This is the first of five papers at this session on the interagency efforts to upgrade and augment the Continuous Work History Sample (CWHS). We will introduce the subject by describing the planning activities now underway as part of this project.

Organizationally our presentation is divided into four parts. We begin with some background on the CWHS and a statement of the goals of the proposed upgrading. Part 2 details some of the pilot work being completed on the content items that we intend to add to the CWHS system. Part 3 focuses on data quality issues and Part 4 concludes the paper with some prospects for the future.

1. BACKGROUND AND INITIAL PROJECT GOALS

For over forty years [5] both government and nongovernment researchers have made extensive use of statistical information about American workers derived from the Social Security Administration's Continuous Work History Sample (CWHS). The primary Social Security use made of the CWHS has been in tabulating the characteristics of covered workers to keep track of how this group has changed over time with changes in the Social Security Act and in the demographic mix of the population [e.g., 26]. The Bureau of Economic Analysis has made considerable use of the CWHS as a source of regional workforce characteristics and especially changes in the workforce, both geographical and industrial [22]. Uses by nongovernment researchers have also been extensive, covering the gamut from labor market supply questions to the measurement of lifecycle earnings, [e.g., 9, 20]. Recently in a pioneering effort by Goldsmith and Hirschberg [8] attention has been focused on the CWHS' potential to address industrial and environmental health issues.

While the usefulness of the CWHS data has been demonstrated repeatedly, it islimited in scope, content, and quality by program requirements. These weaknesses would, of course, have to be corrected in order for the files to reach their full potential as a general purpose database for statistical research. The support of present and potential users who recognize the importance of these data will be necessary to bring about the changes which will improve its usefulness [6].

Professionals concerned with epidemiological problems, occupational safety, and general environmental issues are among those interested in an improved, augmented CWHS. In fact, the real start of the Linked Administrative Statistical Sample (LASS) project was a meeting at the National Center for Health Statistics (NCHS) in October of 1978 involving representatives of several agencies, including Social Security, to explore areas of mutual concern that relate to epidemiology studies.

When the United States Congress [15] amended the Public Health Service Act (Public Law 95-623), NCHS's mission for conducting and coordinating research activities aimed at improving all aspects of health services in the United States was greatly broadened. Part of this legislation calls for the development of a plan by the National Center for Health Statistics for the collection and coordination of statistical and epidemiological data on the effects of the environment on health. Therefore, NCHS desired to work with other agencies to find feasible, cost-effective approaches to developing an implementation plan for carrying out its new mandate.

One effective and relatively inexpensive way to achieve this goal is to integrate data already collected by Federal agencies in pursuit of their individual missions. Social Security and the Internal Revenue Service (IRS) are two of the major agencies which have current data that are not generally available for epidemiological studies. The proposed LASS project is an attempt to study whether these data systems can be used for examining the occupational and industrial etiology of disease.

Basic LASS Data Elements—The Linked Administrative Statistical Sample is to retain the same simplicity of design as the CWHS, and takes that sample as its starting point. In particular, it is planned that ultimately a common statistical sample will be created which is based on the ending digits of the social security numbers used to select the one-percenter Continuous Work History Sample. The following data elements are proposed for inclusion in the final linked sample:

1. Mortality information from the National Center for Health Statistics' processing of death certificates. (At a minimum, on a prospective basis the fact of death would be confirmed by matching the National Death Index to the CWHS. Also, the basic demographic items from NCHS's statistical record including cause of death would be added. Retrospectively, similar information might be obtained as far back as the late 1960's for every identified CWHS decedent. Finally, for both the retrospective and prospective efforts, the decedent's usual occupation and industry during his or her lifetime, items not now coded by NCHS, would be obtained from the certificates themselves.)

2. Individual income tax items obtained initially from the Statistics of Income (SOI) program. Eventually, the information will be derived directly as a by-product of IRS Master File processing. (Detailed income, deduction and tax data could be obtained from the Transaction Files now used to update the Master File. Also avail-
able from that source would be any needed residence information. Last, but not least, the occupation entry on the return would have to be transcribed to the statistical records.)

