Development of Diagnostic Data in the 10-Percent Sample of Disabled SSI Recipients

by Satya Kochhar*

The Social Security Administration has created a 10-percent sample data base of blind and disabled recipients from the Supplemental Security Income (SSI) program. Codes showing the primary medical diagnosis were obtained for the sample by matching several files, and by imputing codes to sample cases where no diagnostic codes were found. The data base is updated each year by repeating this matching operation and by bringing forward the diagnostic codes from the previous year's file. This article describes the sources of diagnostic information in the administrative record system; the methodology used in the development of the 10-percent disability data base, and the technique chosen to compensate for missing values.

*Division of Program Management and Analysis, Office of Supplemental Security Income, Social Security Administration. The author would like to thank Barry Bye, Richard A. Bell, and Charles G. Scott for their expert guidance and Fred Tawney for his excellent computer system support.

The Supplemental Security Income (SSI) program provides monthly cash payments to blind and disabled persons with limited income and resources. An individual is considered disabled if he or she is unable to engage in any substantial gainful activity (SGA) because of any medically determinable physical or mental impairment. The impairment must be expected to result in death or have lasted or be expected to last for a continuous period of not less than 12 months. An individual is considered blind if he or she has either central visual acuity of 20/200 or less in the better eye with the use of correcting lens, or has tunnel vision of 20 degrees or less. In December 1990, about 3.4 million blind and disabled persons received payments under the SSI program. The average monthly payment was $337.

This article is one of a series of articles describing the statistical samples available at the Social Security Administration (SSA).[1] The article focuses on the methodology used to assign diagnostic codes where actual codes were not available from the existing sources. The analysis performed before the imputation provides information that may be of interest to SSI program analysts concerning the variation in diagnostic distribution with several key case characteristics. Having a sample file with imputed codes permits consistent presentation of program statistics in a variety of analytical contexts.

The information collected as part of the disability determination process includes the primary diagnosis that identifies the principal medical cause of the individual's disabling condition. Currently, no single SSA administrative record system contains complete information on diagnoses, demographic characteristics, and payment data for blind and disabled SSI recipients.

Before 1982, diagnostic information was not carried in the Supplemental Security Record (SSR), the major administrative file for the SSI program. Diagnostic codes and other medically related information obtained during the application and determination process were retained in the individual claims folders. Diagnostic information on all individuals receiving SSI payments at a given point in time therefore was not available. Since July 1982, diagnostic codes have been added to the SSR for newly awarded recipients.

To obtain diagnostic information for blind and disabled SSI recipients, SSA's current statistical process extracts a 10-percent sample of the SSR and matches it to other administrative records containing this information. This article describes the sources of diagnostic information in the administrative record system; the methodology used in the development of the 10-percent data base; and the technique chosen to compensate for missing values.

In 1986, SSA established a 10-percent data base. The 10-percent sample is based on the 8th and 9th digits of the Social Security number and is extracted twice a year—June and December—from the SSR. The sample is large enough to obtain reliable estimates for moderately small subsets of the population.

The file permits direct access to microdata and provides flexibility in compiling data for specific research needs. Each record included in the sample carries current information on demographic characteristics, payment data (including payments to representative payees), living arrangements, earned and unearned income, resources, and many other program related statistics on the SSI recipient population. It also allows the Office of Supplemental Security Income to provide State-specific data for a variety of purposes, including budget requests from State agencies, and requests for estimates of the impact of proposed legislation. These files are also widely used for in-house studies

[1] See page 21 for a list of available reprints from the Technical Article series.

about the types of disabilities of SSI recipients.

Beyond the quick access to cross-sectional and longitudinal information about SSI recipients, these sample files have an added use. Because the criteria for selecting cases is identical to the 1-percent Continuous Work History Sample (CWHS), the 1-percent and 10-percent Old-Age, Survivors, and Disability Insurance (OASDI) samples, and the Health Care Financing Administration's Continuous Medicare History Sample (CMHS), it permits record linkage for special studies.

Sources of Diagnostic Information

Obtaining diagnostic data for the 10-percent sample requires access to a number of administrative record systems at the Social Security Administration. The extent of the diagnostic information in these files is described below.

Supplemental Security Record

In July 1982, diagnostic codes for all new awards were added to the SSR except for two types of cases: those whose claim was denied initially but who subsequently were awarded payments after an appeal, and those who were awarded payments and had previously been entitled to Disability Insurance (DI) benefits. The semiannual 10-percent sample extract contains those diagnostic codes that exist on the SSR.

