How Large Are the Classification Errors in the Social Security Disability Award Process?

Hugo Benítez-Silva, *SUNY-Stony Brook*
Moshe Buchinsky, *UCLA, NBER, and CREST-INSEE*
and
John Rust *University of Maryland and NBER*[*]

February, 2003

Abstract

This paper presents an "audit" of the multistage application and appeal process that the U.S. Social Security Administration (SSA) uses to determine eligibility for disability benefits of the Disability Insurance (DI) and Supplemental Security Income (SSI) programs. We study a subset of individuals from the Health and Retirement Survey (HRS) who applied for DI or SSI benefits between 1992 and 1996. We compare the SSA's ultimate award decision $\tilde{a}$ (i.e., after allowing for all possible appeals) to the applicant's self-reported disability status $\tilde{d}$ (recorded at the first HRS survey *after* their initial application for benefits). We use these data to estimate classification error rates under the hypothesis that applicants' self-reported disability status $\tilde{d}$ is the relevant measure of "true disability" and the SSA's ultimate award decision $\tilde{a}$ is a noisy but unbiased indicator of $\tilde{d}$. This "truthful, accurate reporting hypothesis" allows us to estimate the magnitude of classification errors in the SSA award process and obtain insights into the patterns of self-selection induced by varying delays and award probabilities at various levels of the application and appeal process. Overall we find that 28% of SSI/DI applicants who are ultimately awarded benefits are not disabled, and that 61% of applicants who were denied benefits are disabled. We construct a "computerized" disability screening rule using a subset of "objective" health indicators that the SSA uses in making award decisions that results in significantly lower classification error rates than does SSA's current award process. This suggests that there may be cheaper, faster, and more accurate ways to make disability determinations than the SSA's current disability award process. We also estimate classification errors under the assumption that both $\tilde{a}$ and $\tilde{d}$ are noisy but unbiased indicators of an (unobserved) underlying indicator of "true disability," $\tilde{\tau}$. However, the estimated classification error rates remain virtually unchanged under this alternative hypothesis.

Keywords: Social Security, disability, Health and Retirement Study, classification errors.
JEL classification: H5

[*]Corresponding author: please send comments or questions by email to jrust@gemini.econ.umd.edu. This work was supported by NIH grant AG12985-02. We have benefited from feedback from participants of the Cowles Foundation Seminar, the Conference on Reforming Social Security Organized by the Fundación BBV in Madrid, the NBER Summer Institute on Aging, the UCLA applied micro seminar, and the Maryland Population Research Center seminar series. We are grateful for research assistance by Paul Mishkin, and from Hiu-Man Chan and Sofia Cheidvasser for helping prepare the data in the early stages of this project. We also thank the staff of the University of Michigan Survey Research Center (SRC) for answering numerous questions about the HRS data.

1 Introduction

This paper provides an "audit" of the multistage application and appeal process that the U.S. Social Security Administration (SSA) uses to determine eligibility for disability benefits under the Disability Insurance (DI) and Supplemental Security Income (SSI) programs.[1] We seek to quantify the magnitude of "classification errors" in the award process, that is, what fraction of applicants who are awarded benefits are not really disabled, and what fraction of applicants who are denied benefits are really disabled? This is a difficult task since it would appear to require an objective definition of "true disability" as well as an independent, verifiable procedure for reviewing SSA's award decisions to determine which applicants are truly disabled.

Ever since the inception of the DI program in 1956 and the subsequent introduction of SSI in 1976, SSA has made disability determinations according to the same basic definition of "disability", namely

> *The inability to engage in any substantial gainful activity (SGA) by reason of any medically determinable physical or mental impairment, which can be expected to result in death, or which has lasted, or can be expected to last, for a continuous period of at least 12 months.*[2]

While this appears to be a reasonably "objective" definition of disability, in practice it is very difficult to determine whether or not a particular individual is capable of substantial gainful activity. For example, there are hundreds of objectively verifiable medical conditions, cataloged in SSA's "Blue Book", that it regards as sufficiently severe to automatically qualify an applicant for benefits without further consideration of their "residual functional capacity," or the possibility of accommodations that could enable the person to continue working in their current job or some other less demanding jobs. Examples of these "listing conditions" include blindness, multiple sclerosis, and AMS. However, it is easy to cite examples of people who suffer from these conditions who can, and do, work.[3] Thus, even the most obvious objectively verifiable disabling conditions do not seem to admit any objective, error-free procedure that determines whether specific individuals suffering from these conditions are capable of working. There appears to be intangible, difficult to measure characteristics such as intelligence, motivation, and determination that enable certain people to work in spite of severe handicaps.

[1] Although the DI and SSI programs serve different target populations both programs use the same disability determination process and the same underlying definition of disability (presented below). SSI is a means-tested social assistance program that pays a flat benefit, whereas DI is an insurance program for workers that pays benefits related to average earnings.

[2] DI recipients can work without loss of benefits provided that their earnings do not exceed a limit known as the "SGA threshold." During the period of our study (1992-1996) the SGA threshold was $500 per month. This amount was increased to $800 on January 1, 2003, and will increase in the future to keep up with the national average wage index.

[3] Examples include economists Roberto Serrano, Walter Oi, and Curtis Taylor (who are blind) and physicist Stephen Hawking (who has AMS).

In addition to the inherent difficulties in making disability determinations on a case by case basis, an analysis of time series and state level data on SSA disability award rates makes it hard to escape the conclusion that the implementation of its definition of "disability" is subject to political and social influences that can cause disability award rates to vary widely over time, and across states at a given point in time. For example, aggregate disability award rates (the fraction of awards in a year divided by the number of applications in that year) have ranged from a low of 29% in 1982 (during the Reagan Administration) to a high of 52% in 1998 (under the Clinton Administration). It is unlikely that these wide swings in acceptance rates are due to changes in the characteristics of the applicant pool. In the next section we discuss the initial disability determinations, made by state-level bureaucracies known as "Disability Determination Service" (DDS) centers. The DDS award rates also vary widely across states at a given point in time, again in a manner that is difficult to ascribe entirely to differences in the characteristics of the applicant pool. For example, DDS award rates for DI applicants in 2000 ranged from a high of 65% in New Hampshire to a low of 31% in Texas, and award rates for SSI applicants ranged from a high of 59% in New Hampshire to a low of 27% in West Virginia.[4]

The variability in the implementation of the SSA's basic definition of disability over time and across states creates additional difficulties for our evaluation of the accuracy of the award process. We do not want our analysis to be clouded by subjective political judgments about whether the SSA's standards for awarding disability benefits are "too lenient" or "too tough" at a particular point in time. Fortunately, it is possible to separate the question of the *accuracy* of the disability award process from the question of its *leniency* under whatever socio-political "regime" is in place at a particular point in time.

