The Development of the Project NetWork Administrative Records Database for Policy Evaluation

*by Kalman Rupp, Dianne Driessen, Robert Kornfeld, and Michelle Wood**

This article describes the development of SSA's administrative records database for the Project NetWork return-to-work experiment targeting persons with disabilities. The article is part of a series of papers on the evaluation of the Project NetWork demonstration. In addition to 8,248 Project NetWork participants randomly assigned to receive case management services and a control group, the simulation identified 138,613 eligible nonparticipants in the demonstration areas. The output data files contain detailed monthly information on Supplemental Security Income (SSI) and Disability Insurance (DI) benefits, annual earnings, and a set of demographic and diagnostic variables. The data allow for the measurement of net outcomes and the analysis of factors affecting participation. The results suggest that it is feasible to simulate complex eligibility rules using administrative records, and create a clean and edited data file for a comprehensive and credible evaluation. The study shows that it is feasible to use administrative records data for selecting control or comparison groups in future demonstration evaluations.

* The authors are, respectively, Senior Analyst, Division of Policy Evaluation, Office of Research, Evaluation and Statistics, Office of Policy, Social Security Administration; Senior Manager, Fu Associates; Senior Economist, Abt Associates; and Project Director, Abt Associates.

Acknowledgments: The authors are indebted to numerous data processing professionals at SSA, Abt Associates, and Fu Associates for their contributions to the data development reflected in this article. Charlie Scott, Mary Barbour, Jeff Shapiro, Russ Hudson, and Joel Packman of the ORES staff have made especially substantial contributions to the database development effort. For useful comments and suggestions the authors are indebted to Barbara Butrica, Paul Davies, Susan Grad, Howard Iams, Paul O'Leary, Charlie Scott, and Peter Wheeler.

I. Introduction

The purpose of this article is to describe the development of the Project NetWork administrative records database for policy evaluation. Project NetWork has been the first large-scale return-to-work field experiment targeting both the Title II Disability Insurance (DI) and Title XVI Supplemental Security Income (SSI) disabled beneficiary population with severe disabilities. The demonstration was initiated in 1991, to test the feasibility and effects of providing intensive outreach and case management services, and also included work incentive waivers. The demonstration was conducted at eight sites between 1992 and 1994. A comprehensive evaluation component was included in the demonstration design, including the random assignment of 8,248 participants who volunteered for the demonstration to a "Treatment" group receiving case management services, and a "Control" group of persons who did not receive case management services.

The evaluation design called for the collection of data on nonparticipating eligibles at the demonstration sites as well, in order to provide information on the factors affecting the selection of participants among project eligibles—such data being central to the understanding of the role of self-selection and targeting in producing return-to-work outcomes. It also called for the efficient combination of relying on routinely collected SSA administrative records, data from Project NetWork demonstration MIS systems (the Case Management Control System or CMCS[1]), and supplementary survey data collection. It envisioned the followup of Project NetWork participants and nonparticipant eligibles over a period of several years to facilitate the accurate measurement of net outcomes. In addition, case studies were conducted at the eight demonstration sites to provide information on the operational aspects of project implementation. All of these

various sources of information were deemed essential for a comprehensive evaluation of the net outcomes (also referred to as "net impacts") of the demonstration and for the understanding of the institutional processes and selection decisions eventually responsible for the production of these outcomes.

The development of a comprehensive database from SSA administrative records was a key component of this overall evaluation strategy for several reasons:

- SSA administrative records form an excellent source of precise information on two key outcomes of interest in any analysis of return to work among SSA disability beneficiaries: the receipt of DI and SSI benefits. These data are available on a monthly basis, providing complete benefit history during the pre-demonstration period, as well as for the post-demonstration period. If necessary, the post-demonstration period can be extended to the indefinite future at relatively low cost, because SSA continues to maintain benefit receipt information regardless of participation in specific demonstration initiatives. Such high-quality monthly data cannot be easily obtained from other sources, such as surveys, due to the high cost of survey data collection, and substantial nonresponse and recall error.

- SSA administrative records provide information on another important outcome variable–Social Security covered earnings. While the Master Earnings File (MEF) contains only annual earnings information and is limited to Social Security covered earnings, many of the advantages of this data set are similar to the features of the benefit records databases discussed earlier due to its nature as an administrative record.

- SSA administrative records contain important pieces of other information, such as diagnostic variables and basic demographics.

- SSA administrative records cover 100 percent of demonstration participants and can be used to derive information on 100 percent of other relevant universes, such as nonparticipating eligibles. In contrast, survey data collection is constrained by cost and related feasibility considerations, typically resulting in the use of statistical sampling. The use of sampling may have substantial undesirable effects on the precision of estimates, and sample data collection invariably introduces attrition and item nonresponse biases that might greatly complicate analyses. Boruch (1997, p. 185) cites Project NetWork as an example of utilizing complementary information from administrative records and survey data collection in assessing the effects of randomized experiments for planning and evaluation.

The major objective of the Project NetWork administrative records database has been to serve as an important tool for evaluating the Project NetWork demonstration from two perspectives:

- analysis of factors affecting the voluntary decision to participate among demonstration eligibles, and

- analysis of net outcomes experienced by volunteers in terms of earnings and benefit receipt.

In addition, the Project NetWork administrative records database provides a rich array of information items that may facilitate empirical analyses that go beyond the particulars of the narrow objectives of the Project NetWork demonstration intervention. The database, for example, may serve as a tool for future analytic studies related to factors affecting participation in vocational rehabilitation, targeting for employment strategy interventions, or the use of appropriate risk-adjustors in reimbursing providers of vocational rehabilitation services. It can be used for empirical studies testing various nonexperimental methodologies of net impact estimation against experimental evidence. This has been done in the context of other employment and training program evaluations, but never before in the context of interventions targeting severely disabled DI and SSI beneficiaries. Finally, the database development experience itself can be looked upon as a useful model for future studies where administrative records serve as the basis of the evaluation.

In theory, the identification of the universe of cases for the development of an SSA administrative records database is straightforward as long as the Social Security numbers (SSNs) for the universe of interest are known. From a demonstration implementation perspective, however, this is not a trivial requirement, because it presumes the development of a data system that correctly identifies the SSNs of both demonstration participants and nonparticipating eligibles. Service delivery systems do not necessarily conform to this requirement. Oftentimes records are not kept of all demonstration *participants who volunteered*, and evaluators may be particularly concerned about systematic underreporting of participation, especially with respect to people whose participation is limited to a few, low-intensity encounters with the service delivery system, or, perhaps more importantly, differential underreporting of participation by the perceived degree of success (more data on "successful placements" and less on "program dropouts"). Data on project *eligibles* who did not volunteer are even less likely to be systematically collected.