Master Beneficiary Record

The Master Beneficiary Record (MBR) is the primary administrative record file for the Old-Age, Survivors, and Disability Insurance program. The MBR contains information on all OASDI beneficiaries. Beginning in July 1982, diagnostic codes for all newly entitled beneficiaries were added to the MBR. It includes diagnostic codes not only for disabled-worker beneficiaries but also for disabled dependents who concurrently receive SSI payments. Approximately one-third of SSI recipients are entitled to title II (Social Security) benefits. Missing diagnostic codes in the SSR among these cases are likely to be found on the MBR.

Disability Determination Files

The SSA maintains annual disability determination files for the DI and SSI programs. These files are derived from the form SSA-831 (disability determination and transmittal form). For individuals applying for DI benefits, files have been maintained since 1955, and for those applying for SSI since 1975. These files contain diagnostic codes, and limited demographic data, but no payment data for SSI recipients.

SSI annual sample files.—These files contain diagnostic information on persons who apply for SSI or who concurrently apply for SSI and DI benefits on the basis of blindness or disability. Between 1975-83, they were based on a stratified sample of form SSA-831's. In the stratified sampling scheme, the size of the sample is determined by the size of the population of a respective geographic area. In this case, the size of the sample ranged from 2 percent in the more populous States to 100 percent for the less populous States. The SSI annual files exclude applicants previously entitled to DI benefits.

DI annual sample files.—These files contain diagnostic information on individuals applying for DI benefits and include codes for SSI recipients who had a previous entitlement to DI benefits. The files for 1955-83 include the SSI 10-percent sample.

In 1984, SSI and DI annual files were combined. The combined annual 831 files contain 100 percent of SSA disability determinations.

Matching Methodology

To develop complete information on the types of disabilities of SSI recipients, a 10-percent SSI disability data base was established in March 1986. The original file was created by extracting SSI blind and disabled recipient cases from the 10-percent sample file for all recipients and supplementing the file with diagnostic codes from other administrative files described above.

Initial March 1986 Match

The 10-percent sample file, created from the March 1986 SSR, contained approximately 490,380 records for persons whose application for SSI was based on blindness or disability. Of the sample applicants, 264,490 were in current-payment status. The remaining cases (225,890) represented claims with other statuses—denial, termination—or had a pending decision on program eligibility or payment amount. Of the 264,490 cases in current-payment status, 65,320, or 25 percent, had a diagnostic code on the SSR.

Matching operation.—The March 1986 10-percent sample file was matched with the MBR, the SSI annual files (1975-84), the DI annual files (1955-84), and the SSI and DI combined 1985 form 831 file. In addition to the diagnostic codes existing on the original March 10-percent sample file for 65,320 persons, the three-file match

produced 129,410 diagnostic codes for 80,540 additional sample persons. There are more codes (129,410) than sample persons (80,540) because of the multiple sources. The algorithm for selecting the best code is described below.

The codes are based on the diagnostic categories from the International Classification of Diseases (ICD). Generally, the ICD codes are 4-digit codes that identify morbidity and mortality conditions for statistical purpose.[2] From time to time the coding scheme is revised. Before January 1979, SSA had used the ICD 8th edition and after this date it adopted the ICD 9th edition. Table 1 shows the sources of the 129,410 diagnostic codes obtained through the match.

To eliminate multiple diagnostic codes and to convert the 8th edition codes to the 9th edition, a two-step operation was performed. Preference was given to the most recent 4-digit code in any of the three files. The diagnostic codes were selected in the following hierarchical order:

Record	Source edition
SSR (4-digit)	9th
MBR (4-digit)	9th
831 annual files:	
DI (4-digit)	9th
SSI (3-digit)	9th
SSI (3-digit)	8th
DI (3-digit)	8th

Of the 129,410 diagnostic codes obtained in the match, unduplicated codes were identified for 80,540 sample persons. Of that total, 43,240 codes were obtained from the 9th edition and the remaining 37,300 came from the 8th edition. To make the diagnostic information uniform, an algorithm was developed to translate the 37,300 codes from the 8th edition into codes for the 9th edition.