Our approach for estimating the classification errors in the SSA disability award process is simple: *we compare the SSA's award decisions to the self-reported disability status of a sample of applicants who, we believe, have provided truthful and accurate reports of their disability status.* We study a subset of individuals from the Health and Retirement Survey (HRS) who applied for DI or SSI benefits between 1992 and 1996. We compare the SSA's ultimate award decision $\tilde{a}$ (i.e., after allowing for the possibility of appeal) to the individual's self-reported disability status $\tilde{d}$ (recorded at the first HRS interview *after* their initial application for benefits). The latter is a binary indicator that is set to 1 if the HRS respondent reports that they have an "impairment or health condition that prevents them from working entirely", and

[4] The effect of political influences on disability determinations is clearly evident in the SSA's treatment of alcoholism and drug addiction. Prior to 1996 these conditions were considered as valid "impairments" that could enable applicants to qualify for DI or SSI benefits. However, in 1996, the law was amended to specifically exclude drug and alcohol addiction as disabling conditions, leading to a sudden one time surge in "recovery" rates. This policy change was likely due in part to political pressure arising from press exposés of disability beneficiaries who were drug addicts and who admitted to using their disability benefits to pay rent and engaging in larceny to support their drug habit. The policy change may also have been prompted by the welfare reform movement and the prevalent attitude that "able bodied people ought to work".

0 otherwise. This is essentially the same as the SSA's definition of disability as an "inability to engage in substantial gainful activity". Although there are semantic differences between the SSA's definition of "disability" and the definition implicit in our use of the self-reported disability question in the HRS interviews, we believe that the definitional differences are of second order relative to the potentially more serious concern that DI or SSI applicants have an incentive to misreport (i.e., exaggerate) the severity of their impairments and to claim that they are incapable of working even when they really can work.[5]

We estimate classification error rates as follows. In addition to an individual's self-reported disability status $\tilde{d}$ and SSA's ultimate award decision $\tilde{a}$, assume there exists a third, latent indicator of "true disability status" $\tilde{\tau}$. Consider first the case where $\tilde{d} = \tilde{\tau}$ (i.e., where we treat $\tilde{d}$ as representing our measure of "true disability"). Using observations on the SSA's ultimate award decision $\tilde{a}$ and the self-reported disability $\tilde{d}$ for our sample of applicants from the HRS, we can estimate the joint distribution, $\Pr\{\tilde{a}, \tilde{d}\}$. The estimated classification error rates can be computed as conditional probability statements using this joint distribution. There are two types of classification errors *award errors* and *rejection errors.* The former is the conditional probability that a person who has been awarded benefits is not truly disabled, i.e., $\Pr\{\tilde{d} = 0|\tilde{a} = 1\}$. The latter error is the conditional probability that an applicant who was denied benefits is truly disabled, i.e., $\Pr\{\tilde{d} = 1|\tilde{a} = 0\}$. We estimate the award error to be 28% and the rejection error to be 61%.[6]

It is possible to estimate the award and rejection error rates without assuming that $\tilde{d} = \tilde{\tau}$, in the more realistic case where $\tilde{a}$ and $\tilde{d}$ are both noisy indicators of the latent indicator of true disability, $\tilde{\tau}$. Based on previous empirical work (to be described in the next section), we argue that both $\tilde{a}$ and $\tilde{d}$ are *unbiased indicators* of true disability status $\tilde{\tau}$. Under the additional assumption that the three binary random

[5] The HRS definition of disability as a "health condition that prevents a person from working entirely" is stricter in some respects than the SSA's definition of disability. Under the SSA's definition a person could be considered disabled even if they were still able to work, provided that their monthly earnings did not exceed the SGA threshold (currently $800 per month). However, there are other respects in which the HRS definition of disability is less strict than the SSA definition. A respondent in the HRS may report that they are unable to work entirely, but due to a temporary health problem that is not expected to last continuously for 12 months or end in death. These individuals would not be eligible for benefits according the SSA's definition. Also it is not clear whether an HRS respondent who reports that their health condition prevents them from working entirely is referring to their *current job* or *any job.* Under the SSA's definition a person is disabled only if they are unable to engage in substantial gainful activity in *any* type of job, even if this might involve retraining or relocation. A final source of differences between $\tilde{a}$ and $\tilde{d}$ may be due to differences in timing between when the award decision was ultimately made and when the individual was interviewed by the HRS. Here we use self-reported disability from the first wave of the HRS *after* the person reported applying for disability benefits, which was on average 6 months after the initial application. It is possible that a person might have recovered from their disability in the time between their application and their interview by the HRS. While the timing difference and the differences in definition can be a source of some discrepancies between $\tilde{a}$ and $\tilde{d}$, they have only "second-order" effects and cannot be responsible for the very large discrepancies that we find in this analysis. The rest of the paper provide evidence that should convince the reader that the main explanation for the large discrepancies between $\tilde{a}$ and $\tilde{d}$ is that SSA has limited information about an applicant's true health status, and this limited information leads to high rates of errors in its award decisions.

[6] The award and rejection errors are different from the usual Type I and Type II errors of hypothesis testing. A Type I error rate is the probability of rejecting H_0, that an applicant is "truly disabled", when it is true, i.e., $\Pr\{\tilde{a} = 0|\tilde{d} = 1\}$, while Type II error rate is the probability of not rejecting H_0 when it is false, i.e., $\Pr\{\tilde{a} = 1|\tilde{d} = 0\}$. Our point estimates of the Type I and II error rates of the SSA award process are of similar magnitude as the award and rejection error rates, i.e. 26% and 63%, respectively.

variables $\tilde{\tau}$, $\tilde{d}$, and $\tilde{a}$ form a trivariate probit system with a correlation structure designed to match the correlation between the observed random variables $\tilde{a}$ and $\tilde{d}$, we can derive formulas for the classification error probabilities by a straightforward application of Bayes Rule (see section 5 for details). Our estimated classification error rates are hardly changed in this case: the award error rate falls to 23%, but the rejection error rate remains at 61%. The Type I and II error rates do not change much either, 23% and 68%, respectively, compared to 26% and 63% in the case where we assume that $\tilde{d} = \tilde{\tau}$ with probability 1.

