DISABILITY RESEARCH USING SURVEY DATA MATCHED TO SOCIAL SECURITY ADMINISTRATION RECORD DATA*

Kalman Rupp, Ph.D.
Social Security Administration

Paul S. Davies, Ph.D.
Social Security Administration

Abstract The Social Security Administration (SSA) maintains a rich system of administrative record databases. These data systems have been extensively used to administer SSA's programs, to compile descriptive program statistics, and as the basis for many research projects conducted at SSA. There has been increasing interest in using administrative records for policy evaluation, particularly in capitalizing on the analytic potential of linking various administrative record data systems with other administrative records and with survey data. This article provides a brief description of the major SSA administrative records databases available for disability research. We also discuss the advantages and disadvantages of using administrative and survey data for disability policy research and evaluation. We then illustrate the relative roles of administrative and survey data by briefly describing two substantially different activities in which SSA uses matched administrative and survey data for disability policy research and evaluation.

I. Introduction

The Social Security Administration (SSA) maintains a rich system of administrative record databases. These data systems have been extensively used to administer SSA's programs: The Old Age and Survivors Insurance (OASI) and Disability Insurance (DI) components of the Social Security social insurance program, and the means-tested Supplemental Security Income (SSI) program. SSA administrative records data also have been used to compile descriptive program statistics appearing in SSA's regular statistical publications (e.g., *Annual Statistical Supplement to the Social Security Bulletin*) and elsewhere (e.g., "Green Book" compiled by the House Ways and Means Committee) for many years. In addition, administrative record data have served as the basis of many research projects conducted at SSA over the years.

More recently, there has been increasing interest in using administrative records for policy evaluation, particularly in capitalizing on the analytic potential of using data linkages among various administrative record data systems and in linking survey and administrative record data. The joint use of multiple records systems and associated survey data is especially intriguing in the case of the disabled segment of SSA's target populations. This is so because the disability determination system affecting participation in SSA's disability programs (DI and the disabled component of SSI) involves complex rules, processes, and a large array of diverse variables. The outcomes of interest tend to be multidimensional, and the target populations are often relatively small. As a result of such complexities, oftentimes no single data system is available as a sufficient

* The views expressed in this paper are the authors' and do not represent the official position of the Social Security Administration or any other federal agency.

information base for a particular policy evaluation. The use of multiple sources of data with or without matching at the person level may become very attractive in order to arrive at a sufficient information base within appropriate time frames and with costs that are not prohibitively high.

While a number of research and evaluation studies focus on the old-age component of Social Security and SSI, this paper focuses on research and evaluation of the disability component of Social Security and SSI. DI focuses on preretirement-age disabled workers, while categorical SSI eligibility on account of disability is limited to preretirement age adults (aged 18-64) and children (aged 0-17). The disability eligibility criteria used in the DI and SSI programs are identical and use the same process. In contrast, the economic criteria for eligibility differ. In the DI program, a certain level of recent work experience is required (measured by the so-called "DI-insured" status), but there is no means test. Eligibility for SSI disabled status depends on means testing (assets and income), but there is no work experience requirement. Some disabled people are "concurrent" eligibles, meaning that they are eligible to receive payments from both the DI and SSI programs.

Administrative records and survey data have relative advantages and complementarities, depending on the specific policy evaluation questions and the specific nature of the available data sets. Nevertheless, some general considerations apply to most situations involving the potential use and combination of the various sources. Administrative records tend to be excellent sources of information items that are related to the specific purpose of the data collection (such as benefit payments), but less reliable if the information is not directly related to some valued administrative function. Administrative record data is typically available for 100 percent of the underlying universe of interest; therefore, they also can be useful as a frame for launching an associated survey, and in assessing the effects of sampling and nonsampling error. Given the decreasing costs associated with data processing, the marginal cost of additional observations from a given administrative records system is relatively small. Finally, given the relatively small number of observations for particular target groups of interest, administrative records can be used to form the sample frame for customized surveys, or can be used as the only source of quantitative data by default.