3. Longitudinal earnings and benefit histories developed at Social Security as part of the Continuous Work History Sample. (The CWHS, as it now exists, can provide basic demographic information for the sampled individuals; details on every covered job by industry and place of work since 1956; total covered earnings since 1936 (by year since 1950); and, for beneficiaries, the nature of their claims and the amounts they and their dependents receive in benefits.)

Some Long-Run Research Goals--There are a number of long-run goals that an enriched CWHS system could serve. Some of these are:

1. To develop a basic source of socio-economic and job-related mortality and morbidity data. An augmented CWHS could be used, for example, to construct mortality rates by age, race, sex, industry, occupation, and place of work or residence. (This could lead eventually to a much greater understanding of the etiological factors associated with cancer and other causes of death. For one thing, by following individuals by occupation, industry and residence over time, it may be possible to separate out the effects of these factors on health from the effect of health on these factors.)[14]

2. To construct longitudinal personal and administrative unit income profiles of the population at the National, State, and Substate regional levels. These income distributions could be studied both before and after the imposition of Federal income and payroll taxes. (Such a dataset would greatly enhance the capability of the Social Security Administration, the Treasury Department and various Congressional committees to simulate the impact of alternatives to the current tax-transfer system.)

3. To improve the existing data available for studying regional labor market conditions: job mobility; turnover rates; workforce characteristics such as age, race, sex, occupation, wages and self-employment earnings. (The existing CWHS is widely used already to provide regional workforce statistics. Major enhancements, however, are needed as will be discussed in Part 4.)

4. To strengthen existing uses of administrative records for making small area synthetic population and revenue sharing estimates. (This would be partially accomplished by reconciling through record linkage conceptual and other coverage differences at the State and Substate level among a number of existing administrative systems at Social Security, the Health Care Finance Administration (HCFA), and elsewhere.)

5. To complement major household samples like the Current Population Survey and the Survey of Income and Program Participation. (All these surveys can, and probably will, be matched to virtually the same administrative records that make up the CWHS. The CWHS (as a much larger sample) for example, can, with suitable adjustment [19], provide post-stratification control totals which should markedly reduce the sampling variances in these surveys and "correct" them for the (seemingly unavoidable) coverage errors that they will be subject to.) For more details on how these goals might be achieved and still other objectives that could be served, see [17, 18, and 25].

Short Run Goals and Proposed Pilot Activities--The short-run goals of the project are centered around feasibility questions. Thus, in planning for the operational phase of LASS, the participants in this research are examining access restrictions and disclosure issues; potential incompatibilities among the systems being linked; and problems of data quality. Access questions are among the most important issues that have to be addressed before the Continuous Work History Sample can be used to its fullest potential. There is a need to consider the disclosure laws with government-wide application such as the Privacy and Freedom of Information Acts; also, each of the participating agencies has its own legal constraints--statutes and regulations--which control access to its microdata [1-3]. At minimum, these need to be coordinated in terms of some unifying principles of interagency data sharing. In addition, some of them may need to be amended. These issues, however, are out of the scope of this presentation. We would like, instead, to concentrate on the opportunities available for planning enhancements to the CWHS.

The suitability of an upgraded CWHS for studying industrial and occupational mortality depends, in part, on the results of a number of pilot activities. These include efforts to:

1. Examine coverage and content differences between SSA and NCHS death information,
2. Determine the problems which arise when adding cause of death and other data from death certificates to the CWHS,
3. Assess the codability and validity of the occupation entry on the individual income tax return,
4. Study the completeness of the W-2 residence information, and
5. Develop procedures for upgrading the CWHS data on industry and place of work.

The next part of this presentation, therefore, will look briefly at the status of each of the proposed enhancements.

2. Enhancing the CWHS

The Linked Administrative Statistical Sample project will employ several linkage techniques to augment the information now available in the CWHS system. Several agencies are cooperating in the undertaking. Among these are the Social
Security Administration, the Internal Revenue Service, the Health Care Finance Administration, the National Center for Health Statistics, the Office of Personnel Management, and the Railroad Retirement Board. The Bureau of Economic Analysis and the Census Bureau also have been very heavily involved in the early planning stages and will play, it is hoped, a key role in implementing the enhancements envisioned. Staff at the Department of Labor have been keeping abreast of our activities, but just how they might be involved is yet to be determined. A number of other agencies have offered assistance or expressed an interest in using the enriched database once developed. Particular mention should be made here of the National Cancer Institute, the National Institute for Occupational Safety and Health, the Environmental Protection Agency, the Small Business Administration, and the Treasury Department's Office of Tax Analysis.