Our estimated classification error rates are higher than those estimated in previous internal and external audits of the SSA's disability award process. These previous studies relied on decisions of independent "experts" who attempted to directly measure true disability status $\tilde{\tau}$. For example, Smith and Lilienfeld (1971) reported the results of an internal audit of DI awards done by the SSA's Bureau of Disability Insurance (BDI). The BDI found that 21.2% of DI awards should have been denied and 22.5% of DI denials should have been awarded. In a seminal study, Nagi (1969) provided an independent external audit of a sample of 2,454 DI cases. Teams of five experts (consisting of a physician, psychologist, social worker, occupational therapist, and a vocational counselor) conducted individual home examinations/interviews for their sample of DI applicants. With the assistance of a moderator, the team arrived at a collective decision about the disability status of the applicant, without knowledge of the SSA's actual award decision. The results, summarized in Table 1 below, are qualitatively similar to our findings. In particular, the implied rejection error rate, 48%, was more than twice as large as the award error rate, 19%. The team of experts was slightly more lenient than SSA, concluding that 68% of applicants were disabled compared to the SSA's award rate of 62%. However, this difference is not large enough to explain the surprising number of classification errors. Overall, the expert team's decisions differed from SSA's award decision in over 30% of the cases considered.

Table 1: Summary of DI Classification Errors from Nagi (1969)

Expert	SSA Award Decision		Total
Team Decision	Awarded	Denied	
Can Work	291	492	783
	(19.3%)	(52.1%)	(31.9%)
Cannot Work	1,219	452	1,671
	(80.7%)	(47.9%)	(68.1%)
Total	1,510	944	2,454
	(61.5%)	(38.5%)	(100.0%)

Unfortunately, to our knowledge, there are no recent studies undertaking a similar assessment of the classification errors in the current DI award process. The rapid growth during the mid-1990s in the number of initial DI denials overturned on appeal could be an indication of a rise in the classification error rates over this period. One likely reason for this is the unprecedented and unsustainable growth rate (over 10% per year) of awards during the early part of the 1990s, overwhelming the processing capacity of the DDSs. The increase in applications also led to substantial growth in the number of appeals to the SSA's Administrative Law Judges (ALJs). The total number of appeals grew from 225,000 in 1986 to about 498,000 in 1996 (U.S. GAO 1997), increasing processing delays and creating a backlog of nearly a half million cases. The GAO study reports that the average processing time for appealed cases rose from about 10 months in 1994 to over one year in 1996. These huge backlogs have naturally led to concern about the quality of evaluations, especially in view of the high "reversal rate" (of initial denials by the DDS centers) by ALJs, as is clear from the summary provided in Table 2. Most importantly, note that this pattern is apparent across all impairment types. The overall award rate of the ALJs, 77%, is more than twice as large as the 30% initial award rate at the DDS level.

Table 2: Summary of DDS and ALJ Award Rates by Impairment Type

Condition	DDS Award Rate	ALJ Award Rate
Physical	29%	74%
Musculoskeletal	16	75
Back cases	11	75
Other	23	76
Other physical	36	74
Mental	42	87
Illness	39	87
Retardation	54	84
All impairments	30	77

The GAO report specifically cites incomplete documentation of the DDS centers as one of the main reasons for denial of benefits, and one of the major factors behind the high reversal rate by the ALJs.[7] However, the report also suggested that some reversals may be due to the limited medical expertise of the ALJs, who are judges not doctors, and who consult independent medical experts in only 8% of cases leading to awards. In 1997, 27% of all awards were due to successful appeals to ALJs. The GAO report documents major inconsistencies between initial disability determinations by the DDS and the ultimate

[7] Specifically, the GAO states that: "In a 1994 study, SSA found that written explanations of critical issues at the DDS level were inadequate in about half of the appeal cases that turned on complex issues. Without a clear explanation of the DDS decision, the ALJ could neither effectively consider it, nor give it much weight" (U.S. GAO 1997, p. 8).

decisions by the ALJs. Nevertheless, this report provides no explicit information on whether the appeal option increases or reduces the award and rejection errors.

These problems motivated the SSA to propose a comprehensive "disability process redesign" plan in 1994 in order to simplify and streamline the sequential evaluation process used by the DDS and to improve the documentation of their reasons for denials. As part of the redesign process, the SSA considered statistical approaches to disability evaluation using functional impairment indices, based on standardized measures of health and functional status. These would be designed to measure, as objectively as possible, an individual's ability to perform a baseline of occupational demands, including principal dimensions of work and task performance such as primary physical, psychological, and cognitive processes. The goal was to provide a more consistent, unified, and objective basis for initial award decisions. Other changes the SSA considered included: collapsing the current five-stage DDS disability evaluation process into two stages, and the use of a single "disability claim manager" who would be responsible for all aspects of a given claim. Under the current system, anywhere from 16 to 26 different DDS employees handle different parts of a single DI application. Although the main objective of the redesign initiative was to reduce delays in making disability determinations, the initiative may have also been motivated by a desire to obtain a more uniform application of the standards for judging whether or not an applicant is disabled.

Our analysis sheds light on these issues. In particular, our results suggest — contrary to the suggestions of the GAO report — that the high reversal rates by the ALJs actually serves to *reduce* classification error rates. We find that the low initial award rate at the DDS level produces a high rate of rejection errors at this stage. The DDS centers appear to behave according to a philosophy of "when in doubt, reject." However, self-selection is operative: we find that applicants who appeal an initial rejection by the DDS are more likely to be truly disabled than the initial pools of applicants that the DDS evaluated. Therefore, the relatively high acceptance rate by the ALJs, combined with the self-selection in the decision to appeal an initial rejection, significantly reduces the rate of rejection errors without increasing the rate of award errors.

We also consider our version of a "statistical approach" to making disability determinations. We construct an econometrically estimated "computerized screening rule" that results in a significantly lower level of classification errors than the SSA's current disability award process. This suggests that there may be cheaper, faster, and more accurate ways of making disability determinations than the SSA's current disability award process. The use of a statistical screening rule (at least for initial determinations, with human judges still handling appeals) could be the most promising part of the SSA's disability redesign initiative. Unfortunately, the new SSA administration seems to have abandoned the redesign initiative.

The remainder of the paper is organized as follows. Section 2 discusses previous empirical work justifying the hypothesis that HRS respondents who applied for DI provided truthful, unbiased, and "accurate" reports of their disability status. Section 3 provides a brief description of the DI and SSI programs and the disability award process. Section 4 describes the HRS data used in this analysis and provides the results of tabulations demonstrating that self-reported health status, $\tilde{d}$, constitutes an approximate "sufficient statistic" that better predicts variation in a list of "objective" health indicators than does the SSA's ultimate award decision $\tilde{a}$. Section 5 presents the results of an analysis of the award and rejection errors at each stage of the disability award process, while Section 6 describes the computerized disability screening procedure. Section 7 offers some conclusions and qualifications of our research findings, and suggestions for further research.

