a computational error in the calculation of the probability ratios shown in the example. All the remaining errors are, of course, my responsibility.

COMPUTATIONAL NOTE

The Probability Ratios shown in the table above were calculated as follows:

Race and Sex Agree (Race is Black)

$$\frac{99}{100}\cdot\frac{999}{1000}\Big/\left(\frac{1}{10}\cdot\frac{1}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)=197.8020$$

Race and Sex Agree (Race is Nonblack)

$$\frac{99}{100}\cdot\frac{999}{1000}\Big/\left(\frac{9}{10}\cdot\frac{9}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)=2.4420$$

Race Agrees, Sex Does Not (Race is Black)

$$\frac{99}{100}\cdot\frac{1}{1000}\Big/\left(\frac{1}{10}\cdot\frac{1}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)=0.1980$$

Race Agrees, Sex Does Not (Race is Nonblack)

$$\frac{99}{100}\cdot\frac{1}{1000}\Big/\left(\frac{9}{10}\cdot\frac{9}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)=0.0024$$

Sex Agrees, Race Does Not

$$\frac{1}{100}\cdot\frac{999}{1000}\Big/\left(\frac{9}{10}\cdot\frac{1}{10}+\frac{1}{10}\cdot\frac{9}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)=0.1110$$

Neither Agree

$$\frac{1}{100}\cdot\frac{1}{1000}\Big/\left(\frac{9}{10}\cdot\frac{1}{10}+\frac{1}{10}\cdot\frac{9}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)=0.0001$$

REFERENCES

Alexander, L. and Jabine, T.
1978 Access to Social Security Microdata Files for Research and Statistical Purposes: An Overview, Social Security Bulletin, U.S. Social Security Administration.

Alexander, L.
1983 There Ought to be a Law..., Proceedings, Section on Survey Research Methods, American Statistical Association.

Alvey, W. and Aziz, F.
1979 Mortality Reporting in SSA Linked Data: Preliminary Results, Social Security Bulletin, U.S. Social Security Administration.

Alvey, W. and Scheuren, F.
1982 Background for an Administrative Record Census, Proceedings, Social Statistics Section, American Statistical Association.

Aziz, F., et al.
1978 Studies from Interagency Data Linkages (Report No. 8), U.S. Social Security Administration.

Barron, E.
1978 The Survey of Low-Income Aged and Disabled: Survey Design and Data System, Policy Analysis with Social Security Research Files, U.S. Social Security Administration.

Beebe, G.
1985 Why Are Epidemiologists Interested in Matching Algorithms? Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Bentz, M.
1985 The Intergenerational Wealth Study: Prospects for Data Analysis and Methodological Research, presented at the Canadian Conference in Tax Modelling, September 1985.

Bishop, Y., et al.
1975 Discrete Multivariate Analysis: Theory and Practice, MIT Press: Cambridge.

Bixby, L.
1970 Income of People Aged 65 or Older: Overview from the 1968 Survey of the Aged, Social Security Bulletin, U.S. Social Security Administration.

Buckler W. and Smith, C.
1980 The Continuous Work History Sample (CWHS): Description and Contents, Economic and Demographic Statistics, U.S. Social Security Administration.

Butz, W.
1985 The Future of Administrative Records in the Census Bureau's Demographic Activities, Journal of Business and Economic Statistics, American Statistical Association.

Cartwright, D.
1978 Major Limitations of CWHS Files and Prospects for Improvement, Policy Analysis with Social Security Research Files, U.S. Social Security Administration.

Childers, D. and Hogan, H.
1984 Matching IRS Records to Census Records: Some Problems and Results, Proceedings, Section on Survey Research Methods, American Statistical Association.

Clark C. and Coffey, J.
1983 How Many People Can Keep a Secret? Statistical Data Exchange Within a Decentralized System, Proceedings, Section on Survey Research Methods, American Statistical Association.

Cobleigh, C. and Alvey, W.
1975 Validating the Social Security Number, Studies from Interagency Data Linkages (Report No. 4), U.S. Social Security Administration.

Committee on National Statistics
1985 Sharing Research Data, National Academy of Sciences.

Committee on National Statistics
1979 Privacy and Confidentiality as Factors in Survey Response, National Academy of Sciences.

Cox, B. and Folsom, R.
1984 Evaluation of Alternate Designs for a Future NMCUES, Proceedings, Section on Survey Research Methods, American Statistical Association.

Cox, L., et al.
1985 Confidentiality Issues at the Census Bureau, Proceedings of the First Annual Census Bureau Research Conference, U.S. Bureau of the Census.

Cox, L. and Boruch, R.
1985 Emerging Policy Issues in Record Linkage and Privacy, presented at the 45th Session of the International Statistical Institute.