On the other hand, administrative records systems are rigid; it is hard or impossible to customize data items for evaluation purposes. They are often not collected in the standardized and pretested way survey data are collected. Oftentimes, the reporting of information is adversely affected by "moral hazard" – the incentive for respondents to provide information that might produce the best administrative outcome for them. (Documentation requirements and legal sanctions for untruthful reporting may have opposite effects.) Finally, administrative record data (with a few notable exceptions, such as birth and death records) are truncated to the subuniverse of interest – in our case persons applying for or awarded disability benefits.

Therefore, any evaluation looking at factors affecting participation has to include some external data source, such as a survey. The advantages and disadvantages of relying on survey data alone tend to represent the opposites of the pros and cons of relying on

administrative records alone. Obviously, in many situations the combination of multiple sources and especially the combination of survey and administrative record data provides a superb alternative to relying on a single source.

In the next sections of this article, we provide a brief description of the major SSA administrative records databases for disability research, followed by an example of the use of administrative and survey data in two key areas of disability policy evaluation. We conclude with a discussion of outstanding issues and promising areas for future work.

II. Major Administrative Records Data Bases for Disability Research

SSA maintains a number of record systems used to administer the DI and SSI programs. For purposes of disability research, the most important administrative records data bases are the Master Benefit Record (MBR), the Supplemental Security Record (SSR), the National Disability Determination System (NDDS), the Master Earnings File (MEF), and the Social Security Account Number Identification (NUMIDENT) file. Using these files, a researcher can prepare extracts that provide program participation and benefit information for long time-series or for particular cross-sections of interest. This type of information is often useful in forming control or comparison groups. An important limitation stems from the fact that these systems are designed and maintained primarily for the purpose of program administration. Many times, fields that would appear to be useful for research purposes are not updated or are not filled in at all if they are not directly relevant to program administration. Furthermore, some fields are overwritten as new events occur. Therefore, the researcher must use great caution in selecting data fields for analysis to ensure that they contain the appropriate information. There also are issues of privacy, confidentiality, and ownership with which one must grapple when using administrative records data bases for research. The most relevant example is the MEF. The data in the MEF are actually owned by the Internal Revenue Service (IRS). Therefore, SSA employees must abide by IRS privacy regulations when using data from the MEF. SSA's contractors and grantees are not allowed access to the MEF. In the following paragraphs, we provide a brief description of SSA's administrative records data bases that are relevant for disability research.

Master Beneficiary Record (MBR)

The MBR covers OASDI applicants and beneficiaries. It contains approximately 133 million person records grouped into roughly 93 million units based on the Social Security number of the primary wage earner in the unit. Each beneficiary in the unit – spouses, children, widows, parents, etc. – can be identified by his/her own Social Security number (SSN). From the MBR, we can compile an individual's payment eligibility history for many years in the past. In addition, the MBR includes variables such as: SSN of the primary wage earner, SSN of the individual, gender, race, date of birth, primary insurance amount, average indexed monthly earnings, state and county code, current date of entitlement, date of filing, type of claim, diagnosis code, and dual entitlement status code.

Numerous extracts are available from the MBR. The 810/811 extract is an historical format file containing complete histories of benefit amounts that are accurate as of the

date of extraction. The MBR Universe file is repeated every 6 months and represents the entire universe of MBR cases – approximately 133 million individuals – but only a small number of fields are available. In addition, each month a 1-percent and a 10-percent extract file are available. These periodic extract files are snapshots of the MBR as of a point in time and do not contain historical information on benefit amounts or eligibility.

Supplemental Security Record (SSR)
The SSR covers all SSI applicants and recipients. It contains approximately 57 million person records grouped into approximately 42 million units based on the configuration of the family/household as it relates to SSI eligibility (e.g., individual, individual with eligible spouse, individual with ineligible spouse, child, child with ineligible parent(s), etc.). A person may have several records on the SSR, primarily because new records are required whenever a new application is filed or when other types of events occur, such as a change in the structure of the SSI unit. Data items on the SSR include: SSN, date of birth, gender, race, noncitizen information, date of claim, primary and secondary disability diagnosis code, state and county of residence, countable earned and unearned income amounts, payment status, amount of federal SSI benefit, current amount of state supplementation, and date of first payment.