2 Do People Truthfully Report Their Disability Status?

There is a large academic literature that questions the validity of self-reported disability as a measure of "true disability" due to the presumed incentive to misrepresent one's health status in order to "rationalize" non-participation in the labor force (for survey respondents), or to increase the odds of being awarded benefits (for DI or SSI applicants). While we agree that SSI and DI applicants have an incentive to misrepresent their health status to the SSA, we believe that *due to strong guarantees of confidentiality, HRS survey respondents had no incentive to misrepresent their health or disability status, and provided truthful answers to these questions.* The HRS survey was conducted by the University of Michigan Survey Research Center (SRC) and not by SSA, and respondents were given strong assurances that their identities would not be revealed.[8]

One piece of evidence supporting our *truthful reporting hypothesis* is the fact that among HRS respondents who reported receiving DI and SSI benefits in the income section of the HRS survey, 18% of these also reported that their health condition did *not* prevent them from working entirely in the disability section of the survey. It is hard to reconcile these responses with the prevailing view that respondents exaggerate health problems in order to rationalize labor force non-participation or receipt of disability benefits. It seems more likely that the DI and SSI recipients felt no stigma in admitting that their health problem did not prevent them from working, and that they believed the HRS's guarantees of anonymity and confiden-

[8] In the second wave of the HRS respondents were asked for permission to link their survey responses to specific types of administrative data held by the SSA, including earnings histories from SSA and a limited amount of beneficiary data. However, they were given legally binding guarantees that the linkage would occur only for these data items and that SSA would not retain any information that would make it possible to link their survey ID to their Social Security number or take any action that would in any way jeopardize their Social Security benefits. Based on these strong guarantees, over 9,000 of the original 12,652 HRS respondents contacted by the HRS in wave 1 agreed to allow these administrative data linkages to be made.

tiality were credible. Otherwise it would not make sense for a DI or SSI recipient to make such admissions, especially if they believed their responses could be detected by the SSA (based on a linkage of their survey ID to their social security number), since these reports would presumably be grounds for the SSA to order audits known as "continuing disability reviews" (CDRs) to remove them from the roles.[9]

The validity of our analysis also depends on an additional assumption, which we refer to as the *accurate reporting hypothesis.* Even if individuals truthfully report whether they are capable of working or not, they may be using a different standard or "threshold" of disability than the SSA. For example, individuals may, on average, have an internal standard for judging disability that is too "lenient" in comparison with the standard that the SSA employs. In previous work (Benítez-Silva, Buchinsky, Chan, Rust, and Sheidvasser (BBCRS) 2003), we have shown that $\tilde{a}$ and $\tilde{d}$ are unbiased indicators of each other, i.e., we were unable to reject the hypothesis that $E\{\tilde{a} - \tilde{d}|x\} = 0$, where x is a vector of "objective indicators" of health problems and activities of daily living (ADLs).[10]

We interpret this "unbiased reporting" result, $E\{\tilde{d} - \tilde{a}|x\} = 0$, as providing strong empirical support for the "accurate reporting hypothesis". If the HRS respondents had a more lenient disability threshold than the SSA, we would expect their self-reported disability status to be upward-biased relative to the SSA, i.e., we would expect to find that $E\{\tilde{d} - \tilde{a}|x\} > 0$.

The BBCRS study was also unable to reject a more specific hypothesis about how applicants' self- reported disability status relates to SSA's ultimate award decision, the *rational unbiased reporting hypothesis* (RUR). The SSA's award decision can be approximated by a *threshold rule* of the form $\tilde{a} = I\{x\beta_a + \varepsilon_a > 0\}$, where I denotes the usual indicator function, x is a vector of observable (and verifiable) characteristics of the applicant including indicators of health conditions, ADLs, etc. and ε_a is a random variable representing the effect of "bureaucratic noise" and other unobserved factors that affect the SSA's award decision. Thus, the SSA awards the applicant if $x\beta_a + \varepsilon_a > 0$ and denies the applicant if $x\beta_a + \varepsilon_a \leq 0$. The parameters of the vector β_a represent the relative weights (or importance) of various health conditions in the SSA's

[9] Although it is possible that some of these 18% had experienced a medical recovery, in fact fewer than 1% of all DI beneficiaries ever leave the roles voluntarily as a result of medical recoveries despite strong incentives such as a 9 month "trial work period" during which beneficiaries are allowed to work without fear of being removed from the roles or losing any benefits (Muller 1992). It is a puzzle why so few DI recipients voluntarily return to work if a significant fraction of them have either recovered or were never truly disabled in the first place. However, we show in Section 3 below that most DI recipients are very poor with much lower educational attainment compared to non-recipients. The after-tax wages that recipients could expect to earn from returning to work may not be substantially higher than their DI or SSI benefits, and they would also lose their access to Medicare or Medicaid coverage if they were to leave the roles. For this reason there appears to be a clear incentive for individuals with low wage prospects to remain on the roles even if they are capable of earning in excess of the SSA's SGA threshold.

[10] Our previous study also provides strong evidence that our inability to reject this hypothesis is not due to limited numbers of observations or statistical tests that have low power. We were able to decisively reject the hypothesis that $E\{\tilde{a} - \tilde{d}|x\} = 0$ when $\tilde{a}$ is identified as *initial award decision* by the DDS (i.e., the initial award decision made by the DDS before the possibility of appeals are considered).

award decision. As discussed above, there are certain physical conditions in SSA's "Blue Book" that it deems sufficiently severe to result in automatic qualification for DI benefits. We can account for this with a threshold rule with sufficiently large positive values for the components of β_a that correspond to the indicators of the various "listing conditions".

We can also represent the individual's self-reported disability status as a threshold rule of the form $\tilde{d} = I\{x\beta_d + \varepsilon_d > 0\}$, where x is the same vector of verifiable health indicators that the SSA observes, β_d is a vector of weights that the individual assigns to various health conditions and ADLs in making their own judgment of whether or not they are capable of working, and ε_d is a random variable reflecting the net effect of other information that the applicant observes, but the SSA (and the econometrician) does not observe. Under the normalization that ε_a and ε_d are normally distributed random variables with mean 0 and variance 1, it is possible to estimate the vectors of weights β_d and β_a that represent the SSA's and the applicants' threshold rules determining $\tilde{a}$ and $\tilde{d}$. The *rational unbiased reporting hypothesis* corresponds to the hypothesis that $\beta_a = \beta_d$, i.e., the SSA and the individual use the same weights in their threshold rules for $\tilde{a}$ and $\tilde{d}$, respectively. In BBCRS we were unable to reject this hypothesis, which, due to its parametric nature, implies a stronger restriction on the data than the unbiased reporting hypothesis, $E\{\tilde{a} - \tilde{d}|x\} = 0$, so the corresponding hypothesis tests had even greater power.