Table 1.—Number of diagnostic codes obtained from match operation, by source of code, March 1986

Record	International Classification of Diseases		
	Total	9th edition	8th edition
Total	129,410	90,550	38,860
Master Beneficiary Record	26,920	26,920	...
831 annual files:			
DI	29,660	18,880	10,780
SSI	72,830	44,750	28,080

Match results.—The number of cases with diagnostic codes in the March 1986 10-percent disability file more than doubled because of the administrative file record matches described above. As shown in table 2, the number of blind and disabled recipients who were in current-payment status, and who had diagnostic codes in their records, increased from 65,320 to 145,860 (123 percent).

Excluding Disabled Recipients Aged 65 or Older

The 145,860 records with a diagnostic code represented about 55 percent of the SSI blind and disabled population in March 1986. At this point, a decision was made to exclude blind and disabled recipients aged 65 or older from the analysis done for this article. These recipients are carried on SSA records as disabled although they meet the categorical eligibility requirements of aged recipients. For most analytical purposes, these older disabled recipients are included with aged recipients. Of the 211,774 recipients who were under age 65, the matched file contained 128,558 records with codes (or 61 percent of all blind and disabled recipients under age 65).

After the initial March 1986 match, new 10-percent sample files were constructed for 1987 and 1988, using the same matching techniques. This process is producing a steadily rising percentage of records with diagnostic codes because most of the new awards have codes.

There is, however, one complication. The diagnostic codes for new awards are not the pure ICD 9th edition codes. The SSA has developed a shortened set of roughly 180 codes for the most common impairments to simplify the coding process. The numeric values of the new codes have some relationship to the first 3-digits of the ICD code.[3]

In January 1986, SSA issued instructions that the disability determination services were to begin coding according to the new impairment codes.[4] Although

[2] **International Classification of Diseases**, 9th edition , Clinical Modification, Volume I, Diseases Tabular List, World Health Organization, Geneva, 1977.

[3] **Disability Evaluation Under Social Security** (SSA Publication No. 05-10089) Office of Disability, Social Security Administration, February 1986.

[4] Detailed information on the impairment codes can be found in the **Program Operations Manual System, POMS Part 4, Disability (DI)**, Office of Directives Management, Office of Policy, Social Security Administration, 26510.015.

impairment codes are easier to use, they lack the specificity of the ICD codes. Newly awarded cases with impairment codes are counted in the "with code" categories (table 3).

Imputation Methodology

Despite the matching operation employed for 3 consecutive years there remained a sizable group of cases for which no diagnostic code could be obtained. Using an imputation procedure, it is possible to assign a diagnostic code/category to cases without codes. In establishing a procedure for imputing codes to these cases, it was necessary to decide on the form of the imputed results and to identify a basis for assigning the imputed code to the records without codes.

The form of the imputed result could have simply been a 4-digit ICD code. However, as mentioned above, some records contained an impairment code for which the ICD code equivalent was not always identifiable. These dual coding schemes made it necessary to assign a diagnostic category, rather than a specific code. Both coding schemes could be translated into 13 major categories, used widely in publications to describe differing diagnoses. These 13 categories are shown in table 4. The actual code imputed was simply a 3-digit code representing the most frequently encountered code within each of the 13 categories.

As long as the cases with missing codes had a distribution similar to those with codes, there would be no bias in estimates for the total population based purely on the distribution of known cases. However, there are reasons to believe that the distributions differ.

Table 2.—Presence of diagnostic codes in March 1986 disability sample file from pre- and post-administrative record file matches

File status	Records for blind and disabled recipients in current-pay status			
	Pre-match		Post-match	
	Number	Percent	Number	Percent
Total records	264,490	100.0	264,490	100.0
With diagnostic codes	65,320	24.7	145,860	55.2
Without diagnostic codes	199,170	75.3	118,630	44.8

Table 3.—Presence of diagnostic codes in 1986, 1987, and 1988 disability data files from pre- and post-administrative record file matches

File status	Records for blind and disabled receipients under age 65 in current-pay status			
	Pre-match		Post-match	
	Number	Percent	Number	Percent
March 1986				
Total records	211,774	100.0	211,774	100.0
With codes	64,248	30.3	128,558	60.7
Without codes	147,526	69.7	83,216	39.3
December 1987				
Total records	232,872	100.0	232,872	100.0
With codes	93,318	40.1	156,631	67.3
Without codes	139,554	59.9	76,241	32.7
December 1988				
Total records	242,486	100.0	242,486	100.0
With codes	108,376	44.7	184,306	76.0
Without codes	134,110	55.3	58,180	24.0

In order to identify a basis for assigning imputed codes, it was necessary to determine how the mix of the 13 major categories differed from those with and without codes, and to identify the variables that were most useful in explaining the absence of a diagnostic code.