Fortunately, the Project NetWork demonstration included a good quality management information system (CMCS), that allowed for the identification of 100 percent of project *participants* who volunteered and were randomly assigned to "Treatment" and "Control" status. The quality of this information was undoubtedly enhanced by the fact that the SSNs formed the basis of random assignment that was performed offsite by Abt Associates, the SSA evaluation contractor. Offsite automated random assignment by an independent entity provided an extra layer of assurances concerning the integrity of random assignment and the accuracy of recording this crucial piece of information.

Despite the implementation of these quality data collection systems and protocols for *participants* who volunteered, a significant problem arose in the data collection process for *nonparticipants* who did not volunteer. While SSA electronically generated mailing lists, these electronic records were not retained for evaluation purposes during the early phase of the demonstration. Although this problem was corrected for subsequent mailings, the lack of the early mailing list information for purposes of evaluation resulted in a truncated universe of project eligibles.[2] Therefore, a substantial portion of the database development effort reported in this article focused on the simulated re-creation of the universe of Project NetWork eligibles. This was performed by applying the Project NetWork eligibility rules to SSA's national databases. While the need to perform this simulation was rooted in an obvious problem of real-life demonstration implementation, this challenge turned into an intriguing opportunity to test the feasibility of developing analytic samples for policy evaluation by applying fairly complex rules of sample inclusions and exclusions to a national SSA database. Such exercises will be relevant for future policy evaluations, particularly in situations when the retrospective creation of comparison groups from administrative records may be the only feasible way of obtaining data relevant to pressing policy questions.

This article is part of a series of papers on the evaluation of the Project NetWork demonstration. Previous publications included a description of the demonstration design (McManus, Rupp, and Bell 1993), a review of the experimental and evaluation design (Rupp, Bell, and McManus 1994), analysis of self-selection and targeting for participation based on survey and case study data (Rupp, Wood, and Bell 1996), and a summary of the process analysis results (Leiter, Wood, and Bell 1997). Forthcoming pieces will focus on participation among eligibles using the now complete administrative records database and supplementary survey information, net incomes on earnings and benefit receipt, and the overall assessment of the benefits and costs of the Project NetWork demonstration. Several reports prepared under contract with SSA provide detailed documentation of key aspects of the database development effort addressed in this article. Fu Associates (1998a) provides a documentation of the development of SSI benefit variables and Fu Associates (1998b) describes the derivation of DI benefit variables. Abt Associates (1998) documents the derivation of eligibility, diagnostic, and demographic variables.

The database development effort that forms the basis of this article reflects the work of teams of analysts, systems analysts, and programmers at the Social Security Administration, Abt Associates, and Fu Associates over a period of several years.

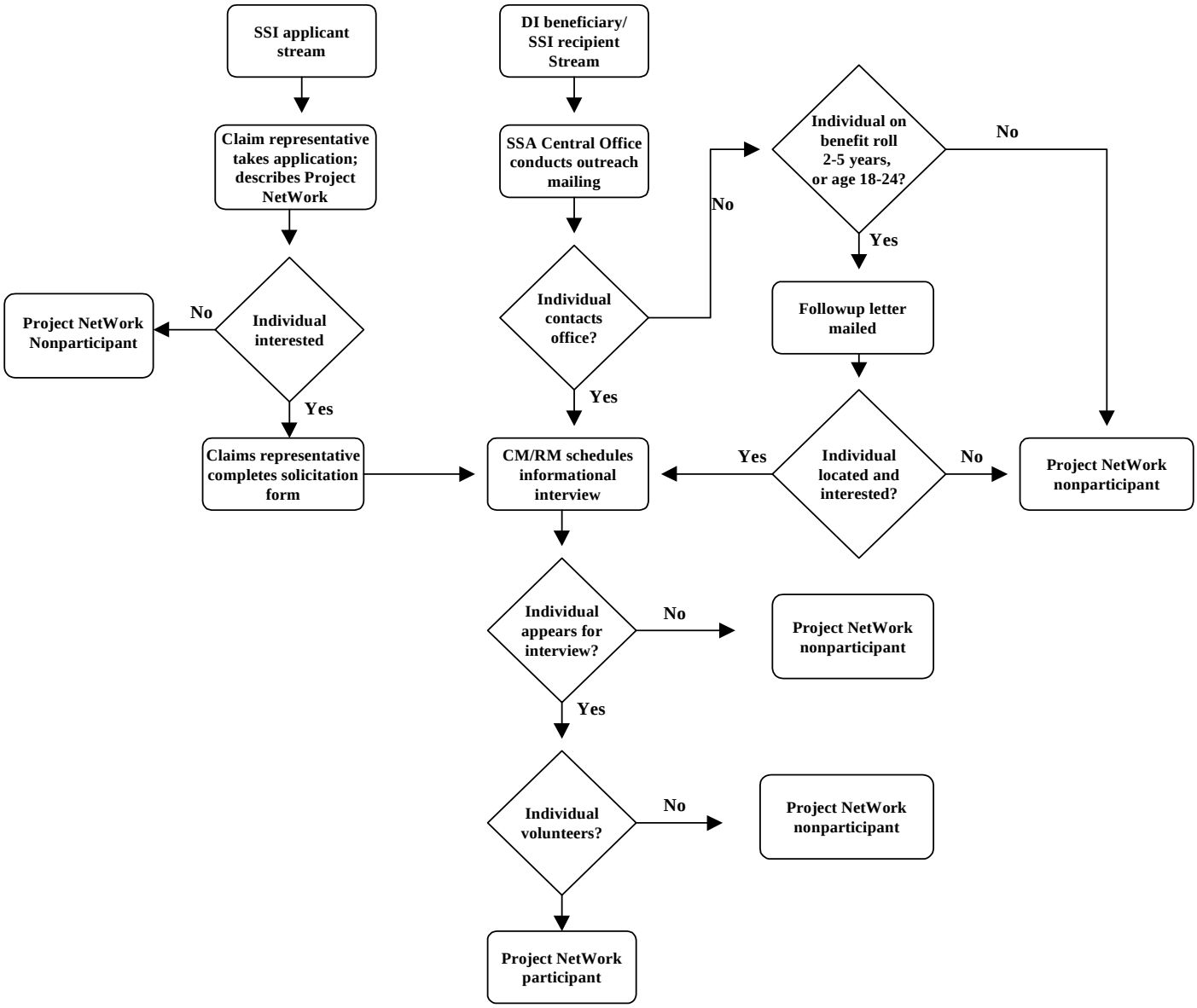
This article is organized as follows. In the next section we describe the study universe of interest (Project NetWork eligibles and participants), the methods, and results of the eligibility simulation used to derive the universe of SSNs included in the final analysis files. Next we describe the major analytic variables and their derivation. Then, we provide a brief description of linked survey data files. We conclude the article with a discussion of the utility of the database and database development methods for policy evaluation and implications for future database development activities focusing on the evaluation of current or planned SSA return-to-work demonstrations.