We view the RUR hypothesis as providing strong evidence that DI and SSI applicants are aware of the SSA's disability award process and adopt the SSA's standard in their own self-classification of whether or not they are disabled. The only reason why the SSA's ultimate award decision $\tilde{a}$ differs from the applicant's self-reported disability status $\tilde{d}$ is that $\tilde{a}$ is affected by the "bureaucratic noise" ε_a, whereas the individual's self-reported disability status $\tilde{d}$ depends on the applicant's private information ε_d about other unobserved health conditions and mental factors (i.e., the intangible "motivation" or "determination" factors discussed previously) that affect whether or not the applicant can work in spite of his/her physical impairments. Changes in the social/political regime that affect the degree of leniency in the SSA's standard for judging whether a DI or SSI applicant is disabled can be represented as changes in the β_a weights in the threshold rule representing its ultimate award decisions. If regimes change sufficiently slowly so that potential applicants can learn and adapt to the new disability standard, then the RUR hypothesis should hold independent of the particular regime and overall degree of leniency in the SSA's award process. Additional evidence supporting this hypothesis is provided by an analysis by Bound and Waidmann (1992) who compared time series data on self-reported disability to DI awards and roles. They concluded that "The changes in the fraction receiving benefits seem to have closely mirrored changes in the number of men (self) identified as disabled." (p. 1416).

Although the period of our study, 1992-1996, can be regarded as a "lenient regime" (since this was under the Clinton Administration when aggregate award rates were much higher than the historical average, reaching an all time high of 52% in 1998), our procedure for separating the question of leniency from the question of accuracy of the DI award process prevents us from making any judgments about whether the SSA's acceptance rates were "too high" during this period. However, there is an aspect of our analysis that could be interpreted as "stacking the cards" against the SSA in our evaluation of its accuracy. This is due to our interpretation that the unobservable component ε_a affecting SSA's ultimate award decision $\tilde{a}$ represents "bureaucratic noise" (i.e., random mistakes) whereas the unobservable component ε_d affecting an individual's self-reported disability $\tilde{d}$ represents "private information" on health conditions that are unobserved by the SSA and by the HRS interviewers.

An alternative interpretation of the RUR hypothesis is that ε_a represents "private information" that SSA has about the true health condition of the applicant that we do not observe, and ε_d represents idiosyncratic "errors" or deviations in individuals' personal thresholds for reporting whether they are disabled or not. Under this alternative interpretation, we could regard $\tilde{a}$ as representing a measure of "true disability". Of course, under this alternative interpretation the SSA would have no classification errors, and the deviations we observe between $\tilde{a}$ and $\tilde{d}$ are entirely due to idiosyncratic errors made by individuals in their self-reported disability status.

We dismiss the possibility that the SSA's ultimate award decision $\tilde{a}$ corresponds to the relevant notion of "true disability" because it seems implausible that the SSA would ever be able to gather all the relevant private information that affects whether an individual is capable of working or not. In particular, our example earlier in the introduction suggests that there are important intangible factors, such as "motivation" or "determination," that have an effect on whether a person is capable of working. These types of factors would be known by the applicant, but would be difficult, if not impossible, for the SSA to observe.

We do, however, acknowledge the possibility that, in addition to private information on health conditions and intangible factors such as motivation or determination entering the residual term ε_d in our model of self-reported disability, ε_d could also reflect other individual-specific "biases," or errors in judgment, about whether the person is really disabled. In this case it would be wrong to ascribe all of the discrepancies between $\tilde{a}$ and $\tilde{d}$ to mistakes on the part of the SSA. To handle this case we introduce a third binary latent variable $\tilde{\tau}$ representing the applicant's "true disability status." We assume that both $\tilde{a}$ and $\tilde{d}$ are noisy but unbiased indicators of $\tilde{\tau}$ so that we have $E\{\tilde{\tau}|x\} = E\{\tilde{a}|x\} = E\{\tilde{d}|x\}$. We also assume that $\tilde{\tau}$ can be represented by a threshold rule $\tilde{\tau} = I\{x\beta_\tau + \varepsilon_\tau > 0\}$. Finally, we assume that the RUR hypothesis holds and that the latent indicator of true disability is determined according to the same social/political standard

as the SSA's award decision. Thus, we have that $\beta_a = \beta_d = \beta_\tau$. In this case both $\tilde{a}$ and $\tilde{d}$ are noisy but unbiased indicators of an applicant's true disability status $\tilde{\tau}$. While the residual term in the SSA's award decision ε_a may reflect "bureaucratic noise" and the residual term in an individual's self-reported disability $\tilde{d}$ may reflect idiosyncratic biases and judgment errors, we assume that the residual term ε_τ, in the equation for true disability $\tilde{\tau}$, contains only unobserved private information on the applicant's health status and is free from noise or idiosyncratic judgmental biases. In general there may be common components in the three residual terms that cause these variables to be correlated. We normalize the marginal distributions of ε_τ, ε_a and ε_d to be $N(0,1)$, and assume that the correlations between these variables is given by

$$\text{corr}(\varepsilon_\tau, \varepsilon_a) = \text{corr}(\varepsilon_\tau, \varepsilon_d) = \text{corr}(\varepsilon_a, \varepsilon_d) \equiv \rho. \tag{1}$$

We can estimate the common correlation ρ, as well as β, from a bivariate probit model for the observed variables $(\tilde{a}, \tilde{d})$. We then use these estimates to compute Bayes estimates of the classification errors under the assumption that both $\tilde{a}$ and $\tilde{d}$ are noisy but unbiased indicators of true disability $\tilde{\tau}$.

3 The Social Security Disability Award Process

The Social Security Disability Insurance (DI) and Supplemental Security Income Disability Insurance (SSI) account for increasingly large components of social insurance spending in the U.S. Each program providing benefits to 6.7 million individuals in 2001, at a cost of $55 billion for DI, and $32 billion for SSI.[11] Although DI and SSI have the same total number of beneficiaries, 6.7 million, the DI program is nearly twice as expensive due to the fact that the average monthly DI benefit, $786, is twice as large as the average monthly SSI benefit, $385. Overall, these programs constitute approximately one-fifth of the SSA's total annual expenditures, and 75% of its administrative budget, 5% of the Federal Budget, and nearly 1% of U.S. GDP.