The major enhancements being proposed for the CWHS include the addition of demographic, programmatic, geographic, and financial variables from various administrative sources. Some of the proposed CWHS enhancements are being pilot tested at the present time as the other speakers at this session will explain; other tasks are being deferred temporarily due to resource constraints. In what follows below, a brief overview of these activities will be given, both ongoing and upcoming. First, pilot work will be described as well as the progress to date. Those items proposed for future testing will also be discussed.

Examining SSA-NCHS Death Reporting Differences—There are two key questions that must be answered if the SSA death reporting system is to be used to study industrial mortality differentials:

1. How complete is the reporting of deaths to SSA?
2. Are there differences in the information shown on death certificates and SSA records?

The reporting of deaths to Social Security is not required for persons who are not OASDI beneficiaries; however, financial incentives, like the lump sum death benefit, make such reports common practice. In order to determine the characteristics of persons whose deaths are not "captured" by SSA, a cooperative project—the 1975 Mortality Coverage and Content Study—was initiated with the National Center for Health Statistics and the National Cancer Institute; this study took as its starting point a sample of 23,000 deaths reported to NCHS for 1975. To date SSA has obtained the death certificates of these decedents and has nearly finished matching them to agency records. A paper presenting preliminary results was given at last year's American Statistical Association Annual Meetings [4]. Another paper by Linda DelBene and Faye Aziz will be given today on this project, providing more results for our analyses of content and coverage differences between SSA and death certificate data.

Adding Data from Death Certificates to the CWHS—To add cause of death to the CWHS it is necessary to supply each State with lists of the decedents identified using SSA information on name, social security number, race, sex, date of birth and date of death. Each State vital records office will then have to search its (microfilm) files and send copies of the death certificates to Social Security.

Several unanswered questions exist about this process. Among these are—

1. Will all the States be able to cooperate?
2. Will SSA's information be sufficient for the States to attempt a search?
3. What will be the quality of the searching?
4. What will be the total cost in money, time and staff?

A pilot test is now underway which should help address these questions. Information on decedents in the CWHS who were identified as having died in 1975 has been sent to the States for death certificate searching. The CWHS decedents were combined, before being sent, with a subsample of NCHS cases already returned as part of the 1975 Mortality Coverage and Content Study.

Merging the two sets of decedents so they are simultaneously searched will make it possible to measure the quality of the work done in each State. (The NCHS cases were previously located by the States using death certificate numbers; now they will be located using SSA identifying information which does not include the certificate number.) Some more details on this effort are included in the paper by Faye Aziz and Warren Buckler to be delivered later at this session.

Adding Occupation to the CWHS—For a number of years professionals concerned with epidemiological problems, occupational safety, and general environmental issues have shown a continuing (and growing) interest in augmenting the Continuous Work History Sample with an occupational variable. One approach for obtaining occupational data for earners in the CWHS is to use the information from individual income tax returns. This will be possible if the occupation item continues to be available on the tax return and if taxpayers provide codable entries on the form.

One of the activities undertaken in preparation for the LASS effort was to compile the many studies [28] which have been done on the reporting of occupation on tax returns in order to make the case that this very important content item be transcribed routinely as part of the Statistics of Income (SOI) program. The evidence from these studies suggests that at the major group level, IRS occupation data may be roughly comparable in quality to that in the decennial censuses. [12]

As part of their Statistics of Income Tax Year 1979 program (under authority of Section 6108(b) of the Tax Reform Act), IRS has agreed to pick up occupation information. This effort will be supported by SSA with the ultimate objective of determining the feasibility and cost of coding occupations for the entire set of tax returns in the CWHS.

At present a collaborative pilot study of the SOI procedures is now underway involving a systematic sample of 6,700 returns. The fourth
paper to be given in this session, by Peter Sailer, Harriet Orcutt, and Phil Clark, will provide a progress report of this work. 