Variation of Diagnoses By Date of Application

Table 4 compares the distribution of the 108,376 blind and disabled recipients for whom diagnostic codes existed on the SSR with that of 75,930 recipients whose codes were obtained through matches. Certain diagnostic groups—musculoskeletal system, mental retardation, congenital anomalies, and nervous system—had proportions that were higher among those where codes were obtained through matches than those where codes existed on the SSR. The other groups such as neoplasms, infective and parasitic diseases, and psychiatric disorders, had proportions that were lower in the former distribution than that in the latter.

The main explanation for this phenomenon is that the match primarily picked up codes for pre-1982 cases obtained from the 831 files, and that these persons were likely to have diagnoses that permitted them to live longer. Before the matching operation, of the 108,376 persons with diagnostic codes on the SSR, 93 percent were for newly awarded recipients who came into the program beginning July 1982, and only 7 percent were for those who were recipients before that date (table 5). The matches picked up 75,930 additional diagnostic codes. Of these, about 62 percent were for pre-1982 recipients and about 38 percent were for the new cases.

One hyphothesis for the phenomenon shown in tables 4-5 is that among these long-term disabled recipients, fewer individuals diagnosed with diseases related to neoplasms, or respiratory, or parasitic and endocrine system would have survived. Similarly, in diagnostic groups where the mortality or the recovery rate is expected to be low, a larger number of persons still would be receiving SSI payments—for example, persons with disabling conditions such as mental retardation or congenital anomalies. However, persons with psychiatric disorders are expected to have low mortality rates, but the recovery rate among them is comparatively higher, especially among those not having psychotic or neurotic disorders.

The findings of a study for a 1972 cohort of DI program beneficiaries by SSA's Office of Disability support this hypothesis.[5] The death rate prior to age 65 among persons with neoplasms was projected to be 84 percent, compared with 24 percent of those with mental disorders. Persons with neoplasms remained in the program on the average of 3.4 years compared, with 15.6 years for persons with mental disorders.

From table 5, it is clear that most cases with missing codes were for persons who entered the program before 1982 and therefore, one can reasonably assume that the distribution of diagnoses of the 58,180 records with missing codes after the matches is closer to the distribution of the pre-1982 records (46,701) than it is to the post-1982 records (29,229). More than 88 percent of the records with missing codes had an application date prior to 1982.

The preceding analysis demonstrates that estimates of diagnostic distributions based only

Table 4.—Number and percentage distribution of sample cases with diagnostic codes that existed on the SSR and those obtained from matches, by diagnostic group, December 1988

Diagnostic group	Records for blind and disabled recipients under age 65		
	Existed on SSR	Obtained from matches	Combined
Total number	108,376	75,930	184,306
Total percent	100.0	100.0	100.0
Infective	1.2	1.0	1.2
Neoplasms	2.2	1.7	2.0
Endocrine	4.7	3.4	4.2
Mental disorders:			
Mentally retarded	25.0	31.3	27.6
Psychiatric	28.7	20.4	25.3
Nervous system	11.1	14.2	12.3
Circulatory	7.8	7.2	7.5
Respiratory	3.4	2.8	3.1
Digestive	.8	.9	.8
Musculoskeletal	7.2	7.9	7.5
Congenital anomalies	1.6	2.5	2.0
Injury and poisoning	3.0	4.2	3.5
Other	3.2	2.6	2.9

[5] John C. Hennessey and Janice M. Dykacz, "Projected Outcomes and Length of Time in Disability Program," **Social Security Bulletin**, September 1989, pp. 2-41.

on the known codes by simply using a straight reweighting of all cases with a valid code would result in a significant bias.

Random Imputation Approach

Kalton and Santos describe a procedure for reducing the bias by random imputation of missing data.[6] For statistical purposes, their approach assigns a value to sample cases in which the data are missing by using the value of sample cases where the data are present. As applied to this study, the total sample is partitioned into disjoint and exhaustive subsets called "imputation classes" defined by program and demographic variables. The variables that are used in forming imputation classes should be associated with the incidence of missing codes and with different diagnostic distributions. Random imputation is a procedure that assigns a value to records with missing data from records containing diagnoses and belonging to the same imputation class.