There are several commonly used extracts from the SSR (Pickett and Scott, 1996). The monthly Characteristic Extract Record (CER) is a 10-percent sample selected based on the last three digits of the SSN. It is unique in that it provides the federal SSI benefit amount based on payment eligibility as well as the actual payment amount received. This is an important distinction for many research and evaluation studies. The 1-percent SSI Longitudinal Extract is drawn every six months. It contains monthly eligibility and payment history for each record back to the beginning of the SSI program in 1974.

National Disability Determination System (NDDS)
The NDDS contains records of all disability decisions made by the state-run Disability Determination Services. Disability determinations are most commonly the result of an SSI or SSDI application, but also may result from a disability redetermination, a continuing disability review, or an appeal of a denied application or benefit termination. Both adult and child disability claimants are included in the NDDS. Relevant fields include: SSN, beneficiary identification code, filing date, type of claim, date of DDS or SSA decision, result of determination, date of birth, primary and secondary impairment codes, date disability period began, gender, and race. These files are generally used to supplement information on the MBR and SSR.

Master Earnings File (MEF)
The MEF contains full (not capped or top-coded) annual earnings for workers based on their SSN. There are nearly 400 million earnings records in the MEF. Each individual's earnings for a given calendar year are posted to the MEF by the November following the close of the tax year (e.g., calendar year 1998 earnings are posted to the MEF by November 1999). Annual earnings totals are available on the MEF from 1951 through the present. Other variables on the MEF include: gender, race, date of birth, date of death, first year of earnings, last year of earnings, FICA earnings, and Medicare earnings.

The information in the MEF is provided to SSA by the IRS, and is therefore subject to very strict confidentiality and privacy regulations. These restrictions notwithstanding, the MEF represents a tremendously rich source of information for research and evaluation, especially when linked to benefit information from the MBR and SSR.

Social Security Number Identification File (NUMIDENT)
The NUMIDENT file contains information on all persons who have ever submitted an application for a Social Security number. The application contains the individual's name, date of birth, city of birth, state or country of birth, gender, race, mother's and father's name and Social Security Number, evidence used to establish citizenship, and date of death. This file contains approximately 689 million records for 389 million people. It is a useful supplement or cross-check for basic demographic information available on other administrative record data bases, such as citizenship, race, and date of death.

III. Use of Administrative Record Data Bases and Survey Data for the Evaluation of Employment-Oriented Demonstrations

An important disability policy objective at the Social Security Administration over many years has been the promotion of employment among disabled SSI and DI beneficiaries. To qualify for disability benefits, applicants have to demonstrate severe disabilities that are expected to last for at least one year or to result in death. Applicants who earn above the "Substantial Gainful Activity (SGA)" level (currently $700 per month) are automatically denied benefits. Therefore, newly awarded disability applicants are either not employed at all, or have fairly low levels of earnings. Because of the severity of disabilities and the SGA screen, newly awarded disability beneficiaries face substantial barriers to employment. Nevertheless, the hope manifested in agency employment policies is that some disability beneficiaries – with support and appropriate incentives – may be able to enter (or re-enter) employment with sufficiently high earnings potential to facilitate a reduction of the receipt of DI and SSI benefits. This objective is a tall order. Hence, over the years SSA has supported a number of demonstration evaluations testing the feasibility and efficacy of the interventions proposed.

In a typical demonstration evaluation, the implementation is tested at a number of localities. At these demonstration sites, volunteers are solicited, and depending on the evaluation design may be randomly assigned to "treatment" and "control" group status. An evaluation component is designed geared towards the description of the intervention tested (process analysis), the analysis of factors affecting the decision to volunteer among targeted beneficiaries, and the estimation of the net effects of the demonstration. Two demonstration evaluations that included such an evaluation design are the recently completed evaluation of the Project NetWork Return-to-Work Experiment (Rupp et al., 1994, 1996 and 1999) and the previous Transitional Employment Training Demonstration evaluation (Decker and Thornton, 1995). Several ongoing or planned SSA demonstrations call for similar overall designs. Because of the large-scale and comprehensive nature of the Project NetWork evaluation, we focus on this experience in the discussion below.