**CWHS Geographic (Residence) Information**

Residence information is presently available on the CWHS for OASDI program recipients. Plans are to add residence to the sample from SSI and Medicare records. For a few years (1972-74), residence is available in the CWHS from the individual income tax return. Plans call for adding, with Census Bureau help, more IRS residence data for both earlier and later years. Beginning with the advent of annual reporting in 1978, the W-2 addresses also are available for use in the CWHS. These addresses are not now being "captured" electronically for administrative purposes.

An experimental pilot study is now underway at SSA to examine W-2 quality and codability issues. In the pilot, an attempt is being made to go back to microfilm copies of the original source documents from the employers. Microprints will be made and then examined for legibility and completeness. If the address data proves adequate, the W-2 could be valuable as a source of residence information for the CWHS.

**CWHS Demographic Items**

The demographic variables on the upgraded CWHS will include sex, race (white, black, other), ethnic origin, date and place of birth, date and place of SSN issuance, and date and place of death. Sex, race, month and year of birth have always been part of the CWHS. Recently added from the Application for a Social Security Number (SSN) were date and place of birth plus the issuance information.

Ethnic origin will be imputed to the sample based on surname and place of birth. A new Census-supplied Spanish surname coding scheme will be applied to the (full) last names of all account number holders. In addition, data on birthplace will enable information to be collected on the earnings and program experience (OASDI and SSI) of immigrants to the United States from foreign countries and migrants to the continental U.S. from Puerto Rico [21].

The newly added data will extend the availability of minority group information from SSA administrative records to persons born in Spanish countries, those from Asian nations, as well as migrants from Puerto Rico.

On an experimental basis attempts may be made, for persons coded other than white or black, to distinguish between those of Asian and American Indian origin. Testing is nearing completion on changes in the application for a social security number, making it possible to obtain ethnic origin directly as a statistical item on all new applicants [13, 16].

**CWHS Financial and Programmatic Information**

Taxable earnings data have been available in the CWHS in one form or another from 1937 on. From 1937 to 1977 nonagricultural wage information is provided for OASDI program recipients. Presently, taxable self-employment income is available since 1955, when it first was subject to FICA taxes. Data on the self-employed by industry exist from 1960 on [5].

Beginning with 1978, "total wages" is included from all employment subject to income tax withholding (not just for social security covered workers). Thus, under the new system, it will be possible to obtain total FICA covered wages for each job directly, rather than have to estimate it, as in the past [e.g., 10]. Also non-FICA wages will be available routinely for the first time. Finally, as a result of the changeover to annual reporting, the CWHS could be augmented by other (W-2 and W-2P) items of information not included formerly: for example, income tax withheld, non-SSA pension coverage, and non-SSA pension income.

From SSA program records, benefit income information is already being provided for OASDI recipients in the CWHS. SSA payment information will soon be added (plus SSI eligibility information about the assets and other income of program participants). Linkage to the Medicare records maintained by HCFA will make it possible to obtain extensive medical and disability data. Longstanding plans for obtaining wage, payment, and other program data from the Office of Personnel Management and the Railroad Retirement Board are currently being deferred but could be reconsidered, especially if there is a Presidential proposal to fully integrate these systems with social security, something now under intense study [24]. Matching the CWHS to the planned Department of Labor's Continuous Wage and Benefit History (CWBH) sample is another project that offers promise but on which active planning is not now underway.

3. ASSESSING DATA QUALITY

One of the key weaknesses in the CWHS as it now exists is that establishment reporting by employers is voluntary. Even with a maximum effort it is unlikely that the Establishment Reporting Plan (ERP) will ever achieve better than a 90 percent participation rate [5]. At present, participation is only about 80 percent and the likelihood is that it could fall sharply as a result of the 1978 conversion of employer wage reporting from a quarterly to an annual system. (The W-2 is replacing the Form 941 as the source of wage data.) Because of SSA's start-up problems, it is still not possible to estimate how many employers will, in the changeover to the new procedures, discontinue ERP. Preliminary data suggest a sizeable reduction in participation.

An important goal of the LASS effort is to upgrade the quality of the CWHS coding of industry and place of work. To this end, there must be a further strengthening of the existing cooperative efforts between the Bureau of Economic Analysis (BEA) and SSA in thoroughly examining the data quality problems which exist in the CWHS and which, by no means, are confined solely to failures in ERP [5,6]. Equally important is the need to revitalize and expand the longstanding cooperative arrangements between the Census Bureau and SSA.