[6] Graham Kalton and Robert Santos, **Compensating for Missing Survey Data**, Survey Research Center, Institute for Social Research, University of Michigan, 1981.

Constructing Imputation Classes

To find potential variables for constructing imputation classes, a variety of program variables were tested by running a stepwise regression analysis. The presence of diagnosis (1, if present; 0, if absent) served as the dependent variable. Independent variables included age, race, sex, representative payee status, living arrangement, age at the time of application, type of DI benefit receipt, and length of time in the SSI program.

The variables that showed the greatest power to explain the presence or absence of a diagnostic code were chosen as candidates for use in imputing diagnosis to the records without a diagnosis. Association of these variables with diagnostic categories was confirmed by running a multivariate analysis of variance with diagnosis as a dependent variable.

Length of time in program.—As suggested by the analyses in the previous section, the shorter the period a recipient is in the SSI program, the more likely it is that the recipient's case record will have a diagnostic code. Of recipients who were receiving SSI payments for less than 5 years, almost 95 percent had diagnostic codes (table 6). This percentage changed as length of time in the program increased. For those who were SSI recipients for 5-9 years, it was 82 percent and for those who were on the rolls for 10 years or more, it was 44 percent. The reasons for the large number of missing codes among those recipients with longer program durations and the differences in diagnostic distributions was discussed above.

Table 7 shows the pattern of diagnoses by duration in the SSI program. Long-term recipients were more likely to be mentally retarded. Among recipients who received payments for 10 years or more, 39 percent were mentally retarded, in contrast with 24 percent of those who received SSI for less than 5 years. However, the pattern was reversed for those with a psychiatric illness. The longer a recipient stayed in the SSI program, the less likely the individual would have a psychiatric disorder. Less than 19 percent of those who remained in the program for 10 years or more had psychiatric disorders, compared with 28 percent of those who were in the program for less than 5 years.

Receipt of DI benefits.—Supplemental Security Income recipients can be divided into two program groups—SSI only and concurrent recipients (those who

Table 5.—Number and percentage distribution of sample cases for SSI recipients under age 65 with diagnostic codes, by match status and application date, December 1988

Source of diagnostic code	Date of application			
	Total number	Total percent	Before 1982	1982 and after
Total	242,486	100.0	43.6	56.4
Before match	108,376	100.0	7.0	93.0
Obtained through match	75,930	100.0	61.5	38.5
Without codes	58,180	100.0	88.5	11.5

receive SSI payments and DI benefits). The difference between the proportions of those with codes among the two major groups was not that great—73 percent for SSI-only cases, compared with 84 percent for concurrent cases (table 8). But when the concurrent cases were further distributed by type of DI benefit receipt, proportions were much more divergent for one subset of concurrent cases. Almost 93 percent of disabled-worker beneficiaries have a diagnostic code. Only 69 percent of disabled adult children (who may be of any age) and 77 percent of other disabled dependents (mainly, widows and widowers) have codes.

As shown in table 9, the proportion of mentally retarded (28 percent) recipients who receive only SSI was not that different from those recipients who also concurrently receive DI benefits (26 percent). However, when the concurrent group was further divided into subgroups, proportions of those with mental retardation differed remarkably. The incidence of mental retardation was highest among disabled adult children (56 percent). It was 15 percent for both disabled workers and other disabled dependents. Diseases of the musculoskeletal system and connective tissue were more prevalent among disabled workers

Table 6.—Number and percentage distribution of sample cases for SSI recipients under age 65 with diagnostic codes and duration in the program, December 1988

Time in program	Records for blind and disabled recipients under age 65			
	Total number	Total percent	With codes	Without codes
Total	246,486	100.0	76.0	24.0
Under 5 years	119,318	100.0	94.6	5.4
5-9 years	46,402	100.0	81.9	18.1
10 years or more	76,766	100.0	43.5	56.5

Table 7.—Number and percentage distribution of sample cases for SSI recipients under age 65 with diagnostic codes, by diagnostic group and duration in the program, December 1988