Crane, J. and Kleweno, D.
1985 Project LINK-LINK: An Interactive Database of Administrative Record Linkage Studies, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Dalenius, T.
1983 Informed Consent or R.S.V.P., Incomplete Data in Sample Surveys (Volume I), Academic Press.

DelBene, L.
1979 1972 Augmented Individual Income Tax Model Exact Match File, Studies from Interagency Data Linkages (Report No. 9), U.S. Social Security Administration.

Fellegi, I.
1985 Tutorial on the Fellegi-Sunter Model for Record Linkage, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Fellegi, I. and Sunter, A.
1969 A Theory of Record Linkage, Journal of the American Statistical Association, vol. 64, pp. 1183-1210.

Flaherty, D.
1978 The Bellagio Conference on Privacy, Confidentiality and the Use of Government Microdata, New Directions in Program Evaluation, vol. 4, pp. 19-30.

Gastwirth, J.
1986 Discussion comments to paper by George Duncan and Diane Lambert, A Model for Statistical Disclosure Control Based on Predictive Distributions and Uncertainty Functions, Journal of the American Statistical Association, American Statistical Association.

Gonzalez, M. and Scheuren, F.
1985 Future Work by the Conference of European Statisticians on Population and Housing Censuses, presented before the Thirty-Third Plenary Session of the U.N. Conference of European Statisticians.

Holik, D. and Kozielec, J.
1984 Taxpayers Age 65 or Older, 1977-81, Statistics of Income Bulletin, U.S. Department of the Treasury, Internal Revenue Service.

HEW Secretary's Advisory Committee
1973 Records, Computers and the Rights of Citizens, U.S. Department of Health, Education and Welfare.

Howe, G. and Lindsay, J.
1981 A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies, Computer and Biomedical Research, vol. 14, pp. 327-340.

Irelan, L. and Finegar, W.
1978 Surveys Relating to Retirement and Survivorship, Policy Analysis with Social Security Research Files, U.S. Social Security Administration.

Irwin, R. and Herriot, R.
1982 An Initial Look at Preparing Local Estimates of Household Size from Income Tax Returns, Proceedings, Section on Survey Research Methods, American Statistical Association.

Jabine, T.
1985 Properties of the Social Security Number Relevant to Its Use in Record Linkages, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Jabine, T. and Scheuren, F.
1985 Goals for Statistical Uses of Administrative Records: The Next Ten Years, Journal of Business and Economic Statistics, American Statistical Association.

Jaro, M.
1985 Current Record Linkage Research, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Jaro, M.
1972 UNIMATCH--A Computer System for Generalized Record Linkage Under Conditions of Uncertainty, AFIPS-Conference Proceedings.

Jensen, P.
1983 Towards a Register-Based Statistical System--Some Danish Experience, Statistical Journal of the United Nations, vol. 1, pp. 341-365.

Johnston, D. et al.
1984 1980 AHA Hospital and National Natality/Fetal Mortality Survey Linkage Methodology, Proceedings, Section on Survey Research Methods, American Statistical Association.

Kasprzyk, D.
1983 Social Security Number Reporting, the Use of Administrative Records and the Multiple Frame Design in the Income Survey Development Program, Technical, Conceptual and Administrative Lessons of the Income Survey Development Program, Social Science Research Council: New York.

Kelley, R.
1985 Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Kestenbaum, B.
1985 The Measurement of Early Retirement, Journal of the American Statistical Association, vol. 80, pp. 38-45.

Kilss, B. and Alvey, W.
1985 (Ed.) Record Linkage Techniques -- 1985, U.S. Department of the Treasury, Internal Revenue Service.

Kilss, B. and Scheuren, F.
1980 Goals and Plans for a Linked Administrative Statistical Sample, Proceedings, Section on Survey Research Methods, American Statistical Association.

Kilss, B. and Scheuren, F.
1978 The 1973 CPS-IRS-SSA Exact Match Study, Social Security Bulletin, U.S. Social Security Administration.

Klein, B. and Kasprzyk, D.
1983 Designing an Integrated Disability Data System from Social Security Administrative Records, Proceedings, Section on Survey Research Methods, American Statistical Association.

Kirkendall, N.
1985 Weights in Computer Matching: Applications and an Information Theoretic Point of View, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Maddox, G.: Fillenbaum, G. and George, L.
1978 Extending the Uses of the LRHS' Data Set, Policy Analysis with Social Security Research Files, U.S. Social Security Administration.

Marks, E., et al.
1974 Population Growth Estimation: A Handbook of Vital Statistics Measurement, The Population Council: New York.

Newcombe, H., et al.
1959 Automatic Linkage of Vital Records, Science, vol. 130, pp. 954-959.