**BEA-SSA Efforts to Upgrade CWHS Industry and Place of Work Data**

With respect to the BEA-SSA relationship, plans include a detailed set of procedures to "perfect" the CWHS files for the period 1957-1977. A comprehensive approach to the handling of misreported (and/or missing) data is anticipated from this joint BEA-SSA effort. The data editing and imputation tasks are expected to be
quite formidable indeed. Joint BEA-SSA plans are also being developed to handle the new (post 1977) data quality problems that are being encountered in the changeover to annual wage reporting. The final paper in this session, by Bruce Levine, discusses progress of ongoing work aimed at increasing the accuracy of employer-reported geography and industry in the CHS.

Census-SSA Plans—It is also expected that the Census Bureau will participate in the CHS upgrading. This effort, however, will have a different focus from the plans developing with the Bureau of Economic Analysis. Traditionally, Social Security has provided industry and place of work data for new employers to the Census Bureau, in connection with the Bureau's Standard Statistical Establishment List (SSEL) program [23]. After each Economic Census, the Bureau has returned to Social Security updated industry data for use in the CHS. For single establishment employers, the incorporation of this data in the CHS is fairly routine. For multi-establishment employers, real difficulties arise because of Census and SSA differences in the identification of establishments. Problems exist also due to failures by SSA to obtain establishment-level information from some employers at all. Two major changes in this arrangement are being proposed: (1) that SSA be provided with the Bureau's SSEL annual updates on place of work codes for single-establishment employers (again, if the confidentiality issues can be worked out); and (2) that, for multi-establishment employers, an experimental study be undertaken to see if the SSEL information on employer place of work can be combined with the employee's residential address (from the individual income tax return or the W-2), in order to create synthetic establishment identification codes for CHS cases where the voluntary SSA establishment reporting plan is not working properly. (The SSEL information is on an establishment basis, while SSA and tax return records are a mixture of establishment and firm basis.)

The synthetic establishment assignment process, as it is envisioned to date would use a Census Bureau address coding scheme to determine the distance between the employee's home and all the establishments of his employer. The establishment closest to residence could be chosen as the establishment that was "most likely" to be the employee's place of work. Complications caused by address changes over time would have to be overcome; but the scheme, in our opinion, offers real promise and should be tested.

Future Possibilities—A number of other general recommendations for potential improvements in data quality could be made. Most of these are a "natural" consequence of the enrichment of the CHS by data from new sources and have been described in the preceding sections. However, there is one further point that bears on the quality issue which at least should be touched on here.

The LASS planners have been giving serious consideration to designing a new sample for the CHS which would supplement the existing 1% data. This sample would be based on employers. It would be stratified by industry and firm size and longitudinal in nature, i.e., subsamples of the employees whose employers fall in the sample would be followed longitudinally. Such a design would allow resources to be concentrated in the voluntary establishment reporting plan on just the sampled employers. Improved participation rates and other quality gains would be expected.

4. FUTURE PROSPECTS FOR LASS

In order to mount the proposed Linked Administrative Statistical Sample project, a high degree of cooperation is essential among the agencies involved. Most of the technical problems which must be faced have already been touched on in this paper. Perhaps the hardest problems though, as in any large endeavor, are organizational or managerial in nature. Meetings with both the potential producer and user agencies have been held with some frequency since October 1978; however, the LASS project is still in its initial planning phase. (Full details on the progress to date may be found in LASS Working Notes Series [27] or in the publication Statistical Uses of Administrative Records with Emphasis on Mortality and Disability Research [29].) It will be some time before all the options have been laid out and the costs estimated. Ordering priorities will be a difficult process since each participating organization has its own missions, research goals and administrative procedures. There is also a concern about the ability of each of the participating agencies to obtain the new staff and budget that will be required. Throughout the next year efforts to overcome these obstacles will continue in the hopes that an enhanced CHS can be made available to researchers in the near future.

REFERENCES

[8] Goldsmith, J. and Hirschberg, D., "Mortality and Industrial Employment (I)," J. Occupational Medicine, Vol. 18, pp. 161-164, 1976. (There were also two other papers by Goldsmith in this journal and an important letter commenting on the results by Pierre De Couffle.)


