# 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.

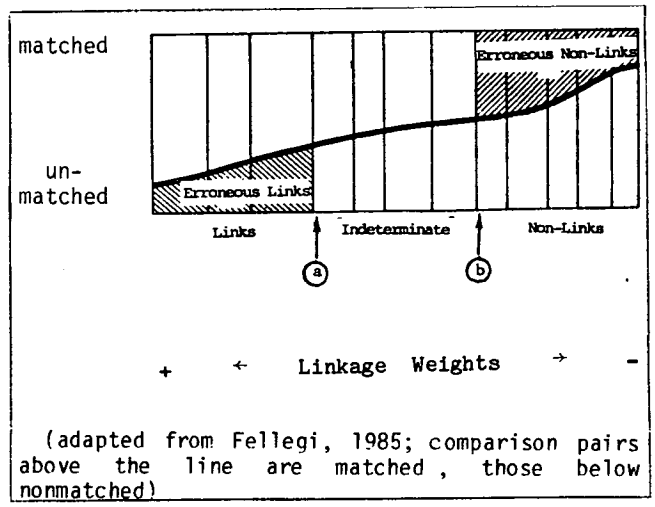
| Outcome | Probability Ratio | Base 2 Log of Ratio |
|---|---|---|
| Race and sex agree: | | |
| Race is black..... | 197.8020 | 7.6279 |
| Race is nonblack..... | 2.4420 | 1.2881 |
| Race agrees, sex does not: | | |
| Race is black..... | 0.1980 | -2.3364 |
| Race is nonblack..... | 0.0024 | -8.7027 |
| 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 inkage 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

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

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% |

---

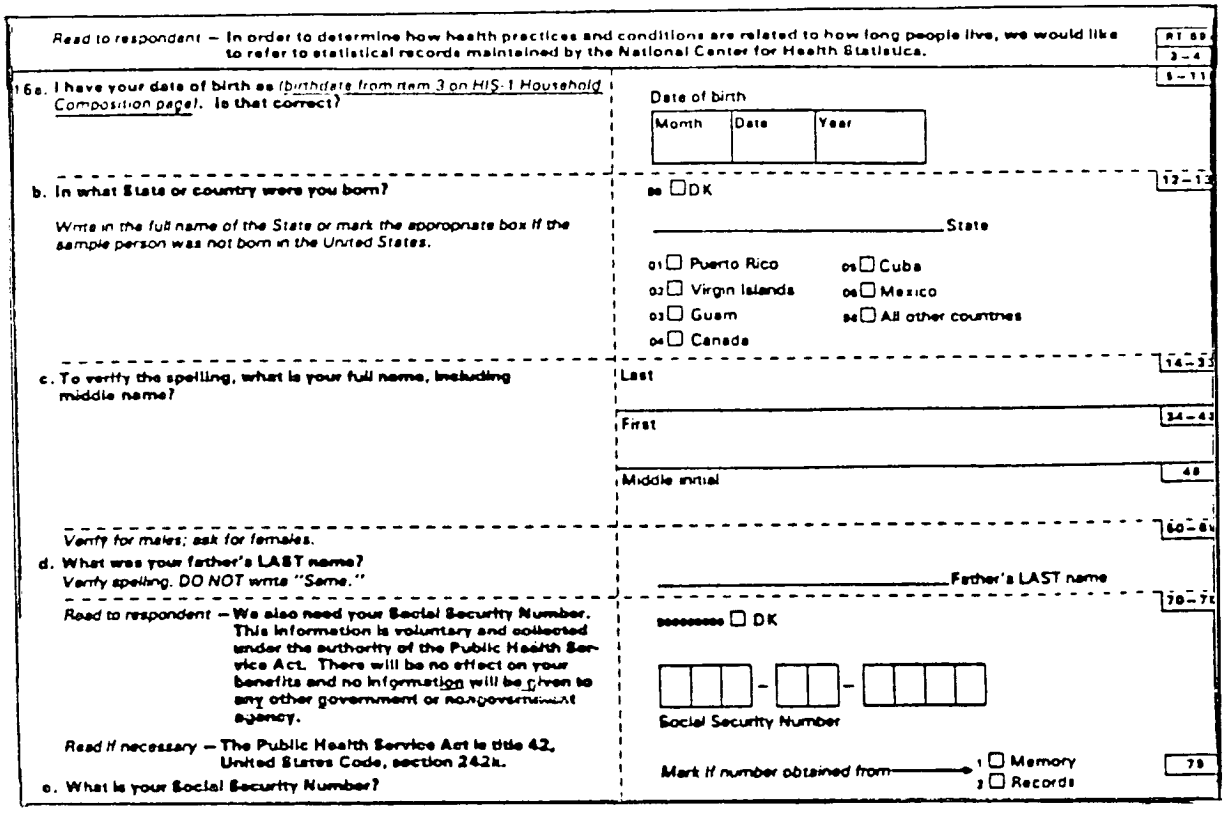
Author's Note:

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)



*Read to respondent* – In order to determine how health practices and conditions are related to how long people live, we would like to refer to statistical records maintained by the National Center for Health Statistics.

16a. I have your date of birth as *(birthdate from item 3 on HIS-1 Household Composition page)*. Is that correct?

Date of birth: Month | Date | Year

b. In what State or country were you born?

*Write in the full name of the State or mark the appropriate box if the sample person was not born in the United States.*

88 ☐ DK

_____ State

01 ☐ Puerto Rico
02 ☐ Virgin Islands
03 ☐ Guam
04 ☐ Canada
05 ☐ Cuba
06 ☐ Mexico
98 ☐ All other countries

c. To verify the spelling, what is your full name, including middle name?

Last

First

Middle initial

*Verify for males; ask for females.*

d. What was your father's LAST name?
*Verify spelling. DO NOT write "Same."*

_____ Father's LAST name

*Read to respondent* – We also need your Social Security Number. This information is voluntary and collected under the authority of the Public Health Service Act. There will be no effect on your benefits and no information will be given to any other government or nongovernmental agency.

999999999 ☐ DK

___ - ___ - ____

Social Security Number

*Read if necessary* – The Public Health Service Act is title 42, United States Code, section 242k.

e. What is your Social Security Number?

*Mark if number obtained from* → 1 ☐ Memory 2 ☐ Records

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

PCM 2

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 ☐ Perma 4 ☐ Area 5 ☐ Special place

5a Interviewer code

6h 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

Form SIPP-4003 (6-28-83)

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) & may be seen only by sworn Census employees and may be used only for statistical purposes.

## 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. (PCM 3) | 18 PERSON NUMBER (Assign 101, 102, etc. for move 1; 201, 202, etc. for move 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 for 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)? Examples: 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 |
|---|---|---|---|---|---|---|
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |
| | | | | | | ☐1 ☐2 |

| 21 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: 01-20-1963, 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 18 OR OLDER: 26a MARITAL STATUS — Is ... now – 1-Mar. 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 25a. If none enter "999" | 27 DESIGNATED PARENT OR GUARDIAN — Fill for "Never married" persons under 18. 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 16 OR OLDER: 31a EDUCATION — 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. Yes \| No | ARMED FORCES: 32a Did ... ever serve on active duty in the U.S. Armed Forces? 1-Yes 2-No. Yes \| No | 32b When did ... serve? See codes below | If "Yes" in 32a ask: 32d Is ... now in the Armed Forces? 1-Yes 2-No. Yes \| No | GO TO ITEM 33a |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |
| | | | | | | | ☐1 ☐2 | | | | | | | |

**23 HOUSEHOLD ROSTER COVERAGE**

• FIRST INTERVIEW

23a 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. (PCM 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 |
|---|---|---|
| 23b 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 |
| 23c When did ... move to here? Enter appropriate code and date in item 23 | | |
| 23d 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 |
| 23e Which of these persons (own/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 B for the person or group of persons not living or eating with RP

Ask if not apparent:

22c Does any other household eat the group by live OR eat with this household? ☐ Yes – Exclude 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 |
|---|---|---|---|---|---|---|---|---|
| 23f 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 |
| 23g 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 – (Also 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-Asian (Black or Negro), 22 – Another group not listed, 23 – Don't know

**CODES FOR 31a**

00 – Never attended or kindergarten
01-08 – Elementary
09-12 – High school
13-18 – College (Academic)

**CODES FOR 32b**

Enter cards 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)

**CODES FOR 28**

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

| Person No. | Specify race |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |
| | |

**SOCIAL SECURITY**

33a Person number — Enter person No. from item 18

33b ★ What is ...'s Social Security (or Railroad Retirement) number? If don't know, ask: If I could knock, could ... 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

| 33a | 33b | 33c |
|---|---|---|
| | | ☐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.