Oh, H. L. and Scheuren, F.
1984 Statistical Disclosure Avoidance, presented before a May 1984 meeting of the Washington Statistical Society.

Oh, H. L. and Scheuren, F.
1983 Weighting Adjustments for Unit Nonresponse, Incomplete Data in Sample Surveys (Volume 2), Panel on Incomplete Data, National Academy of Sciences.

Oh, H.L. and Scheuren, F.
1980 Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data, Proceedings, Section on Survey Research Methods, American Statistical Association.

Paass, G.
1985 Disclosure Risk and Disclosure Avoidance for Microdata, presented at the May 1985, meetings of the International Association for Social Service Information and Technology (IASSIST).

Parnes, H., et al.
1979 From the Middle to Later Years: Longitudinal Studies of the Preretirement and Postretirement Experiences of Men, Ohio State University.

Patterson, J. and Bilgrad, R.
1985 The National Death Index Experience: 1981-1985, Record Linkage Techniques -- 1985, U.S. Department of the Treasury, Internal Revenue Service.

President's Reoganization Project for the Federal Statistical System
1981 Improving the Federal Statistical System: Issues and Options, Statistical Reporter.

Redfern, P.
1983 A Study of the Future of the Census of Population: Alternative Approaches, commissioned by the Statistical Office of the European Communities.

Rodgers, W.
1984 An Evaluation of Statistical Matching, Journal of Business and Economic Statistics, American Statistical Association, vol. 2, pp. 91-102.

Rogot, E., et al.
1983 The Use of Probabilistic Methods in Matching Census Samples to the National Death Index, Proceedings, Section on Survey Research Methods, American Statistical Association.

Scheuren, F.
1983 Design and Estimation for Large Federal Surveys Using Administrative Records, Proceedings, Section on Survey Research Methods, American Statistical Association.

Scheuren, F.
1980 Methods of Estimation for the 1973 Exact Match Study, Studies from Interagency Data Linkages (Report No. 10), U.S. Social Security Administration.

Scheuren, F. and Herriot, R.
1975 The Role of the Social Security Number in Matching Administrative and Survey Records, Studies from Interagency Data Linkages (Report No. 4), U.S. Social Security Administration.

Scheuren, F. and Oh, H. L.
1975 Fiddling Around with Nonmatches and Mismatches, Proceedings, Social Statistics Section, American Statistical Association.

Sirken, M. and Greenberg, M.
1983 Redesign and Integration of a Population-Based Health Survey Program, presented at 44th Session of the International Statistical Institute.

Smith M.
1982 Development of a National Record Linkage Program in Canada, Proceedings, Section on Survey Research Methods, American Statistical Association.

Smith, W. and Scheuren, F.
1985a Multiple Linkage and Measures of Inexactness: Methodology Issues, presented at the Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9-10, 1985.

Smith, W. and Scheuren, F.
1985b Some New Methods in Statistical Disclosure Avoidance, presented at the 1985 Annual Meetings of the American Statistical Association, in a session sponsored by the Section on Survey Research Methods.

Steinberg, J. and Pritzker, L.
1967 Some Experiences with and Reflections on Data Linkage in the United States, Bulletin of the International Statistical Institute, vol. 42, pp. 786-805.

Storey, J.
1985 Recent Changes in the Availability of Federal Data on the Aged, report prepared for the Gerontological Society of America.

Tepping, B.
1968 A Model for Optimum Linkage of Records, Journal of the American Statistical Association, vol. 63, pp. 1321-1332.

U.S. Bureau of the Census
1973 The Medicare Record Check: An Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census, PHC(E)-7.

U.S. Health Care Financing Administration
1983 Medicare Statistical Files Manual.

Wilson, O. and Smith, W.
1983 Access to Tax Records for Statistical Purposes, Proceedings, Section on Survey Methods, American Statistical Association.

Winkler, W.
1985 Preprocessing of Lists and String Comparison, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Appendix A

SUPPLEMENTAL BIBLIOGRAPHIC SOURCES

In this paper we have cited some of the literature on exact and statistical matching when the discussion warranted. Further bibliographic material can be found in the following publications:

- Record Linkage Techniques--1985 (1985), U.S. Internal Revenue Service. (Edited by Beth Kilss and Wendy Alvey.) Many of the citations in the present paper come from this volume, which contains the proceedings of the Workshop on Exact Matching Methodologies, held May 9-10, 1985, in Arlington, Virginia.

- Statistical Working Paper Series (1977-1985), Federal Committee on Statistical Methodology. (Produced under the general editorial guidance of Maria Elena Gonzalez.) See especially, No. 5, on "Exact and Statistical Matching," and No. 6, on the "Statistical Uses of Administrative Records." Some of the publications in the Series were prepared by the U.S. Department of Commerce; more recently the publications have been issued by the U.S Office of Management and Budget.