The volume of applications combined with the complexities involved in screening and adjudicating applications and appeals makes administration of these programs expensive and time consuming. For example, in 1998, the SSA processed more than two million applications for DI and SSI benefits, and over 500,000 appeals, at a cost of over four billion dollars, more than 67% of the SSA's total administrative

[11] DI was enacted in 1956 to insure covered workers, their spouses, and dependents against loss of earnings due to disability, under the strict definition of "disability" given in the introduction. Workers over the age of 31 are disability-insured if they had 20 quarters of coverage during the last 40 quarters and are fully insured. They are fully insured if they had at least one quarter of coverage for each year between 1950 (or age 21, if later) and the year they reached age 62 (or became disabled, if earlier). SSI was enacted in 1974 in part to cover gaps in coverage to people such as housewives, divorcees, and others who do not have sufficient work history to be covered under DI. The SSI program is more akin to welfare: it is mean-tested, and average monthly benefits are only about 50% as high as DI benefits. See Benítez-Silva et al. (1999), Apfel (1999), Bound and Burkhauser (1999), Haveman and Wolfe (2000), and the Social Security Advisory Board (1998) for more detailed information about these programs.

costs. There is a large bureaucracy involved in making disability determinations, including over 15,000 employees at the SSA's 54 DDS centers and over 1,000 Administrative Law Judges (ALJs) who handle the first stage of appeal beyond reconsiderations by the DDSs. The average cost of running this bureaucracy is about $2,000 per application. However, SSA may have ample justification for running this expensive and complicated "monitoring technology." According to the Social Security Advisory Board (1998, p.1): "It is estimated that a young, average-earning disabled worker and his family will receive about $285,000 over the course of their lifetime. ... nearly one out of three young men and nearly one out of four young women who are now age 20 will become disabled before reaching age 67."

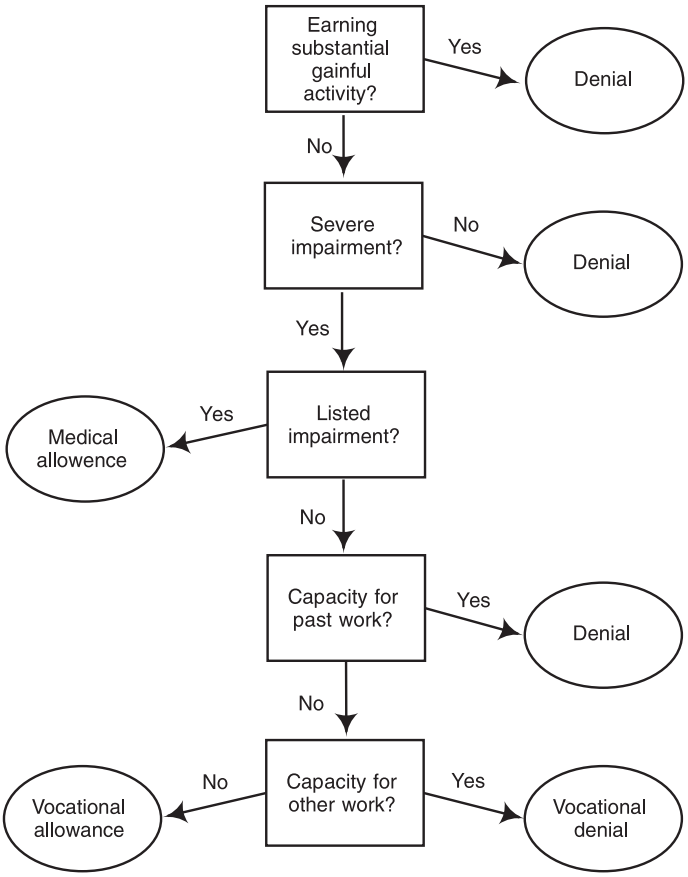
We now briefly describe the process by which SSA makes disability determinations. The process starts when an applicant files an initial application for DI benefits at one of the SSA's 1,300 field offices. This application is forwarded to one of the 54 DDS centers for processing, usually in the state where the claimant resides. The DDS makes initial or "first-stage" accept/reject decision according to a sequential five-stage screening procedure illustrated in Lahiri et al. (1995), which we reproduce in Figure 1. This procedure is designed to weed out inappropriate cases quickly, so that resources can be devoted to judging difficult cases where the determination of physical or psychological problems is less clear-cut. In 2001, the mean time for an initial decision by the DDS was about 90 days. Since the decision at each of the five stages is supposed to be independent of the decision made in previous stages, each of the five stages is handled by different personnel who are specialists in a particular stage of the process. This is the reason why large number of different people can be involved in evaluating a particular applicant.

In the first stage, the DDS determines whether or not the applicant has engaged in substantial gainful activity (SGA) subsequent to the claimed date of onset of disability. Any applicant who is found to earn in excess of the SGA threshold has demonstrated an ability to engage in SGA and is denied benefits at this stage. At the second stage, the severity of the physical or psychological problem is assessed. Applicants are denied if the impairment is not expected to last longer than 12 months or end in death. The third stage consists of a determination of whether the applicant's impairment is one of several hundred severe impairments, in the *Listing of Impairments* in SSA's "Blue Book." If the applicant's impairment appears in this list, then the applicant is automatically granted a *Medical Allowance.* Applicants who are denied a Medical Allowance are referred to the fourth stage, at which the DDS evaluates residual functional capacity, in order to determine whether or not the disability prevents the individual from being engaged in any element of his/her previous work. If this is found not to be the case, the applicant is denied benefits. Otherwise, the application is passed on to the fifth and final stage where the DDS evaluates the applicant's capacity for other work. Applicants deemed capable of engaging in another type of SGA

receive a *Vocational Denial*, while those found unable to do so receive a *Vocational Allowance.*

Applicants who are awarded benefits cannot begin receiving their benefits until the end of a five-month waiting period.[12] DI beneficiaries are entitled to Medicare coverage two years after the date of successful application, even if they are younger than 65, the normal eligibility age for Medicare coverage of Old Age insurance beneficiaries. The current average disability benefit is around $786 per month, approximately the same as the poverty line in the U.S.