II. Derivation of Study Universe: Project NetWork Eligibles and Participants

A key objective of the database development was the derivation of the appropriate universe for the Project NetWork administrative records database. Conceptually, our goal was to identify the universe of SSNs containing all Project NetWork eligibles and participants at the demonstration sites. Identifying Project NetWork participants by SSNs was easy, because the management information system (the CMCS) identified all participants assigned to treatment and control status. The most important challenge of the database development was to identify the universe of Project NetWork eligibles, regardless of participation status. Without an appropriate universe of eligibles, a substantial portion of the planned analyses would not be feasible. In particular, deriving a consistently defined universe of participant and nonparticipant eligibles was to be an important pre-condition of analyzing factors affecting participation in the demonstration.

In our simulation, we attempted to identify those individuals who were directly solicited to participate in Project NetWork during the demonstration period, and to identify the date of solicitation as a reference time point in analyzing the decision to participate in Project NetWork. The solicitation strategy was based on the empowerment outreach philosophy of the demonstration: DI beneficiaries and SSI disability applicants were to be solicited for the demonstration regardless of factors, such as age and the severity of disabling conditions, that were believed to have been used by the traditional vocational rehabilitation system to "screen out" people labeled to be poor candidates for successful return-to-work interventions. Thus, Project NetWork was meant to use a wide net to "screen in" a full cross-section of DI beneficiaries and SSI disability applicants. The starting point of the development of our eligibility simulation strategy was the recruitment and intake process (chart 1). This process is analyzed in detail by Rupp, Wood, and Bell (1996). The essential observation for purposes of simulating eligibility for Project NetWork is that the outreach and recruitment effort focused on two major sources of eligibles: SSI applicants and SSI/DI beneficiaries. These two groups (represented by boxes at the top of chart 1) were subject to individual-level outreach through a variety of techniques–information provided by claim representatives to SSI applicants, outreach mailing focusing on DI and/or SSI beneficiaries, and targeted followup mailings to some beneficiaries. We note that there was a third group, newly awarded DI beneficiaries, who were subject to some level of individualized outreach, which was not explicitly considered in our eligibility simulation. These people were to have been informed of Project NetWork through a telephone call or letter mailed by the local field office. We did not have a flag identifying newly awarded DI beneficiaries. However, all awardees,

Chart 1.—Project NetWork recruitment and intake process



including new beneficiaries, were considered in the eligibility simulation.

The principal routes towards Project NetWork participation, as represented by chart 1, included a sequence of self-selection and targeting decisions that started with these two broad target groups of SSI applicants and SSI/DI beneficiaries. We note that the sites also conducted outreach activities that were not individually targeted (for example, media advertisements), and also accepted walk-ins. Nevertheless, these alternative routes to program participation were thought to be supplementary to the individualized outreach targeting SSI applicants and SSI/DI beneficiaries. In our eligibility simulation, we were focusing on SSI applicants and SSI/DI beneficiaries who were subject to individualized outreach of some variety. We note, however, that our eligibility algorithm was consistently applied to both participants and nonparticipants, and therefore we are able to identify participants who did not satisfy the simulated eligibility rules. Some of these may have been individuals who were not subject to the individualized outreach, but may have learned about the demonstration through other means.

Since our eligibility simulation is anchored to the distinction between the SSI/DI beneficiary and SSI applicant streams, our database development strategy had to be informed by the basic rules of solicitation for these two streams of eligibles. Chart 2 describes the schedule of mail solicitations to SSI/DI beneficiaries for each site, while chart 3 describes the schedule of solicitation for new SSI applicants.

As can be seen from chart 2, the mail solicitations were based on demonstration site and the last digit of the SSN. The last digit of the SSN was used to assure an even flow of volunteering throughout the Project NetWork intake period at the given demonstration site. Assuming a constant stock of beneficiaries, the basic design of the mail solicitations would have assured that all beneficiaries were solicited at some point, albeit on different targeting dates—the solicitation date was to differ based on the last digit of the person's SSN. However, deviations from this simplified conceptual model could occur partly because the assumption of a constant stock of beneficiaries is incorrect due to people moving in and out of beneficiary status over time, and also because of migration in and out of the demonstration area. Moreover, there were deviations in the actual implementation from the simple model of the basic design of the mail solicitation outreach effort. A variety of such deviations are noted in the footnotes to chart 2. Additional variation in mail solicitations may have been introduced through likely variations in the actual date of computerized file extraction relative to the target date represented in the exhibit for each mailing, and perhaps in the nature of the source file used for the mailings.

Chart 2.—Mail solicitations of existing SSI recipients and DI beneficiaries, by extract date, last digit of SSN, and site

June '92[1] SSN= 0,9	Sept. '92[1] SSN= 1,8	Dec. '92[1] SSN= 2,7	Mar. '93[1] SSN= 3,6	June '93 SSN= 4,5	Sept. '93 SSN= 5,9	Dec. '93 SSN= 1,8	Mar. '94 SSN= 2,7
Dallas	Dallas	Dallas	Dallas	Dallas			
Ft. Worth	Ft. Worth	Ft. Worth	Ft. Worth	Ft. Worth			
		Minn.	Minn.	Minn.	Minn.	Minn.	
		Phx/LV	Phx/LV[2]	Phx/LV	Phx/LV[3]	Phx/LV	
		N.H.	N.H.[4]	N.H.[5]	N.H.	N.H.	
			Rich.	Rich.	Rich.	Rich.	Rich.[6]
		Tampa	Tampa	Tampa	Tampa	Tampa[7]	Tampa[8]
		Spk/CdL	Spk/CdL	Spk/CdL	Spk/CdL	Spk/CdL[7]	

Note: Ft. Worth = Ft. Worth, TX; Minn= Minneapolis, MN; Phx/LV= Phoenix, AZ and Las Vegas, NV; N.H.= New Hampshire; Rich= Richmond, VA; Spk/CdL= Spokane, WA and Coeur d'Alene, ID.

[1] Partial mailings for selected ZIP codes may have been sent in the sites shown.

[2] Mailings may not have been sent in Phoenix/LV and Coeur d'Alene.

[3] Only selected ZIP codes were mailed from this extract: 85032 and below in Phoenix, and 89109 and below in Las Vegas.

[4] May have sent letters only to SSI recipients.

[5] Only selected ZIP codes were mailed from this extract: all ZIP codes for Keene, Littleton, Manchester, and Nashua, plus 03603, 03743, 03773.

[6] May have sent letters to only current SSI recipients between 15 and 18 years old, plus those on the rolls 2 to 5 years, for *all* SSNs.

[7] Individuals solicited for the pilot demonstration may have been excluded.