- Statistics of Income and Related Administrative Record Research (1981-1984), U.S. Internal Revenue Service. (Edited by Beth Kilss and Wendy Alvey.) This annual publication series contains numerous papers on record linkage topics and is a successor to the Social Security publications: Statistical Uses of Administrative Records With Emphasis on Mortality and Disability Research (1979) and Economic and Demographic Statistics (1980), which also may be useful.

- Statistical Uses of Administrative Records: Recent Research and Present Prospects (1984), U.S. Internal Revenue Service. (Edited by Thomas Jabine, Beth Kilss and Wendy Alvey.) This handbook of recent work includes many papers on data linkage, most of which are also found in the series listed above.

- Studies From Interagency Data Linkages (1973-80), U.S. Social Security Administration. (Produced under the general editorial supervision of Fritz Scheuren.) Of special interest may be the bibliography by Scheuren, F. and Alvey, W. (1975), "Selected Bibliography on the Matching of Person Records from Different Sources," which will be found in Report No. 4 in the Series, pages 127-136.

- Policy Analysis with Social Security Research Files (1978), U.S. Social Security Administration. (Edited by Wendy Alvey and Fritz Scheuren.) Most of the research files described are based on data linkage methodologies.

- Accessing Individual Records from Personal Data Using Non-Unique Identifiers, National Bureau of Standards, NBS Special Publication 500-2.

Additional citations to the recent literature on disclosure which may be of value are given below. Some of these are of interest as general background; others focus specifically on disclosure barriers to data linkage.

Crank, S. (1985)
Evaluation of Privacy and Disclosure Policy in the Social Security Administration, Social Security Bulletin, U.S. Social Security Administration.

Dalenius, T. (1985)
Privacy and Confidentiality in Censuses and Surveys, Proceedings, Section on Survey Research Methods, American Statistical Association.

Hansen, M. (1971)
The Role and Feasibility of a National Data Bank, Based on Matched Records and Alternatives, Federal Statistics, Report of the President's Commission (vol. II).

Spruill, N. (1984)
Protecting Confidentiality of Business Microdata by Masking, The Public Research Institute: Alexandria, VA.

Spruill, N. (1983)
The Confidentiality and Analytic Usefulness of Masked Business Microdata, Proceedings, Section on Survey Research Methods, American Statistical Association.

Young, P. (1984)
Legal and Administrative Impediments to the Conduct of Epidemiologic Research, Task Force on Environmental Cancer and Heart and Lung Disease: Washington, DC.

Appendix B

TAXPAYER OPINION QUESTION
ON SHARING IRS DATA

Yankelovich, Skelly and White, Inc. (1984)
1984 General Purpose Taxpayer Opinion Survey

60a. As you may know, the IRS has been required by law to keep all of their records confidential. However, some people feel the IRS should share this information with other government departments in order to save money and reduce bureaucratic waste since those departments also need this information to do their work. Others feel that the taxpayer's right to privacy is more important. For which, if any, of these departments or purposes do you think it would be all right for the IRS to provide information?

a. The Census Bureau	24%
b. Major criminal investigations (such as drugs and organized crime)	43%
c. Investigations of illegal aliens	34%
d. Welfare fraud investigations	48%
e. Draft Boards or Selective Service	17%
f. Other U.S. Federal departments	12%
g. State governments	13%
h. Child support investigations	38%
i. Fraud and embezzlement investigations	43%
j. Other	1%
k. None (should keep records private)	31%
l. Don't know/no answer	4%