Administrative record data are at the core of SSA demonstration evaluations. This is because project eligibles and participants can typically be identified based on administrative record systems that cover the entire national universe of people who ever applied for or were awarded disability benefits. Many key personal characteristics and outcomes can be tracked based on administrative records for an indefinite follow-up period with low marginal cost and without the need to worry about attrition bias. National survey data are not promising candidates for the evaluation of such demonstrations simply because the demonstrations are operated at a limited number of diverse localities across the country, such as Dallas, TX, Tampa, FL, Las Vegas, NV; national surveys typically do not have the necessary number of observations for any meaningful evaluation. Customized surveys, however, can be launched and tailored to provide information that is complementary to the administrative records data. These data tend to be costly, and budget constraints may result in limitations of statistical power when measuring key demonstration outcomes.

The Project NetWork demonstration evaluation relied on the development of a comprehensive administrative record database with two waves of supplementary survey data. The following paragraphs describe the major features of this data system; more detail is provided in Rupp et al. (1999). The universe of interest included project eligibles who accepted an invitation to volunteer for Project NetWork (project participants, randomly assigned to "treatment" and "control" status), and eligibles who were invited to participate at the 8 demonstration sites but did not volunteer. Because some of the electronic mailing lists SSA generated for the outreach were lost, an important part of the evaluation included the simulation of the Project NetWork eligibility rules using a combination of SSA's national administrative record databases. This simulation identified 138,613 eligible nonparticipants. An additional group of 8,248 participants were identified based on SSA's automated local management information system. These volunteers were randomly assigned to the "treatment" and "control" groups to complete the analytic universe of 146,861 cases. To assure the integrity of randomization, assignment to treatment and control groups was performed off-site by an independent evaluation contractor, Abt Associates, using a computerized random assignment process.

Outreach, record matching among the various SSA source files, and creation of the sample frame for linked survey data (collected by Abt Associates) used the SSN as the source of identifying the appropriate person record. The analytic variables of interest were derived, cleaned and edited from the MBR, SSR, NDDS, MEF and Numident. From the administrative record data, monthly SSI and DI benefit information and annual earnings were used for the experimental net outcome evaluation for several post-randomization years. Importantly, the measurement of these outcomes was feasible without attrition. The SSI and DI benefit, MEF earnings data, and demographic and diagnostic information derived from NDDS and other administrative records were also utilized to analyze the factors affecting the decision to participate in Project NetWork.

A baseline and a follow-up survey were conducted to provide supplementary information on noneconomic variables (e.g. health, functional limitations, self-esteem, measures of cognitive and affective functioning, hospitalizations) that may affect the decision to

volunteer and/or net outcomes, labor market variables that go beyond the limited annual earnings information that is provided by the MEF, attitudes toward work and the demonstration, the cost of vocational rehabilitation services from sources other than Project NetWork, and perceptions concerning program rules. Baseline interviews were completed with 3,439 Project NetWork participants and nonparticipants. The response rate was 87 percent for participants, but only about 50 percent for nonparticipants. A total of 1,521 follow-up interviews were completed (with participants only), for a final response rate of 83 percent. The survey information was useful in providing a rich array of data that are not available from administrative records. However, survey attrition (especially for nonparticipants) and the relatively low statistical power of impact estimates for noneconomic variables were important limitations. Administrative records were useful to gauge the nature of survey attrition.

The Project NetWork evaluation confirmed the usefulness of administrative record data as a core of SSA demonstration evaluations, and through the simulation of complex eligibility rules, indicated the feasibility of using administrative record data systems for the creation of customized comparison and control groups in demonstration evaluations. Survey data formed a supplementary role, useful for descriptive purposes, but were of more limited value due to the potential of Type II error in the net impact analysis arising from limited resources for relatively expensive survey observations.