[8] Unclear whether any mailing occurred; if so, only for SSI recipients age 15-18 and SSI recipients on the rolls 2-5 years, for *all* SSNs.

As chart 3 shows, the solicitation of new SSI applicants was a function of demonstration site (district office) and the date of SSI application, with some uncertainty concerning the actual solicitation period during implementation at selected sites.

The location, solicitation period, and SSN ending digit information included in charts 2 and 3 are key to the establishment of Project NetWork eligibility, and to the simulation of the universe of eligible SSNs and the solicitation reference date. Importantly, the key eligibility variables are available in SSA administrative records systems.

The actual development of the simulation of eligibles included several major steps:

- creation of a source file of "potential" Project NetWork eligibles and participants,
- cleaning of key variables relevant to the eligibility algorithm,
- development of relevant eligibility flags, and
- development of analytic universe.

A source file of "potential" Project NetWork eligibles and participants was derived from various sources. The objective at this stage was to assure that the SSNs of all potential eligibles were included: therefore the net was to be set wide at this stage, even at the expense of including SSNs of persons who later may be judged not eligible upon a more rigorous and refined application of eligibility rules. The three principal sources of data at this stage included the Supplemental Security Record (SSR), the Master Beneficiary Record (MBR), and a "finder file" assembled by Abt Associates.

The first two of these data files also included information from SSA's application record system (National Disability Determination System or NDDS—commonly known as the "831" system). SSA's national data systems were searched to identify SSNs of persons who were either on the DI or SSI rolls or who were SSI applicants at the demonstration sites during the relevant period of time for mail or applicant solicitation. The process used to search the SSR and MBR systems differ somewhat, in part as a reflection of the differences in the data systems. For example, to identify persons residing in the relevant catchment areas of the demonstration, the MBR search used field office codes, while the SSR search used ZIP code information. The SSR search resulted in the identification of 171,645 potentially eligible person records, while the MBR search identified 170,972 records.

The third principal source of information was the Abt Finder File. Abt Associates assembled this file, containing 101,519 records over the course of their evaluation from a variety of

Chart 3.—New SSI applicants solicited for Project NetWork, identified, by district office code and date of application

Project NetWork site	District office codes	Date of SSI application[1]
Dallas	757, 798, 813, 814, 819, 835, 851, 853, 854, A73, A76, B44, D20, E64, M97, M99	June 1, 1992 to Sept. 30, 1993
Forth Worth	816, 821, 836, B47, E35, M98	June 1, 1992 to Sept. 30, 1993
Minneapolis	544, 675, 676, 677, 678, 679, 683, 685, C39, C40, D22	Oct. 20, 1992 to Sept. 6, 1998 Oct. 19, 1993 to April 7, 1994[2]
Phoenix	367, 455, 864, 907, 911, 913, 914, 929	Oct. 23, 1992 to July 6, 1993; and Aug. 31, 1993 to Jan. 5, 1994
Las Vegas	059, 459, 908, 945, 946, A10, D49, I61	Oct. 23, 1992 to June 1, 1993; and Aug. 2, 1993 to Jan. 5, 1994
New Hampshire	010, 011, 012, 013, 014, 015 035, 041, 050, 084, 954	Oct. 13, 1992 to April 15, 1994[3]
Richmond	285, 293, 300, 303, A33, A38 A41, C82	Jan. 6, 1993 to June 29, 1994[4]
Spokane	868, 915, 918, 924, 927, Q50	Oct.13, 1992 to March 17, 1994
Coeur d' Alene	189, 872, 893, 895, 898, 953 Q51	Oct. 13, 1992 to March 26, 1994
Tampa	257, 266, 656, 657, 658, 659, 665, 666, 673, 674, 867, 949, A43, A98, C19, Q48, Q49	Oct. 13, 1992 to April 1, 1994

[1]It is unclear whether solicitations of new SSI applicants occurred during the pilot phase in each site. Except for Dallas and Fort Worth, the starting dates for the pilot are listed in each site as the earliest possible starting point for solicitation. Solicitation intervals are inclusive of the listed dates.

[2]Minneapolis may not have solicted during all or part of December 1993.

[3] Expected solicitation stop date. Actual stop date never reported to evaluation contractor. Random assignment ended April 30, 1994.

[4] Expected solicitation stop date. Actual solicitation stop date never reported to evaluation contractor. Random assignment ended June 30, 1994.

sources. Importantly, the Abt Finder File, included all Project NetWork participants randomly assigned to "Treatment" and "Control" status. The file had substantial known limitations, however, in terms of the coverage of the universe of nonparticipating eligibles as a result of the lack of access to electronic mailing list information previously discussed.

The net result was a source file[3] of SSNs covering 210,226 potentially eligible person-level observations.

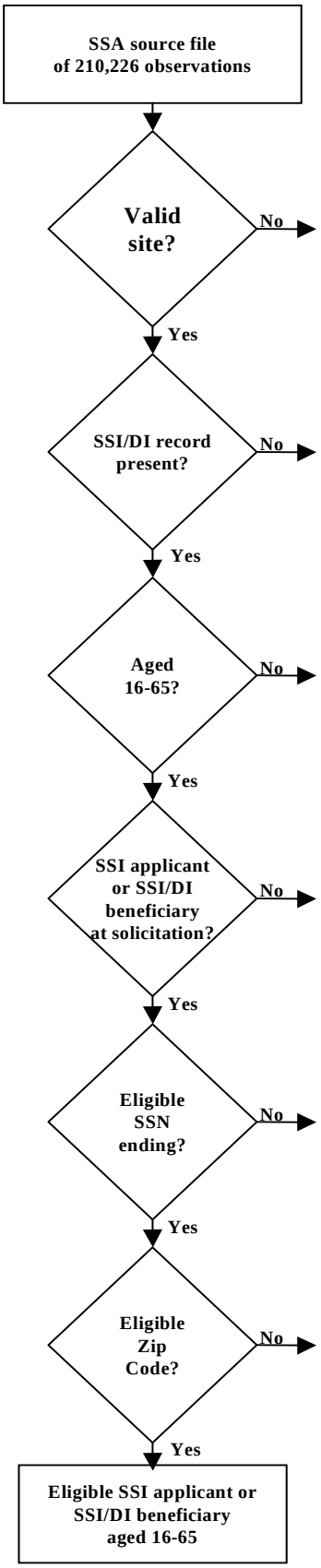
Our next step was the cleaning and editing of analytic variables. This effort, which is described in more detail later in this article, included a vast array of analytic variables, most of them having no implication for the eligibility simulation. However, a few variables were important for our eligibility simulation. These included demonstration site, age, and variables relevant for identifying a case as a potential mailing list or potential SSI applicant case.

After cleaning and editing the relevant variables, we developed a series of flags that are necessary to derive the pool of simulated eligibles, and performed the eligibility simulation. Chart 4 provides a schematic description of the eligibility simulation.