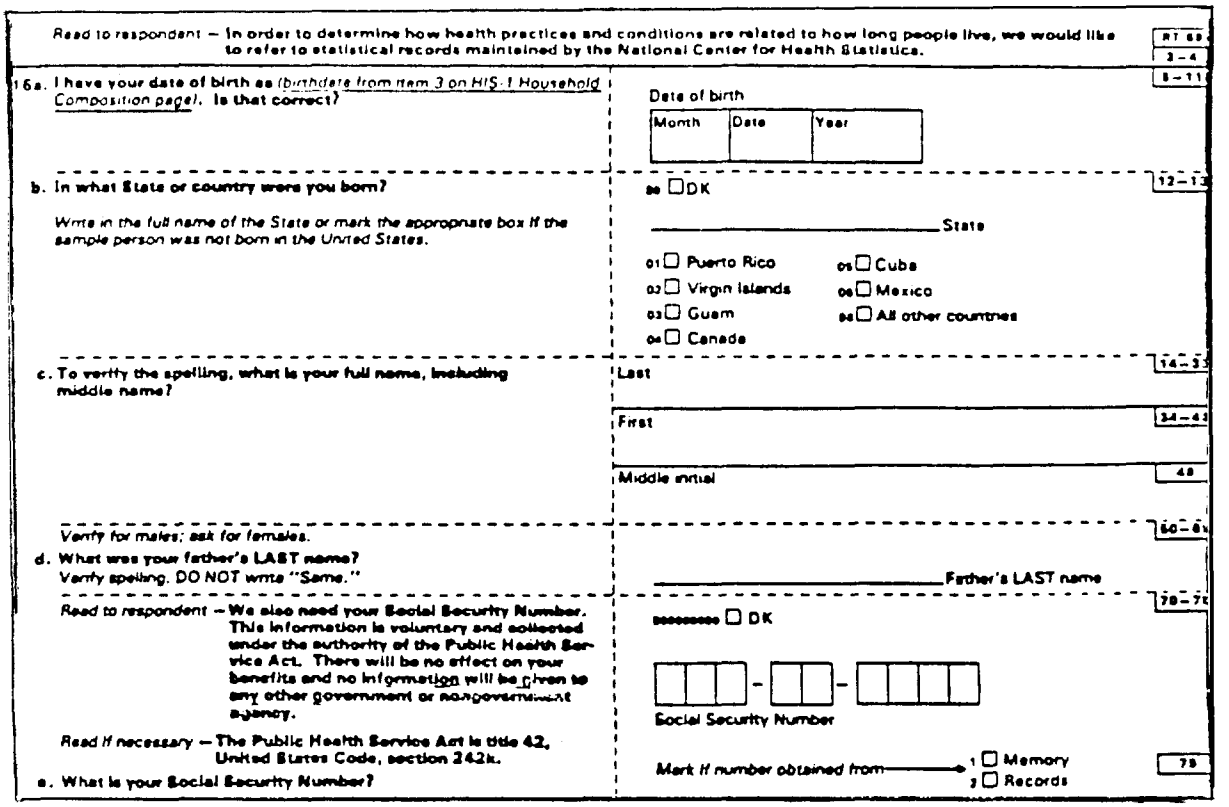
Author's Note:

Tom Jabine, Dan Kasprzyk and others have commented on the many problems this question may have had when it was asked. In my opinion the responses are far from definitive, but they do make the main point I wished to make--that we need more and better research on this issue.

Appendix C

RECORD MATCHING INFORMATION FOR HIS

(Question 16)



Instructions

1. Read the introductory statement above item 16 to explain the purpose of obtaining the information.

*2. When asking 16a, insert the birthdate from the HIS-1, Household Composition Page. If the birthdate recorded in the HIS-1 is in error, make no changes to the HIS-1 entry, but enter the correct birthdate in the answer space in 16a and note "Date verified." If you determine that the person is actually under 55 years of age, footnote the situation and continue the interview. Do not make any changes to the HIS-1(D16-2) or to the supplement. Mark Check Item S2 in Section S based on the original HIS-1 age.

3. Enter the full state name on the line in 16b; do not use abbreviations. If the sample person was not born in one of the 50 states or the District of Columbia, mark the appropriate box in 16b, leaving the state line blank.

4a. If questions arise in 16c, we want the name the sample person is legally known by. If the person has more than one middle name, enter the initial of the first one given. Some women use their maiden name as a middle name: accept the response as given. Be sure to verify the spelling and record the last name first in this item.

*4b. It is acceptable to record an initial as the first name in 16c if this is how the person is legally known. Even if such a person uses their full middle name, only the middle initial is necessary. For example, G. Watson Levi would be recorded as Levi, G., W. in 16c. Do not record name suffixes such as "Sr.," "Jr.," "III," etc.

5a. When verifying 16d for males, ask "Was your father's last name _____?" Always ask the question for females, regardless of their marital status. Be sure to verify the spelling.

5b. Enter the last name of the sample person's father in the answer space, whether it is the same as the person's name or not. Always verify the spelling, even if the names sound alike. If it is volunteered that the person was legally adopted, record the name of the adoptive father.

NOTE: Take special care to make the entries in 16b-d legible. Printing is preferred.

6. Read the introduction to 16e to all respondents. If you are asked for the legal authority for collecting social security numbers, cite the title and section of the United States Code, as printed below the introduction. If you are given more than one number, record the first 9-digit number the respondent mentions, not the first one issued. If the number has more than 9 digits, record the first 9-digits. Do not record alphabetic prefixes or suffixes.

7. After recording the social security number, mark the appropriate box indicating whether the number was obtained from memory or records.