Figure 1: Five Stage DDS Disability Determination Procedure

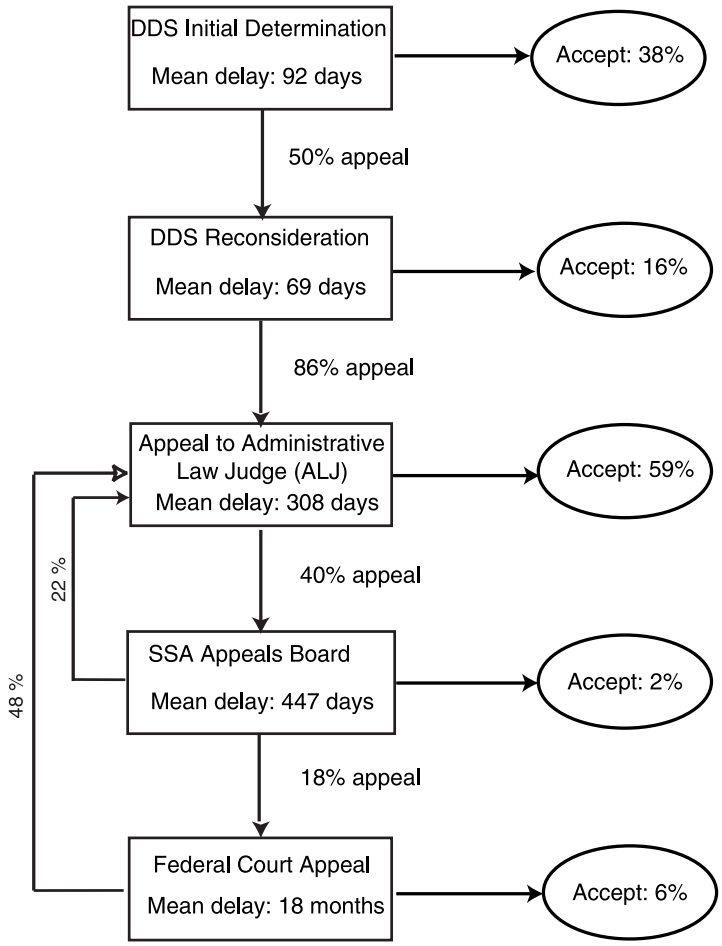


An applicant can appeal an initial rejection. There are four different appeal stages, illustrated in Figure 2. The first level of appeal is known as *Reconsideration* and is performed by the same DDS that made the initial determination. An application for reconsideration must be filed within 60 days of receipt of an initial denial notice. According to Social Security, 50% of denied applicants request a reconsideration, and the mean time required by the DDS to reach a decision on an SSDI reconsideration was 69 days. The acceptance rate at the reconsideration stage is 16%, lower than the 38% acceptance rate for initial determi-

[12] The start of waiting period is the later of (a) the date of disability onset; and (b) the date the applicant first attained disability insured status. It is waived if the applicant had a period of disability in the five years prior to the onset of the current disability.

nations. An applicant who is denied benefits at the reconsideration stage has 60 days to exercise the option to appeal to an ALJ. According to Social Security, approximately 86% of applicants who were denied at the reconsideration stage decided to appeal to an ALJ. In 2001, the mean time for a decision from an ALJ was 308 days, and the acceptance rate at this stage increases to 59%. An applicant who was denied benefits by the ALJ has 60 days to file a request for consideration at the central Appeals Board in Washington. According to Social Security, the Appeals Board considers about 40% of all ALJ dispositions, including cases it reviews on its own initiative. The mean duration for a decision from the Appeals Board is 447 days and the award rate is only 2%. An additional 22% of the cases heard by the Appeals Board are remanded back to the ALJ. After this stage the only remaining recourse is an appeal to Federal Court. These appeals involve average delays in excess of 18 months, substantial legal fees, and an acceptance rate of only 6%. It is also worth noting that 48% of the appeals to the Federal Court are remanded back to the ALJ level.

Figure 2: Summary of SSA's Disability Application and Appeal Process



The other Federal program providing disability benefits is the SSI, a means-tested cash assistance program enacted in 1974. Unlike the DI, there is no work requirement for SSI benefits. However, SSI applications are evaluated according to the same process as DI benefits and satisfy the same basic definition of disability. Furthermore, SSI is mean-tested with very low earnings and asset thresholds of $545 per month and $2,000, respectively, for a single individual.[13] As a result of different eligibility requirements, the SSI program serves a different "clientele" than does the SSDI program: 55% of disabled adults under 65 receiving SSI benefits are women, whereas 58% of adult SSDI beneficiaries are male. In contrast to DI, SSI recipients are not subject to the five-month waiting period and are immediately eligible for Medicaid benefits. However, monthly SSI benefits are significantly lower, averaging only $385 per month in 2001. Stapleton et al. (1994) show that since the late 1980s, the trends in applications, awards, and acceptance rates for the SSI and DI programs have been very similar. This is fortunate from our perspective, because the HRS data do not allow us to distinguish between the two programs.

4 Measuring Disability and Health Status in the HRS

This section provides a brief description of the Health and Retirement Survey (HRS), the data set we use to measure health and disability status of a sample of older Americans. The HRS provides highly detailed information on health and disability status, making it one of the best available data sets for conducting our analysis. We also include tabulations comparing several objective and subjective characteristics for various subsamples of DI applicants, non-applicants, recipients, and rejectees. Our results confirm previous conclusions by Benítez-Silva et al. (1999) that self-reported disability status, $\tilde{d}$, constitutes a very powerful predictor of application, appeal, and award decisions. We show that $\tilde{d}$ provides a more powerful predictor of a wide range of objective health and functional limitation measures, labor supply, and economic status measures, than the SSA's ultimate award decision $\tilde{a}$.

Indeed, when we classify applicants as disabled or non-disabled according to their self-reported disability $\tilde{d}$ we obtain much better discrimination between the two groups in terms of the degree of severity of an array of objective health status indicators and activities of daily living than we obtain when we classify individuals as disabled or non-disabled using the SSA's ultimate award decision $\tilde{a}$. For example, we find that disabled rejectees (i.e., individuals for whom $\tilde{d} = 1$ and $\tilde{a} = 0$) are much closer in terms of observed characteristics to disabled awardees (i.e., individuals for whom $\tilde{d} = 1$ and $\tilde{a} = 1$) than they are to non-disabled rejectees (i.e., individuals for whom $\tilde{d} = 0$ and $\tilde{a} = 0$). Similarly, we find that disabled

[13] The asset threshold excludes home, auto, household items, burial plots, and life insurance with face value under $1,500.

awardees are much closer to disabled rejectees in terms of their observed health condition than they are to non-disabled awardees. This suggests that if the SSA had the luxury of being able to observe applicants' truthful self-assessments of their disability status (i.e., $\tilde{d}$), then they would have been able to do a much better job in discriminating among those who are truly disabled from those who are not disabled. Of course, the SSA does not have this luxury, since all DI and SSI applicants would presumably report that they are unable to work in order to maximize the chances of being awarded benefits. However, in Section 6 we show that there is a feasible way for the SSA to improve its ability to discriminate between disabled and non-disabled applicants by using a statistical model that predicts the probability of self-reported disability, $\tilde{d} = 1$, in terms of a large number of objectively verifiable health conditions and activities of daily living.

4.1 Measurement and Data Issues

The data for our study come from the first three interviews of the HRS, a nationally representative longitudinal survey of 7,700 households whose heads were between the ages of 51 and 61 at the time of the first interview in 1992 or 1993. Each adult member of the household was interviewed separately, yielding a total of 12,652 individual records. Waves two and three were conducted in 1994/95 and 1996/97, respectively, using computer assisted telephone interviewing (CATI) which allows for much better control of skip patterns and reduces recall errors. Deaths and sample attrition reduced the sample to 11,596 individuals in wave two of the survey and to 10,970 individuals in wave three.[14]

The HRS has several advantages over the alternative sources of data previously used to analyze the DI award process such as the SIPP data. The HRS is a panel focusing on older individuals, with separate survey sections devoted to health, disability, and employment. The health section contains numerous questions on "objective" and subjective indicators of health status, as well as questions pertaining to activities of daily living (ADLs), instrumental activities of daily living (IADLs), and cognition variables. In the disability section, respondents were asked to indicate the dates they applied for DI benefits or appealed a denial, and whether or not they were awarded benefits.