The top of table 1 describes the source file by the eligibility screen variables and the bottom of the table gives the subfile of Project NetWork participants by our eligibility screen variables.

Let us first discuss the top portion of the table, focusing on the whole source file. It shows how the final file of 145,410 participants and simulated eligibles was derived from the source file of 210,226 potentially eligible records. It also shows the number of cases surviving each screen independently (column 1) and conditionally on surviving all previously considered screens–the cumulative total surviving up to the screen identified in a given line (column 4). The first step was to check whether there was a valid site code identifying one of the

Chart 4.—Project NetWork eligibility algorithm



demonstration areas. Less than 2 percent of the source file failed to survive this screen. Then we checked whether a valid SSI/DI record was present and found that 5.5 percent of the source file failed this screen. Less than 5 percent of the source file was outside of the 16-65 age range at the time of solicitation.

The most important reason for screening out observations from the source file was failure to have a record of either an SSI application or evidence of SSI/DI benefit eligibility at the time of solicitation. More than 20 percent of the source file failed this screen. Column 5 of table 1 shows that when the various eligibility screens are applied in a sequential fashion, the largest drop is associated with this variable. It is to be noted here that the SSI/DI beneficiary status variable is based on a concept of "payment eligibility" reflecting SSA's assessment of whether the person was entitled to receive benefits during the given month. It is possible that some of these people actually received a payment during the solicitation reference month either as a result of payments that were subsequently deemed to have been made in error (overpayments), or as a result of payments that were made during the reference month to correct a previous underpayment. (See Pickett and Scott (1996) for a comprehensive discussion of the broader issue of "payment eligibility" versus "actual payments.") The last two screen variables reflect minor refinements in the eligibility algorithm to account for the fact that certain site and SSN combinations were excluded from the mailings, and certain mailings were limited to a subset of ZIP codes in the demonstration areas.

While the top of table 1 applies the eligibility screens to the whole universe of potential Project NetWork eligibles, the bottom applies the same screens to the group who actually participated in Project NetWork. This is an important test of the eligibility algorithm because the intuitive expectation is that those who participate in Project NetWork should be deemed eligible by program operators. Looking at this data also provides some sense of how the necessarily mechanistic rules of the eligibility algorithm may relate to the way the demonstration was actually implemented.

While one should expect Project NetWork participants to pass the eligibility screens imposed, the expectation of a 100 percent "hit rate" would be naive. There are a number of reasons why legitimate discrepancies may arise. For example, the demonstration solicited "walk-ins" through general media advertisements and other tools. Also, the date of random assignment typically followed the solicitation with some lag—this implies changes in age and possibly beneficiary status. In addition, as previously mentioned, the SSI and DI data files were based on the concept of "payment eligibility," and not on "actual payments." Finally, the "solicitation reference date" developed for our algorithm assumed that the data files were pulled in a systematic fashion for the mailing anchored exactly to benefit status during the month prior to the presumed mailings. In fact, we have no evidence that this actually occurred in an exact manner, and it is unclear whether the files used to generate the mailings reflected the same beneficiary status concepts that were used to generate the file we worked

with. A lag of a few months between pulling the files and actual mailings could account for such discrepancies.

The bottom section of table 1 shows that the file of participants survived most screens with only minor attrition. Of all participants, 254 cases did not have an SSI/DI record present, representing a little over 3 percent of participants. Some of these may have been referrals from other agencies, but it is also possible that some denied DI applicants participated. In any event, the number is very small. With few exceptions, participants survived the age screen and the SSN-ending and ZIP code related refinements.

The single major source of participants not surviving the eligibility screening process is that 1,111 participants (13.5 percent of the total) had no evidence of either an application or an SSI/DI benefit record during the month prior to solicitation. This is the same factor that caused the largest drop in simulated eligibility from the source file. The drop could also be attributable to several additional factors. As a practical matter, the most important reason relates to timing. The vast majority of this group were either SSI applicants or beneficiaries at the time of random assignment. A more detailed analysis indicate that only about 5 percent of participants were neither applicants nor beneficiaries at the time of random assignment—and some of these may have applied or received benefits at some other point in time during the demonstration period. Thus, considering beneficiary status at random assignment instead of solicitation reference date for participants alone would bring the rate of estimated eligibility among participants to well over 90 percent.

Overall, more than 82.4 percent of participants survived the criteria for simulated eligibility. Thus, 17.6 percent of participants were classified as ineligible by the simulation. The following discusses the reasons.

To gain insight into the nature of differences between "simulated eligible" participants and "simulated ineligible" participants, we looked at a selected number of characteristics for these groups, as well as simulated nonparticipant eligibles. Table 2 shows the empirical results. We focus here on the differences between the two groups of participants.

Notably, the New Hampshire site is substantially overrepresented among "simulated ineligible participants." New Hampshire was especially active in enrolling other agency referrals; this supports the notion that one reason for simulated ineligibility is referrals from other agencies of a small number of individuals who may not have had previous exposure to the SSA system. Importantly, from the point of view of using the data for experimental impact estimates, the "Control" and "Treatment" assignment of the two groups of participants is close to 50-50 as expected.

Overall, the demographic and impairment information for the two groups of participants is comparable. The one exception relates to the tendency for "simulated ineligible" participants to have more missing values. This is expected, because the program involvement of this group is weaker as evidenced by the lack of benefit records or lack of program eligibility during the month of solicitation. For age and gender where the consideration of MEF information was available, in addition to the program files (MBR, SSR, and NDDS), as a substantial source of deriving the edited variables presented here, the rate

Table 1.—Distribution of SSA Project NetWork source file and subfile of participants, by eligibility screen variables

Eligibility screen variable	Number passing (1)	Number failing (2)	Percent of total passing (3)	Cumulative total surviving (4)	Cumulative percent of total surviving (5)
	Source file				
Total	210,226	. . .	100.0	210,226	100.0
Valid site	206,232	3,994	98.1	206,232	98.1
SSI/DI record present	198,684	11,542	94.5	198,684	94.5
Aged 16-65	201,287	8,939	95.7	194,672	92.6
SSI applicant or SSI/DI beneficiary at solicitation	167,638	42,588	79.7	148,815	70.8
Eligible SSN ending	203,371	6,855	96.7	148,815	70.8
Eligible ZIP code	204,965	5,261	97.5	145,410	69.2
	Subfile of participants				
Total	8,248	. . .	100.0	8,248	100.0
Valid site	8,248	. . .	100.0	8,248	100.0
SSI/DI record present	7,994	254	96.9	7,994	96.9
Aged 16-65	8,219	29	99.6	7,975	96.7
SSI applicant or SSI/DI beneficiary at solicitation	7,137	1,111	86.5	6,874	83.3
Eligible SSN ending	8,202	46	99.4	6,874	83.3
Eligible ZIP code	8,145	103	98.8	6,797	82.4

of missing information is very small, and the rest of the distribution is very close for the three groups. In the derivation of the other variables, the MBR, SSR, and NDDS files played a substantial role. Therefore, it is not surprising that the rate of missing information for "simulated ineligible" participants is clearly higher than for "simulated eligible" participants. It is notable, that the rate of missing information for "simulated ineligible" participants is only slightly higher than for nonparticipant eligibles for education and impairment.