* Revised February 1984

SENSITIVE QUESTIONS

There are no questions considered to be sensitive on either the core series of items or the supplement. However, certain information may be considered sensitive and the following explanation of the need for the data is provided regarding social security number and the subject of incontinence.

- Social Security Number and National Death Index Match

So that in the future the National Center for Health Statistics (NCHS) may investigate the relationship between the results of the "Supplement on Aging" data and causes of death, the supplement collects the appropriate information (items 11a-11e of questionnaire Section 3, Occupation/Retirement), particularly the social security number, that will enable monitoring the National Death Index records for sample persons.

The cost-effectiveness of this supplement is enhanced by the availability of the National Death Index (NDI). Data on the future mortality of the survey population will be available with minimum expenditures by means of a computer search of the NDI. Information on age at death, cause of death, residence at time of death and place of death can be easily ascertained from a copy of the death certificate obtained from the appropriate vital records office. This additional information can be integrated with data from the original survey to greatly enrich the scope of the analysis. Extensive information on the health status of the elderly is being collected on the original survey. Information obtained from death certificates will allow investigators to relate these health status measures to longevity and cause of death. It will also be possible to determine whether selected behavioral and socioeconomic factors collected at the time of the original survey, such as living arrangements, affect the relationship between health characteristics and mortality.

Several years after the data collection and preparation is completed, a list of all survey respondents will be submitted to the NDI and a search made to determine which respondents had died during the interim period. Additional searches of the NDI will be carried out on a periodic basis. In order to optimize the successfulness and reduce the cost associated with these searches, the following information must be collected as part of the original survey: social security number, full (legal) name, Date of birth, State of birth, race, sex, and marital status. Ascertainment of social security number is most essential. A search of the NDI which uses social security number should produce only one match if the subject is deceased. The other information is then used to verify the match. The result of such a match identifies a death certificate which can be obtained from the State with reasonable certainty that it is in fact for the subject. If a social security number is not available, multiple matches within the age range established will occur, especially for common names. This would necessitate obtaining death certificates from several States and attempting to determine whether any of them is for the subject. These false positives would add both acquisition costs and staff costs to the death search process, as well as introducing error.

Interviewers will verify the person's name and birth date (which may have been provided by the household respondent on the core questionnaire), and obtain the last name of the person's father. The social security number will also be requested and if the person is unable to recall the number, he or she will be asked to check their card. This information is not thought to be sensitive; however, respondents will be reminded of the voluntary and confidential nature of the survey, the purpose of the data collection, the legislative authority under which the information is being collected, and the absence of any penalty for refusal. Nonresponse to any of these items will

not affect most of the analyses planned for the supplement; however, provision of social security numbers allows for future epidemiologic research for this population without the necessity of conducting a separate longitudinal or followback survey.

- Incontinence

NCHS's and NIA's interests in general physical problems of older people, which relate directly to their quality of life, include questions on urination and bowel control (Pretest Questionnaire Section V, Items 6a-6e, 7a-7e). One issue is the relationship of incontinence to the aging process. In this case, incontinence can be viewed as a health problem, independent of other illnesses. In order to examine this issue, it will be necessary to collect data from all persons in the 55-and-over age group (so that their effects can be examined) and from people both with and without other illnesses.

In addition, a substantial part of the interest in the problem of incontinence results from the relationship between incontinence and institutionalization. It is the view of some experts consulted that incontinence is one of the main reasons for the decision to institutionalize an older person.

Considerable effort went into wording these questions both to minimize sensitivity and to assure comparability with similar items proposed for the 1984 National Nursing Home Survey. Attachment VIII presents planned analysis of comparable data for both the institutionalized and noninstitutionalized populations from the two surveys.

Appendix D

RECORD MATCHING INFORMATION FOR SIPP
(Question 33)

CARD B - Continued
COMMON QUESTIONS AND SUGGESTED ANSWERS

I thought that the Bureau of the Census operated only every 10 years, when they counted people. What is the Bureau of the Census doing now?

In addition to the decennial census, which is conducted every 10 years, the Bureau collects many different kinds of statistics. Other censuses required by law are conducted on a regular basis including the Census of Agriculture, the Censuses of Business and Manufactures, and the Census of State and Local Governments. In addition, we collect data on a monthly basis to provide current information on such topics as labor force participation, retail and wholesale trade, various manufacturing activities, trade statistics, as well as yearly surveys of business, manufacturing, governments, family income, and education.

Why does the Census Bureau want to know my Social Security Number?

We need to know your Social Security Number so we can add information from administrative records to the survey data. This will help us avoid asking questions for which information is already available and help to ensure the completeness of the survey results. The information we obtain from the Social Security Administration and other government agencies will be protected from unauthorized use just as the survey responses are protected.