Diagnostic group	Time in program			
	Total	Under 5 years	5-10 years	Over 10 years
Total number	246,486	119,318	46,402	76,766
With diagnostic code				
Total number	184,306	112,905	38,004	33,397
Total percent	100.0	100.0	100.0	100.0
Infective	1.2	1.2	1.0	1.1
Neoplasms	2.0	2.4	1.7	1.3
Endocrine	4.2	4.8	3.8	2.5
Mental disorders:				
Mentally retarded	27.6	24.0	28.2	39.0
Psychiatric	25.3	28.2	22.4	18.6
Nervous system	12.3	11.0	14.4	14.5
Circulatory	7.5	8.1	7.3	5.9
Respiratory	3.1	3.5	3.0	2.1
Digestive	.8	.9	.7	.7
Musculoskeletal	7.5	7.9	7.3	6.3
Congenital anomalies	2.0	1.6	3.1	2.0
Injury and poisoning	3.5	3.3	3.7	4.1
Other	3.0	3.1	3.3	2.0

(12 percent) and other disabled dependents (15 percent) than among SSI-only cases (6 percent).

Age at time of application.—The age of an SSI recipient at the time of application affects both the availability of a diagnostic code and the diagnostic distribution among certain subsets of the SSI population. About 80 percent of those who entered the SSI program before reaching age 18 had a diagnostic code (table 10). These proportions were lower (70 percent) for those who entered the program at ages 25-44. However, for those aged 55-64 at the time of application, this proportion was 94 percent. Almost all of those who were in the group aged 55-64 became eligible after 1982.

Diagnostic distributions for the respective age groups differed. The younger the recipients were when they entered the program, the more likely they were to be mentally retarded (table 11). Almost one-half of those who entered the SSI program before reaching age 24 were eligible because of mental retardation. Among recipients who

Table 8.—Number and percentage distribution of sample cases for SSI recipients with diagnotic codes, by Social Security benefit receipt, December 1988

Type of Social Security benefits	Records for blind and disabled recipients under age 65			
	Total number	Total percent	With codes	Without codes
Total	242,486	100.0	76.0	24.0
SSI only	169,332	100.0	72.7	27.3
Concurrent cases	73,154	100.0	83.5	16.5
Disabled workers	43,007	100.0	92.8	7.2
Disabled adult children	24,112	100.0	68.7	31.3
Other disabled dependents	6,035	100.0	76.8	23.2

Table 9.—Number and percentage distribution of sample cases for blind and disabled SSI recipients under age 65, by diagnostic group and type of Social Security benefit receipt, December 1988

Diagnostic group	Total	SSI only	Concurrent cases			
			Total	Disabled workers	Disabled adult children	Other disabled dependents
Total number	242,486	169,332	73,154	43,007	24,112	6,035
With diagnostic code						
Total number	184,306	123,186	61,120	39,925	16,563	4,632
Total percent	100.0	100.0	100.0	100.0	100.0	100.0
Infective	1.2	1.1	1.2	1.4	.7	1.2
Neoplasms	2.0	2.1	1.9	2.1	1.0	2.6
Endocrine	4.2	4.2	4.1	4.6	1.2	10.3
Mental disorders:						
Mentally retarded	27.6	28.4	26.0	14.8	56.3	14.6
Psychiatric	25.3	25.0	25.7	31.7	13.8	16.6
Nervous system	12.3	13.2	10.7	9.7	13.6	8.5
Circulatory	7.5	6.9	8.9	10.8	1.9	17.9
Respiratory	3.1	3.0	3.5	4.1	.8	7.4
Digestive	.8	.8	.8	1.0	.2	1.0
Musculoskeletal	7.5	6.5	9.5	12.0	2.1	14.8
Congenital anomalies	2.0	2.5	1.0	.6	2.2	.4
Injury and poisoning	3.5	3.0	4.5	4.8	4.3	2.7
Other	3.0	3.3	2.2	2.4	1.7	2.0

entered at a later age, these proportions dropped significantly (21 percent, ages 25-44; 12 percent, ages 45-54; and 8 percent for the group aged 55-64).

However, this pattern did not hold for those who had psychiatric disorders, which were most prevalent (41 percent) among those who entered the SSI program at ages 25-44. Among those who entered the program before age 18, the proportion with psychiatric disorders was only 7 percent.

Among those who entered the program at ages 55-64, the most common reasons for the disability resulted from diseases of the circulatory system and the musculoskeletal system. These proportions were 22 percent and 20 percent, respectively.

Random Imputation of Missing Diagnostic Codes

"Imputation classes" were constructed by dividing the 10-percent sample (separately for both "with codes" and "without codes") into four classes defined by type of DI benefit receipt variable, by three classes of length of time in program, and five classes of age at the time of application variables—a total of 60 classes.