Given the availability of fairly straightforward legitimate explanations for simulated noneligibility among some participants, our analysis shows that the eligibility simulation performed fairly well, especially when beneficiary status at random assignment is considered in the assessment.

III. Development of Administrative Records Analytic Data Files

The development of SSA analytic files involved four major groups of variables:

- SSI benefit variables,
- DI benefit variables,
- demographic and diagnostic variables, and
- MEF earnings variables.

In the creation of each set of analytic variables the appropriate source files have gone through various steps of checking, cleaning, and editing. Output files containing the appropriate set of analytic variables were derived. Essentially, four major output files were created corresponding to the above four groups of variables and were subsequently used for various analytic purposes matched to each other and other files based on the SSN identifier of each individual. The following provides a description of the creation of these major output files.

SSI Benefit Analysis File

The SSI benefit analysis file was created according to specifications from ORES and was implemented by a team of analysts and programmers at SSA, Abt, and Fu Associates. The output file was prepared and detailed documentation was provided to SSA by Fu Associates (1998a), under a subcontract with Abt Associates. The following provides major highlights outlined in the report to SSA.

The objective was to create an analysis file that provides detailed information on monthly SSI benefits for all participants and potentially eligible persons with involvement with the SSI program between January 1974 and December 1996. The output file contained 171,645 observations and 521 analytic variables. The key output variables of interest included monthly SSI payment status (in terms of payment eligibility), optional state supplementation, SSI federal assistance amount, and amount of state supplementation, and chargeable earned and

Table 2.—Project NetWork: Percent of simulated nonparticipant eligibles and simulated eligible and ineligible participants, by basic characteristics

[Percentages may not add to 100 do to rounding]

Basic characteristics	Nonparticipant simulated eligible applicants and beneficiaries aged 16-65	Participant simulated eligible applicants and beneficiaries aged 16-65	Simulated ineligible participants[1]
Total number	138,613	6,797	1,451
Site			
Total percent	99.9	99.8	99.9
Cour d'Alene	2.4	2.9	2.2
Dallas	11.8	14.4	11.6
Forth Worth	9.1	9.5	7.4
Las Vegas	6.6	4.5	5.2
Minneapolis	16.5	13.2	7.9
New Hampshire	6.4	7.9	37.6
Phoenix	10.5	7.9	12.2
Richmond	12.7	15.6	4.6
Spokane	8.2	9.5	4.4
Tampa	15.7	14.4	6.8
Random assignment status			
Total percent	100.0	100.0	100.0
Control	0	49.3	50.9
Treatment	0	50.7	49.1
Neither	100.0	0	0
Age			
Total percent	100.0	99.9	100.2
Missing data	0	0	0.3
Less than 16	0	0	.7
16-17	1.3	.5	.9
18-24	7.9	8.7	9.9
25-40	35.4	44.4	40.6
41-61	55.1	46.0	45.9
62-65	.3	.4	.7
66 or older	0	0	1.2
Gender			
Total percent	100.0	100.0	100.0
Missing data	0	0	1.8
Male	44.7	42.1	41.2
Female	55.3	57.9	57
Race			
Total percent	100.0	99.9	100.1
Missing data	2.7	2.6	22.3
Black	25.2	28.5	15.9
Other	4.9	3.7	3.7
White	67.2	65.1	58.2

See footnote at end of table.

Table 2.—Project NetWork: Percent of simulated nonparticipant eligibles and simulated eligible and ineligible participants, by basic characteristics—*Continued*

[Percentages may not add to 100 do to rounding]

Basic characteristics	Nonparticipant simulated eligible applicants and beneficiaries aged 16-65	Participant simulated eligible applicants and beneficiaries aged 16-65	Simulated ineligible participants[1]
Education (in years)			
Total percent	100.0	100.0	100.0
Missing data	46.2	42.3	54.4
Less than 9	9.0	6.2	4.8
9 to 11	14.3	12.7	9.6
12	22.6	26.1	21.6
13 or more	7.9	12.7	9.6
Impairment			
Total percent	99.9	99.7	99.8
Missing data	10.4	6.1	12.8
Infectious and parasitic	3.4	3.6	1.4
Neoplasms	2.5	1.5	1.3
Endocrine and metabolic	4.1	4.5	4.3
Blood and blood forming organs	.3	.5	.6
Schizphrenia	7.8	10.7	12.3
Psychoses and neuroses	20.9	24.2	22
Mental retardation	9.4	7.3	8.8
Central nervous system	5.4	5.6	6.3
Diseases of the—			
Eye	2.5	2.7	2.1
Ear	1.2	1.2	1.2
Circulatory system	7.1	6.7	6.1
Respiratory system	2.5	2	1.7
Digestive system	1.3	.9	.8
Genitourinary system	1.3	2.3	1.7
Skin and subcutaneous	.2	.2	.1
Musculoskeletal system	13.6	12.7	10.6
Congenital anomalies	.3	.4	.3
Injury	5.7	6.6	5.4

[1] Includes participants who failed the 16-65 age restrictions but may have passed the other eligibility screens.

unearned income. Monthly values were to be provided from January 1990 through December 1996. Summary measures of SSI benefit history prior to January 1990, and a limited number of other variables of interest such as SSI application dates, were also provided.

The source data file contained SSR and matched NDDS data. It contained one record per possible Project NetWork eligible individual as uniquely identified by a Personal Account Number (PAN). The record layout included data variables for up to four applications for each PAN with data unique to that application for SSI benefits, including the SSI Computational History Data describing computed benefits[4] for all months from January 1974 through December 1996. The file also included matched data from the NDDS file (commonly referred to as the "831" file), including detailed diagnostic information that was used in other parts of the analysis.

A substantial amount of data exploration was conducted to check the quality of the source data, identify the presence of invalid data items, and to provide a sound basis for the creation of the final, clean, and edited output file. A series of intermediate files were used to perform diagnostic procedures. The intermediate files, with some data manipulation, were eventually used to generate the analysis file.

The diagnostic procedures, in general, confirmed the high quality of key variables of interest. For example, all date variables were found to have valid data in the year, month, and day fields. Only 101 of 29,373,686 (0.00034 percent) payment status code fields had invalid values. The key payment amount fields all contained legitimate values. The optional state supplementation code contained blank values for 3.1 percent of the cases. Diagnostics on data variable relationships found only a few inconsistencies, none involving the core set of variables. No application date was missing, and only a small number of cases had an application date of zero.