O.M.B. No. 0607-0425. Approval Expires June 30, 1984

FORM SIPP-4001 (6-30-83)

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

CONTROL CARD

SURVEY OF INCOME AND PROGRAM PARTICIPATION

NOTICE – Your report to the Census Bureau is confidential by law (title 13, U.S. Code). It may be seen only by sworn Census employees and may be used only for statistical purposes.

PGM 1

1 REGIONAL OFFICE CODE | 2 CONTROL NUMBER: PSU | Segment | Segment suffix | Serial | Sample 8 | 8 | 4 | 3 ADDRESS I.D. | 4 SEGMENT TYPE: 1 ☐ Address 2 ☐ Unit 3 ☐ Permit 4 ☐ Area 5 ☐ Special place | 5a Interviewer code | 5b Letter sent | WAVE 1 2 3 4 5 6 7 8 9 (☐ Yes ☐ No) | 6a EXTRA UNIT Original unit serial number | 6b Sheet ___ Line ___ | 6c OFFICE USE ONLY | 7 Wave for which Control Card first prepared

HOUSEHOLD RECORD (Card ___ of ___)

FILL ITEMS 17 – 20 FOR ALL PERSONS LIVING OR STAYING HERE

FILL (OR UPDATE AS APPROPRIATE) ITEMS 23 – 33 FOR HOUSEHOLD MEMBERS ONLY – Ask each item for entire household before asking next item

17 ENTRY ADDRESS I.D.	18 PERSON NUMBER Assign 101, 102, etc. for wave 1; 201, 202, etc. for wave 2	19a HOUSEHOLD ROSTER FIRST INTERVIEW AT ORIGINAL ADDRESS – What are the names of all persons living or staying here? Start with the name of the person or one of the persons who (owns, rents) this home. Please include middle and maiden names. Last, first, middle, maiden. FIRST INTERVIEW AT MOVER'S NEW ADDRESS Fill items 17 – 19a, 24 – 33c before the interview.	19b RELATIONSHIP TO REFERENCE PERSON (RP) Ask if not apparent: What is ...'s relationship to (Read name of RP)? Example: Reference person, wife or husband, son, daughter, son-in-law, sister, foster son, partner, lodger, lodger's wife. See codes in Information Booklet (Code / Update)	20 HOUSEHOLD MEMBER Does ... usually live here? 1-Yes 2-No – Probe for URE (Yes / No)	23 DATE ENTERED OR LEFT Complete if instructed by item 21, enter code and date. See codes below (First occurrence: Code, Mo., Day; Update: Code, Mo., Day)	24 BIRTH DATE/AGE a. What is ...'s date of birth? Example: 01-20-1983, 12-01-1924. Verify age using flashcard. b. ... is now (Read age) years old, is that correct? (Mo., Day, Year, Age, Age update)	25 PERSON NUMBER OF PARENT Refer to 18 and 19b. If none enter "999"	ASK IF 15 OR OLDER: 26a MARITAL STATUS Is ... now – Married 1-Sp. Pres. 2-Sp. Abs. 3-Widowed 4-Divorced 5-Separated Or 6-Never married? (Code / Update)	26b PERSON NUMBER OF SPOUSE Refer to 18, 19b, and 26a. If none enter "888"	27 DESIGNATED PARENT OR GUARDIAN Fill for "Never married" persons under 15. If no parent, ask: Who in this household is responsible for ...? Enter person number	28 SEX Ask if not apparent: Is ... 1-Male or 2-Female? (M / F)	29 RACE SHOW FLASH-CARD What is the race of each person in this household? See codes below	30 ORIGIN SHOW FLASH-CARD What is the origin or descent of each person in this household? See codes below	ASK IF 15 OR OLDER – EDUCATION: 31a What is the highest grade or year of regular school ... has ever attended? See codes below (Code / Update)	31b Did ... complete that grade (year)? 1-Yes 2-No	ARMED FORCES: 32a Did ... ever serve on active duty in the U.S. Armed Forces? 1-Yes 2-No	If "Yes" in 32a ask: 32b When did ... serve? See codes below	32c Is ... now in the Armed Forces? 1-Yes 2-No	33a
PGM 2				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	
				1☐ 2☐							1☐ 2☐				1☐ 2☐	1☐ 2☐		1☐ 2☐	

21 HOUSEHOLD ROSTER COVERAGE

• FIRST INTERVIEW

21a I have listed (Read names from item 19a), have I missed – If "Yes," to any of the following, ask name and complete items 17 – 20 above. PGM 4