The assignment of values for sample records with missing codes was accomplished separately for each imputation class. Within each imputation class, records with codes were distributed into the 13 diagnostic groups. Then the assignment of the imputed code to the record with a missing code was performed in such a way that the probability of selecting a code belonging to a diagnostic group equaled the proportion of codes in that group among all records in the imputation class with codes. A random number was generated to assign the imputed codes permanently to each record in each of the imputation classes. The

Table 10.—Number and percentage distribution of sample cases for SSI recipients with diagnostic codes, by age at time of application, December 1988

Age at time of application	Records for blind and disabled recipients under age 65			
	Total number	Total percent	With codes	Without codes
Total	242,486	100.0	76.0	24.0
Under 18	45,029	100.0	79.7	20.3
18-24	47,634	100.0	73.4	26.6
25-44	79,513	100.0	70.5	29.5
45-54	45,646	100.0	75.0	25.0
55-64	24,664	100.0	93.7	6.3

Table 11.—Number and percentage distribution of sample cases for SSI recipients with diagnostic codes, by diagnostic group and age at time of application, December 1988

Diagnostic group	Records for blind and disabled under age 65					
	Total	Under 18	18- 24	25-44	45-54	55-64
Total number	242,486	45,029	47,634	79,513	45,646	24,664
With diagnostic code						
Total number	184,306	35,897	34,968	56,071	32,249	23,121
Total percent	100.0	100.0	100.0	100.0	100.0	100.0
Infective	1.2	.5	.7	1.8	1.4	.9
Neoplasms	2.0	2.0	1.1	1.6	2.9	3.2
Endocrine	4.2	1.8	1.3	4.3	7.2	7.4
Mental disorders:						
Mentally retarded	27.6	49.6	44.4	20.8	12.0	7.8
Psychiatric	25.3	6.8	24.0	41.1	25.7	16.7
Nervous system	12.3	21.0	15.8	9.7	7.7	7.0
Circulatory	7.5	1.0	1.5	4.9	15.0	22.2
Respiratory	3.1	1.4	.5	1.7	6.2	8.8
Digestive	.8	.3	.3	1.0	1.4	1.0
Musculoskeletal	7.5	1.5	2.0	5.7	13.9	19.7
Congenital anomalies	2.0	6.8	1.8	.6	.5	.3
Injury and poisoning	3.5	1.5	4.4	4.3	3.7	3.1
Other	3.0	5.8	2.2	2.5	2.3	1.8

assigned codes were flagged so they could be dropped in the event a diagnostic code becomes available from an administrative record.

Table 12 shows the distribution of sample cases with diagnostic codes that existed on the SSR (pre-match), codes obtained through the matching operation, imputed codes, and final accumulated codes.

The distribution of imputed diagnostic codes for 58,180 sample records with missing codes was closer to that of those records where codes were obtained through the matching operation than to the pre-match distribution (table 9). In both categories, proportions of the sample persons with diseases with low mortality and low recovery rates were higher than others. For example, 32 percent were mentally retarded in the imputed category, compared with 25 percent in the pre-match category. The latter group included more recent cases. However, the opposite was true for those with psychiatric disorders. In the imputed category, only 22 percent had such diagnosis, compared with 29 percent in the pre-match category.

Discussion

Over the past several years, SSA has been successful in creating a sample file containing diagnoses for a cross-section of SSI recipients. The diagnoses were obtained from several administrative computer files, and by imputation, where codes were not available. This 10-percent disability data base will be updated annually. It will be used for special studies, to estimate the impact of proposed legislation, and for ad hoc requests from various sources.

Because the sample file will be used for a variety of statistical analyses, it was decided that the imputation process would be appropriate for compensating for missing codes. This process has several advantages over other missing data approaches. As the analysis above demonstrates, it is not appropriate to use only those cases with known diagnoses to describe the SSI blind and disabled population. Such an approach would lead, for example, to an undercount of about 27,000 recipients who are mentally retarded and an overcount of about 20,000 recipients with psychiatric disorders. For subgroups of the SSI population, the impact could be even more significant.

The missing data problem could be addressed by constructing a special set of case weights for those cases with diagnoses available to adjust for the difference between those cases with and without codes. The case weights could have been based on the same imputation classes as those used for the random assignment described above. That is, within each class, case weights could have been assigned to cases with known diagnoses in each class so that the sum of the weights equaled the total number of cases in the class.