The creation of the final analysis file used a set of consistent rules concerning default values according to ORES specifications. The same default values were used in subsequent merges of this subfile of about 172,000 cases to the total universe of about 210,000 cases. Most importantly, this assured that the relevant SSI payment amounts were consistently set to zero for all cases that were not included in the SSI benefit analysis file of about 172,000 cases (for more detailed documentation, see Fu Associates (1998a)).

DI Benefit Analysis File

In general, the creation of the DI benefit analysis file had similar objectives to the creation of the SSI benefit analysis file. The objectives for the output file were specified by ORES. Fu Associates performed most of the data processing, with substantial input from ORES programmers. The process, quality checks, edits, and results are documented by Fu Associates (1998b) in detail.

The objective was to create an analysis file to provide detailed monthly information on DI benefits and related variables. The file also includes a few additional variables that were utilized in subsequent steps in the data processing. The output file contained 147,868 observations with 202 analytic variables. The main source of information was an extract from the MBR, containing 48 history segments describing DI benefit history with effective dates and eligibility status of individuals from January 1962 to December 1997. Some other files were also used in the analysis. Most notably, ORES provided a separate data file providing the date of old-age conversion, if

any, for each record. Generally, the key source variables performed well on the data quality checks (Fu Associates 1998b).

The major variables of interest in the output file included monthly benefit amount and payment status indicators from January 1990 through December 1997, summary measures of pre-1990 DI benefit experience, and information on the year and month when a beneficiary is expected to reach the early and regular retirement date based on date of birth, and the date of old-age conversion.

Analysis File of Demographic, Diagnostic, and Project NetWork Eligibility Variables

The creation of the final edited demographic and impairment data, as well as the finalization of the Project NetWork eligibility variables, was the result of a collaborative effort of a number of analysts and programmers at ORES and Abt Associates. The main results are documented by Abt Associates (1998). In addition to the description of the creation of demographic variables (except the variable "age" that was created on the SSA mainframe due to the use of MEF records), the report also summarizes the main eligibility variables discussed elsewhere in this article. The final file contains 33 variables and 210,226 observations.

The key demographic variables of interest are age, gender, race, level of education, and type of impairment. While the number of variables is small, a key challenge of this part of the data development work arose as a result of (a) multiple files as possible sources of these variables and (b) missing, conflicting, and potentially inaccurate source variables, especially compared to the completeness and high accuracy of SSI and DI benefit information.

The general strategy applied here was to consider the various potential sources of the given information item and, when conflict arose, use of a logical and consistent process of identifying and using the "best" source of information available. The major data files included extracts from the SSR, MBR, and NDDS systems (and for determining date of birth, the MEF extract). Given that multiple application records are imbedded in the source files, some variables used a large number of potential source fields. For example, the number of potential source fields for primary and secondary impairment was 22 for both variables. In general, the rules considered our judgement as to the likely quality of each field based on administrative context.[5] The rules we established also reflected the timing related to the program intervention (preference was given to the closest impairment information preceding solicitation for Project NetWork). A detailed description of the specific rules applying to each variable is contained in Abt Associates (1998).

As a result of having multiple sources of information, the proportion of missing values was reasonably low for both participants and nonparticipants on the key demographic variables, and on impairment,[6] as was previously reported in table 2.

MEF Earnings Records File

The MEF Earnings Records File was created by SSA staff using standard procedures. This provides summary annual earnings information capped at the Social Security maximum for each year up till 1996 for all 210,226 observations of potential Project NetWork eligibles and participants. In order to provide an opportunity for Abt Associates to do some offsite data processing of earnings information for the participation, net impact and waiver effect analyses within the framework of Privacy Act restrictions, a separate data file containing grouped observations of 10 to 17 cases with bottom coding of low means and standard deviations was also created. As it turns out, this grouped information proved to be unsatisfactory for most of the net impact analysis, as a result of noise introduced by nonhomogenous grouping for important disaggregated analyses. As a result, all of the final net impact analysis data processing was performed by ORES on the SSA mainframe computer. The resulting tables contain aggregate impacts and satisfy Privacy Act disclosure restrictions. This information was shared by ORES with Abt analysts and became a major source of information to include in the final net impact analysis report.

IV. Linked Survey Data Files

The information from the administrative records database has been supplemented by two waves of survey data collection to provide information on important economic and noneconomic variables that are not available in administrative records data files but are believed to have potentially important effects on participant selection and/or outcomes. The survey is documented by Abt Associates (1997). The following provides a brief description of these supplementary survey data files, and discusses the complementary role of administrative records and survey data in the evaluation.

The baseline survey was designed under a contract with Lewin-ICF, Inc. (now The Lewin Group), and it contains questions about education and training, health, functional and activity limitations, employment history, and knowledge of SSA's work incentives for disability beneficiaries. The survey also contains a wide array of questions about emotional stability, drug/alcohol use, and cognitive functioning.

The followup survey questionnaire was designed by Abt Associates and contains questions on health and functional limitations, education and receipt of training and rehabilitation services, transportation and child care, employment, personal attitudes and outlook, and income and benefits. In particular, the followup survey provides information about the receipt of income other than earnings and SSI and DI benefits, such as food stamps, housing assistance, workers' compensation benefits, retirement or survivor income, and unearned income. Together with the baseline survey, the followup survey provides measures of changes in self-esteem, depression, and attitudes toward work. The followup survey also provides respondent assessment of Project NetWork and measures of the extent to which participants understand the rules determining SSI and DI

benefit levels and eligibility, and what effect the demonstration waivers have on these rules.

Interviews were conducted in person with the use of computer-assisted personal interviewing (CAPI) techniques. A total of 3,439 baseline interviews were conducted immediately after the participation decision. A total of 3,439 baseline interviews were completed, including 2,555 with treatment and control group members who volunteered, and 884 with nonparticipating eligibles. Response rates were relatively high for participants (87 percent), and low for nonparticipants. For nonparticipants the response rate was 53 percent for existing beneficiaries, and only 49 percent for new SSI applicants. The followup survey was designed to cover only participants (treatment and control cases). Altogether, 1,521 followup interviews were completed for a final response rate of 83 percent.

The use of survey data in the evaluation was designed to serve to complement the SSA administrative records data system. The respective roles of the two types of data can be summarized as:

- The administrative records data served as a basis of the demonstration outreach and the development of the CMCS management information system. It also was used as the sample frame for the survey. The 100-percent coverage of all project participants and nonparticipating eligibles in SSA's administrative records systems was an important feature facilitating both demonstration implementation, and measurement of evaluation outcomes.

- The administrative records system was the source of information for the eligibility simulation. Such information would not have been feasible to obtain through a survey.