IV. Use of Survey of Income and Program Participation (SIPP) Files Matched to Administrative Record Data to Simulate Eligibility and to Assess Participation in SSA's Programs

Frequently, changes to program eligibility criteria are proposed, or policy makers need to know the potential effects of legislative initiatives that call for changes in eligibility criteria. For example, what would be the effect on the SSI program and SSI recipients of increasing the unearned income disregard from $20 to $125? How will the SSDI and SSI caseload be affected by the recent increase in the SGA level from $500 to $700? Administrative records data bases in isolation are not well suited to answer questions of this nature, primarily because they typically do not contain information about nonparticipants who would be potentially affected by the policy change. Although administrative records data bases offer very detailed program data with which to analyze policy changes, they often lack economic and/or demographic variables that are critical to such evaluations. Conversely, general-purpose surveys do not contain the program data necessary to examine detailed features of eligibility. SSA is working to overcome these limitations by combining the two sources of information. By exact matching survey data with administrative records, one can create a very powerful data base for policy evaluation and analysis. This section provides an overview of efforts at SSA to use general-purpose surveys combined with administrative records to address broad policy questions related to SSA's disability programs. To illustrate, we describe an SSA modeling effort in which the SIPP is used as the core source of information to simulate SSI eligibility. Administrative records are added to refine the model and to expand the range of questions that can be answered. SSI eligibility using SIPP survey data alone was first simulated using the 1984 SIPP (Wixon and Vaughan, 1991; Vaughan and Wixon,

1991). Ongoing research by the authors and colleagues focuses on enhancing the capability of modeling SSI financial eligibility using the 1990s SIPP panels.

Eligibility for the SSI program is based on financial, age, and disability factors. In order to simulate program eligibility, and to evaluate the effects of a policy change, one must construct an eligibility model using a data source that provides substantial detail on income, assets, health/disability, household composition, and age. Availability of other demographic characteristics, such as gender and race, also is desirable in order to describe the populations of interest. For most of this information, the SIPP is an ideal source. It is a panel data set that gathers monthly information on detailed income sources and amounts in the core interviews, and on detailed asset categories, asset amounts, health impairments, and work limitations in the topical modules. With this wealth of information, we are able to develop a SIPP-based eligibility simulation model that can be used to answer broad policy questions related to the SSI program. We can develop a cross-sectional model that utilizes detailed information from the topical modules. We also can develop a temporal or dynamic model that exploits the panel nature of the data, but sacrifices some of the detailed information on assets and disability. Both of these approaches are useful, although one will generally be more appropriate than the other depending on the particular policy question at hand.

Because the SSI program considers the categorical (age, disability) eligibility and financial resources of other household members in determining eligibility, we develop a couples file from the SIPP. The simulation model is composed of a series of modules that are designed to replicate, as closely as possible, the features of the SSI eligibility determination. One module considers the categorical eligibility of the individual and his/her spouse (if present). In other words, this module performs a separate simulation for the individual and his/her spouse to determine their individual categorical eligibility status for the SSI program based on age and disability. The second module considers the assets of an individual and his/her spouse. If countable assets for the SSI unit are below the relevant cut-off level, the SSI unit is determined to be resource eligible for the SSI program. The third module considers the income of an individual and his/her spouse. In this module, countable income is derived based on the eligibility provisions for the relevant type of SSI unit. If countable income is below the relevant cut-off level, the SSI unit is determined to be income eligible for the SSI program. The final module combines the three features of SSI eligibility from the previous sections (categorical, assets, and income) to develop a single indicator of SSI eligibility. Notably, SSI financial eligibility is simulated for categorically eligible units as well as for categorically ineligible units. This is very important for policy evaluation. To conduct policy simulations, one would simply change the appropriate eligibility criteria in the simulation model, re-run the model using the new criteria, and compare the characteristics of the SSI-eligible units under the new eligibility criteria with those of the SSI-eligible units under the original eligibility criteria.

To this point, the model has relied strictly on SIPP survey data. However, by exact matching the SIPP to SSA administrative records data bases, we can refine and enhance the model, and expand the nature of policy questions the model can address. For

example, we can replace survey self-reports of SSI participation with a participation flag from the administrative records. Presumably, the administrative report will be more accurate than the self-report – it is believed that survey respondents sometimes confuse Supplemental Security Income with Social Security, leading to erroneous responses. In addition, the administrative records allow us to follow SIPP respondents not only during the SIPP panel, but also for many years before and many years after the panel. The SSI records, for example, provide monthly program participation information and benefit amounts from January 1974 through the present. These monthly program histories can be used as both independent and dependent variables in analyses of the effects of SSI policy changes. As independent variables, we can consider the effects of past spells of SSI participation on current outcomes. As dependent variables, we can develop duration measures and analyze the dynamics of SSI participation. Since the MBR and NDDS also are matched to the SIPP, we can use administrative reports of date of disability onset and disability diagnosis to improve survey-based measures of categorical eligibility. The MBR also will permit us to identify concurrent SSI and SSDI recipients.