		Yes	No
– Any babies or small children?	0010	1☐	2☐
– Any lodgers, boarders, or persons you employ who live here?	0012	1☐	2☐
– Anyone who usually lives here but is away now – traveling, at school, or in a hospital?	0014	1☐	2☐
– Anyone else staying here?	0016	1☐	2☐

• FIRST INTERVIEW AT MOVER'S NEW ADDRESS

		Yes	No
21b I have listed ... (Read names from item 19a), as now living at this address. Is this correct? If NO, ask – Which person did not move here?	0018	1☐	2☐
21c When did ... move in here? Enter appropriate code and date in item 23			
21d Is there anyone else living or staying here now? If YES, ask name and complete items 17 – 19a, then ask: When did ... begin living here? If moved in before sample person(s), enter code "4" in item 23 and print date. If moved in at the same time or after sample person(s), enter appropriate code and date in item 23.	0020	1☐	2☐
21e Which of these persons (owns/rents) this home? Assign as reference person in item 19b			

22 HOUSING UNIT COVERAGE

• FIRST INTERVIEW

22a INTERVIEWER CHECK ITEM
Unit is: ☐ In a special place – SKIP to 24 above
☐ Not in a special place – ASK 22b

Ask if not apparent:
22b Do all persons in this household live OR eat together?
☐ Yes – ASK 22c
☐ No, neither live nor eat together – Fill Table X for the person or group of persons not living or eating with RP

Ask if not apparent:
22c Does any other household on the property live OR eat with this household?
☐ Yes – Redefine the unit to include space occupied by all persons who live or eat together. Apply merged unit procedures if appropriate.
☐ No
GO TO ITEM 24 ABOVE

• SUBSEQUENT INTERVIEWS

	W2	W3	W4	W5	W6	W7	W8	W9
21f I have listed ... (Read names from item 19a). Are all of these persons still living or staying here? If NO, ask: Who no longer lives here? When did ... leave? Enter appropriate code and date left in item 23. Probe if necessary, for reason left. If entire household moved complete item 23, determine new address (item 34), and end interview.	0022 1☐Y 2☐N	0026 1☐Y 2☐N	0030 1☐Y 2☐N	0034 1☐Y 2☐N	0038 1☐Y 2☐N	0042 1☐Y 2☐N	0046 1☐Y 2☐N	0050 1☐Y 2☐N
21g Is anyone else living or staying here, including new born babies? If YES, ask name and complete items 17 – 20 above, then ask When did ... begin living here? Enter appropriate code and date in item 23. Anyone else?	0024 1☐Y 2☐N	0028 1☐Y 2☐N	0032 1☐Y 2☐N	0036 1☐Y 2☐N	0040 1☐Y 2☐N	0044 1☐Y 2☐N	0048 1☐Y 2☐N	0052 1☐Y 2☐N

CODES FOR 23

Entered
1 – Birth
2 – Marriage
3 – Other
4 – (Use only with item 21d)

Left
5 – Deceased
6 – Institutionalized
7 – Living in Armed Forces barracks
8 – Moved outside of country
9 – Separation or divorce
10 – Person number 201+ no longer living with sample person
11 – Other

CODES FOR 30

01 – German
02 – English
03 – Irish
04 – French
05 – Italian
06 – Scottish
07 – Polish
08 – Dutch
09 – Swedish
10 – Norwegian
11 – Russian
12 – Ukrainian
13 – Welsh
14 – Mexican-American
15 – Chicano
16 – Mexican
17 – Puerto Rican
18 – Cuban
19 – Central or South American (Spanish speaking)
20 – Other Spanish
21 – Afro-Amer. (Black or Negro)
30 – Another group not listed
39 – Don't know

CODES FOR 31a

00 – Never attended or kindergarten
01-08 – Elementary
09-12 – High school
21-26 – College (Academic)

CODES FOR 29

1 – White
2 – Black
3 – American Indian, Eskimo or Aleut
4 – Asian or Pacific Islander
5 – Other – Specify below

Person No.	Specify race

CODES FOR 32b

Enter code in descending order of priority. Thus if person served in Vietnam and in Korea, enter "1".

1 – Vietnam Era (Aug. '64 – April '75)
2 – Korean Conflict (June '50 – Jan. '55)
3 – World War II (Sept. '40 – July '47)
4 – World War I (Apr. '17 – Nov. '18)
5 – May 1975 or later
6 – Other Service (All other periods)

SOCIAL SECURITY

33b Person number Enter person No. from item 18	33c ★ What is ...'s Social Security (or Railroad Retirement) number? If don't know ask: Will you look, would ... be able to provide this information? If "Yes," mark reminder card.	33c Nonresponse Mark the appropriate box 1-Don't know 2-Refused 3-None
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐
		1☐ 2☐ 3☐

INTERVIEWER: Go to first questionnaire interview.