Although this approach would have resulted in a weighted

Table 12.—Number and percentage distribution of sample cases for SSI recipients by diagnostic codes that existed on SSR, obtained through matching operation, inputed codes, and final codes

Diagnostic group	Records for blind and disabled recipients under age 65			
	Pre-match codes	Obtained from match	Imputed codes	Final codes
Total number	108,376	75,930	58,180	242,486
Total percent	100.0	100.0	100.0	100.0
Infective	1.2	1.0	1.2	1.2
Neoplasms	2.2	1.7	1.7	2.0
Endocrine	4.7	3.4	3.3	4.0
Mental disorders:				
Mentally retarded	25.0	31.3	32.1	28.7
Psychiatric	28.7	20.4	22.0	24.5
Nervous system	11.1	14.2	13.7	12.7
Circulatory	7.8	7.2	7.0	7.4
Respiratory	3.4	2.8	2.6	3.0
Digestive	.8	.9	.9	.8
Musculoskeletal	7.2	7.9	7.1	7.4
Congenital anomalies	1.6	2.5	1.9	2.0
Injury and poisoning	3.0	4.2	2.4	3.6
Other	3.2	2.6	4.0	2.8

diagnostic distribution with the same structure as the imputed distribution, this procedure has two principal drawbacks when it is to be used in conjunction with a general purpose data base. First, use of case weights adds complexity to the estimation of population totals and their sampling errors. Second, those analyses that do not focus on diagnosis would use all of the 10-percent sample cases, not just those with known diagnoses. These analyses would result in different estimated totals for subsets of the SSI population than those produced from the weighted data. Although both sets of estimated totals would have reasonably good statistical properties, the fact that they would differ from one analysis to the next could be a source of some confusion. It is better that all analyses begin with the same set of sample cases whether or not the focus of the analysis includes the medical diagnosis.

Additional work may be done on the imputation process. Because the imputed codes that were assigned were limited to codes that were representative of the 13 diagnostic groups, it is still not possible to obtain diagnostic breakdowns in any format other than those 13 categories. Approaches are now being considered that would permit imputation of a more specific set of diagnostic codes. Because analytical needs sometimes require a greater degree of specificity, it will be useful to conduct a thorough review of relationships between the ICD codes and the impairment codes, so that more than 13 diagnostic categories can be identified.

Available Reprints from the Technical Article Series

The Decline in Establishment Reporting: Impact on CWHS Industrial and Geographic Data

This article discusses the impact of the decline of employers' participation in the voluntary Establishment Reporting Plan (ERP) program on the industrial and geographic data in the Continuous Work History Sample (CWHS), SSA's largest file of employment and earnings records. Under the ERP, employers with multiunit businesses supply information that allows each unit to be classified under a separate primary geographic and industrial activity. In the mid-1970's, participation in the ERP began to decline, and the decline has accelerated since 1978 when SSA changed to annual wage reporting. The CWHS area most affected involves geographic data; the impact on industrial classification has been less severe and primarily involves employers in the manufacturing sector.—Linda M. Dill, Adah D. Enis, and Cheryl I. Williams, January 1991, pp. 2-20.

The Monthly OASDI One-Percent Sample File

This article briefly describes the development of SSA's Master Beneficiary Record (MBR) and documents the contents, technical features, and uses of the 1-percent sample file that is developed from it. This sample file is designed to enchance the production of descriptive statistics, simulations of the beneficiary population, and other statistical analysis and research projects. A 1-percent sample file has been generated for each month since January 1985.—Lewis F. Frain, June 1989, pp. 8-15.

The Social Security Administration's Continuous Work History Sample

The Continuous Work History Sample (CWHS) is the oldest major longitudinal sample data source in the Federal statistical system. It was developed to fulfill the need for statistics to be used in planning and operating the old-age insurance system established by the Social Security Act of 1935. This article discusses the current structure of the CWHS data system and files and describes some uses of the sample data and a number of issues that will affect the future of the system. — Creston M. Smith, October 1989, pp. 20-28.

Requests for copies of **Social Security Bulletin** Technical Articles should be addressed to the Office of Research and Statistics, Room 209, Van Ness Centre, 4301 Connecticut Ave., NW., Washington, DC 20008 or telephone (202) 282-7138.