IV. Conclusion

In this paper, we discussed the use of SSA administrative records and matched survey data for disability research and policy evaluation. We provided a brief description of the major administrative records data systems, and discussed two substantially different examples of the use and relative role of matched administrative records and survey data. The relative role and utility of survey data and administrative records depends on the nature of the problem, the availability of observations and relevant data items, and feasibility considerations (such as confidentiality restrictions).

The decisions concerning the role of survey and administrative records fundamentally depend on the nature and size of the target population and the relevant sample sizes that may be available from survey and administrative records. On one end of the spectrum, there are research questions where the focus of attention is on relatively small segments of the disability beneficiary population. In these situations, such as in the context of demonstration evaluations discussed above, administrative records data tend to form the core of the evaluation and survey data might be used in a supplementary role. This is so because SSA's administrative records data systems provide 100 percent coverage and fairly comprehensive data on all persons who have ever participated in the disability program. The experience with the Project NetWork disability case management experiment indicates that administrative records matched to each other are expected to play the core role in demonstration evaluations in the future, and can serve as versatile tools for creating customized analysis samples. Administrative records can play similar roles with respect to evaluation of other policy interventions, focusing on relatively narrow segments of SSA's programs. For example, SSA's evaluation of the effects of the 1996 welfare reform legislation on disabled children and drug addicts and alcoholics are heavily dependent on administrative records.

On the other end of the spectrum are research questions that focus on participation in SSA's disability programs among the general population. Here, national survey data can

play a fundamental role, provided they cover the relevant population segment in sufficient numbers, and contain a reasonable number of observations for participants as well. Administrative records data are of more limited value here, because detailed disability data is typically not available for the nonparticipant segments of the population. Nevertheless matching in administrative records could serve a fundamental role by improving the quality of data on the key participation variable, and also providing an enhanced longitudinal sample of participants. The SSI eligibility simulation we discussed falls into this category. Similar simulations are feasible to assess DI eligibility and participation, or to address a variety of issues related to the effects of the labor market and program environment on participation. The administrative record data can also be utilized in this context to assess the nature and extent of survey attrition. In sum, the use of matched administrative records and survey data will continue to be a promising source of information for disability research and evaluation.

References

Decker, Paul T. and Craig V. Thornton (1995), "The Long-Term Effects of Transitional Employment Services," *Social Security Bulletin* 58(4): 71-81.

Pickett, Clark D. and Charles G. Scott (1996), "Reinventing SSI Statistics: SSA's New Longitudinal File," *Social Security Bulletin* 59(2): 31-56.

Rupp, Kalman, Bell, Stephen H. Bell, and Leo McManus (1994), "Design of the Project NetWork return-to-work experiment for persons with disabilities," *Social Security Bulletin* 57(2): 3-20.

Rupp, Kalman, Michelle Wood and Stephen H. Bell (1996), "Targeting people with severe disabilities for return-to-work: the Project NetWork demonstration experience," *Journal of Vocational Rehabilitation* 7: 63-91.

Rupp, Kalman, Dianne Driessen, Robert Kornfeld and Michelle Wood (1999), "The development of the Project NetWork administrative records data base for policy evaluation," *Social Security Bulletin* 62(2): 1-13.

Vaughan, Denton R. and Bernard Wixon (1991), "Implementing an SSI Model Using the Survey of Income and Program Participation," in *ORS Working Paper Series*, Number 54 "Two Papers on a New SIPP-Based Microsimulation Model of SSI and OASDI," Washington, D.C.: Division of Economic Research, Office of Research and Statistics, Social Security Administration.

Wixon Bernard and Denton R. Vaughan (1991), "Rationale for a SIPP-Based Microsimulation Model of SSI and OASDI," in *ORS Working Paper Series*, Number 54 "Two Papers on a New SIPP-Based Microsimulation Model of SSI and OASDI," Washington, D.C.: Division of Economic Research, Office of Research and Statistics, Social Security Administration.