- The administrative records served as the basis of information on key demographic and impairment variables and central outcomes, including the receipt of monthly DI and SSI benefits and covered earnings. The lack of attrition (100-percent coverage) of administrative records and the high accuracy of administrative records information on the key outcomes were essential ingredients of the success of accurate measurement on key outcomes. The estimates from administrative records were also statistically very accurate due to the large number of observations arising from the 100-percent coverage, and the lack of attrition bias in longitudinal followup.

- The administrative records served as an important tool of assessing the potential statistical estimation problems arising from survey sampling, nonresponse, and survey attrition.

- The administrative records data had limitations in terms of the type of information available. Generally, the quality of payment-related information was highly accurate, but the availability of information on variables that are less important in the administrative process (for example, education) was more limited. For many information items the usefulness of the administrative records was limited by factors such as the timing of data collection. For example, many data items collected by SSA are limited to application for disability benefits. Administrative records earnings information is limited to covered employment, and no information is available on hours and wage rates. Finally, a large number of variables of potential interest, such as those related to motivation, attitudes, and noneconomic outcomes are not available from SSA administrative records.

- The survey data, as described above, provided supplementary information on a large number of important variables that enhanced the overall evaluation.

- The survey data were very useful in providing detailed descriptive information on a large number of noneconomic variables and providing detail on relevant economic variables, such as hours and wage rates.

- The usefulness of survey data was limited by relatively small sample sizes, sample attrition, recall error, and the inclusion of only a single wave of followup for the subset of participants. These limitations were particularly substantial in the context of the measurement of net outcomes and for describing the nonparticipant portion of the eligible population.

- In theory, the usefulness of survey data could be increased by increasing the sample size, number of followup interviews, and other enhancements, but such enhancements might be prohibitively expensive, while the marginal cost of additional administrative records observations is essentially zero.

In sum, the administrative records and survey data had complementary roles in the comprehensive evaluation of Project NetWork. Many of the lessons learned from the Project NetWork experience in this regard are relevant for the design of planned future demonstration evaluations.

V. Conclusion

This article describes the development of the Project NetWork administrative records database for policy evaluation. The results suggest that it is feasible to simulate complex program eligibility rules using administrative records, and to create a clean and edited data file containing detailed background and outcome information relevant for the evaluation of return-to-work demonstrations, including factors affecting participation, subgroup characteristics, and detailed outcome data on key variables of interest to policymakers. The work that has been conducted involved complex database management and editing, and required the translation of programmatic rules of program eligibility to operationally measurable concepts using administrative records data files.

Beyond the contribution of these data files to analyses of

participation and the net outcomes of the Project NetWork demonstration, useful lessons can be learned from this experience for the design and implementation of future demonstration evaluations. In particular, many of the currently planned or anticipated SSA demonstration initiatives involve the potential use of complex selection rules defining appropriate comparison groups or sites and the use of the same or similar source files as have been used in the Project NetWork evaluation. The results of the Project NetWork database development effort reflected in this article shows that it is feasible to use administrative records for the purpose of creating comparison groups on a scale, and degree of flexibility and complexity that would be either not feasible or prohibitively expensive using survey data. Finally, SSA administrative records provide high quality detailed monthly information on the receipt of SSI and DI benefits, demographics, and annual SSA covered earnings—variables that are critical for the net outcome evaluation of any employment-related demonstration intervention.

References

Abt Associates. 1997. *Project NetWork Baseline and Follow-up Survey Data and Documentation*. Report prepared for SSA/ORES in September 1997. Abt Associates: Bethesda, MD.

_____. 1998. *Evaluation of Project NetWork: Documentation of Final File of PNW Eligibles*. Report prepared for SSA/ORES on September 24, 1998. Abt Associates: Bethesda, MD.

Baruch, Robert F. 1997. *Randomized Experiments for Planning and Evaluation: A Practical Guide*. Applied Social Science Research Methods Series, Volume 44. Sage Publications: Thousand Oaks, CA.

Fu Associates. 1998a. *Development of the Project NetWork SSI Benefit Analysis File.* Prepared for SSA on January 12, 1998. Fu Associates, Ltd.: Arlington, VA.

_____. 1998b. *Development of the Project NetWork SSDI Benefit Analysis File*. Prepared for SSA on June 26, 1998. Fu Associates, Ltd.: Arlington, VA.

Leiter, Valerie; Michelle L. Wood; and Stephen Bell. 1997. "Case Management at Work for SSA Disability Beneficiaries: Process Results of the Project NetWork Return-to-Work Demonstration," *Social Security Bulletin,* Vol. 60, No. 1, pp. 29-48.

McManus, Leo; Kalman Rupp; and Stephen H. Bell. 1993. "Project NetWork: A Return-to-Work Project for Social Security Disability Applicants and Beneficiaries." In *Partners for Independence: Models that Work*. Conference Proceedings, The Fifth North American Regional Conference of Rehabilitation International, October 27-29, 1993, Atlanta, GA, pp. 427-435.

Pickett, Clark D. and Charles G. Scott. 1996. "Reinventing SSI Statistics: SSA's New Longitudinal File," *Social Security Bulletin*, Vol. 59, No. 2, pp. 31-56.

Rupp, Kalman; Stephen H. Bell; and Leo McManus. 1994. "Design of the Project NetWork Return-to-Work Experiment for Persons with Disabilities," *Social Security Bulletin*, Vol. 57, No. 2, pp. 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*, No. 7, pp. 63-91.

Notes

[1] The CMCS was an important source of information on type of service and program cost in the process analysis and in the analysis of program cost.

[2] The early mailing lists that were not available for the evaluation contractor electronically, covered the data for all mailings prior to June 1993. Site differences in the resulting truncation were substantial. The Dallas, TX and Forth Worth, TX "SSA Case Manager" model sites were most affected. The Richmond, VA "VR Outstationing" model site was least affected.

[3] It is to be noted that the Master Earnings File (MEF) played a role in creating the file of simulated eligibles, albeit an auxiliary one. Most notably, MEF information played an auxiliary role in narrowing the group of simulated eligibles to the 16-65 age group that has been identified by SSA's materials explaining the rules of eligibility (see Baruch (1997) for more detail). The almost negligible practical role of the MEF records in the eligibility simulation is simply the result of the fact that the MBR and SSR contain high-quality information concerning the key eligibility variables. The MEF, of course, is one of the major sources of the net outcome information, as will be discussed later in the article.

[4] For a detailed conceptual description of the contents of these "payment eligibility" fields and their relationship to actual payments, see Pickett and Scott (1996).

[5] For example, since date of birth has programmatic relevance in the OASDI program, we gave preference to the DI file version of this variable in case of conflict.

[6] The education variable was an exception with relatively high missing variable rates (table 2).