However, the HRS does have some limitations that make it more complicated to study the DI award process. First, unlike the SIPP data, there is no match to the SSA Master Beneficiary Record, so we are unable to verify individuals' self-reported information on dates of application and appeal for SSDI and SSI benefits. Second, the HRS did not distinguish between SSI and SSDI applications, so all questions regarding disability implicitly combine the two programs into a single category.[15] Finally, the HRS did not

[14] Additional individuals, mostly new spouses of previous respondents, were added in waves two and three. We include these respondents in our analysis, yielding a total of 13,142 individual records.

[15] Henceforth, "DI" will denote both SSDI and SSI unless otherwise noted.

include appropriate follow-up questions that would have allowed us to determine whether DI applications or appeals reported in previous surveys had been awarded or denied, or whether they were still pending, resulting in potential censoring of information on appeals and reapplications. Fortunately, we were able to rectify some of these censoring problems using other information in the HRS. For example, the income section of wave two of the HRS included a question about whether respondents received Social Security income, and if so, the type of Social Security Income (DI benefits, retirement, etc.) and the date at which the respondent began to receive those benefits. For a previously pending case, an observed receipt of DI benefits after the application/appeal signified acceptance into the program. If no benefits had been received after 24 months following the application/appeal, we inferred a denial since virtually no cases are pending for longer than two years.[16]

Individual decisions as to when to apply or appeal for disability benefits are made in continuous time.[17] However, we observe individuals' health variables at points in time that are roughly two years apart. To most closely approximate an individuals' characteristics at the time of the application, we restrict our attention to the application/appeal episodes that were initiated within a one-year window surrounding the interview date (six months before to six months after), yielding a total of 387 observations.[18]

4.2 Data Analysis

Benítez-Silva et al. (1999 and 2001) conducted an internal validation of the quality of our "constructed" disability histories and the accuracy of individuals' responses to the HRS questions. In that paper we compared the dates of disability onset, application, and award, with a set of monthly labor supply dummy variables constructed from the work history section of the HRS. Since the two sets of variables were constructed independently using data from separate sections of the survey, there is no guarantee that the dates of the break in labor supply would correspond with the dates of disability onset. Yet, they match almost perfectly. Specifically, the results show a dramatic 50 percentage point drop in the labor force participation rate in the month following the onset of disability, falling from over 60% to under 15%.

[16] Additional strategies used to resolve ambiguous cases are detailed in Benítez-Silva et al. (1999).

[17] Given the panel nature of the HRS, we allow a single individual to yield several application episodes. We observe a maximum of three application episodes per person in the data, but most individuals have only one episode.

[18] When an applicant reported having a disability that prevented them from working entirely and the date of onset of this disability, we used the respondent's information from the closest HRS wave *after* the reported onset date, provided it was not more than 6 months after the onset date. If the respondent did not report having a health condition that prevented them from working entirely, we used health status information from the closest HRS wave *after* the date of application for DI benefits, subject to the constraint that it was not more than 6 months after. There are a total of 393 cases satifying these criteria, but due to missing data in some of the variables, we obtained a sample of 387 observations for which all the health status and socio-economic variables tabulated in Table 4 was available. We have experimented with windows of different length, and although this affects the number of observations, it does not significantly alter our results.

Table 3: Characteristics of DI Applicants and Non-Applicants

Variable	SSI/DI Applicants (1)	SSI/DI Non-Applicants (2)	SSI/DI Applicants		SSI/DI Non-Applicants	
			Non-Disabled (3)	Disabled (4)	Disabled (5)	Non-Disabled (6)
Number of Observations	1,179	27,415	348	831	955	26,460
Age	55.5	57.4	55.1	55.7	60.9	57.2
	(0.4)	(0.4)	(5.5)	(4.3)	(6.3)	(5.7)
White	55.2	75.6	55.7	55.8	63.7	76.0
	(1.4)	(0.3)	(2.7)	(1.7)	(1.6)	(0.3)
Male	43.0	44.9	39.9	44.3	42.0	45.0
	(1.4)	(0.3)	(2.6)	(1.7)	(1.6)	(0.3)
Married	65.5	82.4	66.4	65.1	78.7	82.5
	(1.4)	(0.2)	(2.5)	(1.6)	(1.3)	(0.2)
BA	7.2	23.0	8.9	6.5	7.8	23.5
	(0.7)	(0.2)	(1.5)	(0.9)	(0.9)	(0.3)
Prof. Degree	1.5	8.3	2.0	1.3	1.7	8.5
	(0.4)	(0.2)	(0.7)	(0.4)	(0.4)	(0.2)
Respondent Income	6,478	27,415	8,681	5,541	3,516	25,425
	(12,714)	(71,711)	(13,769)	(12,117)	(12,225)	(72,707)
Net Worth	91,994	265,698	86,180	94,458	161,604	269,130
	(197,857)	(536,364)	(153,228)	(213,933)	(413,295)	(539,602)
Annual Hours Worked	982	1,458	1,151	916	252	1,496
	(40)	(7)	(77)	(47)	(25)	(7)
Ineligible for SSI	19.5	26.4	20.5	19.0	58.6	25.3
	(1.2)	(0.3)	(2.2)	(1.4)	(1.6)	(0.3)
Hospital Stays	1.15	0.18	1.00	1.21	0.83	0.15
	(3.9)	(0.9)	(5.3)	(3.1)	(3.4)	(0.6)
Doctor Visits	14.0	5.0	10.8	15.3	12.1	4.7
	(15.7)	(6.9)	(13.7)	(16.3)	(11.9)	(6.5)
Poor Health	43.0	2.4	29.3	48.7	28.1	1.5
	(1.4)	(0.1)	(2.4)	(1.7)	(1.5)	(0.1)
Health Limitation Prevents Work	70.5	3.5	0.0	100.0	100.0	0.0
	(1.3)	(0.1)	(0.0)	(0.0)	(0.0)	(0.00)

Table 3: (Continued)

Variable	SSI/DI Applicants (1)	SSI/DI Non-Applicants (2)	SSI/DI Applicants		SSI/DI Non-Applicants	
			Non-Disabled (3)	Disabled (4)	Disabled (5)	Non-Disabled (6)
Is it difficult for you to:						
Walk across a room?	14.8	0.8	10.6	16.5	11.9	0.4
	(1.0)	(0.1)	(1.6)	(1.3)	(1.0)	(0.1)
Sit for long time?	45.7	13.7	36.2	49.8	45.0	12.6
	(1.4)	(0.2)	(2.6)	(1.7)	(1.6)	(0.2)
Get up from a chair?	58.6	20.6	47.4	63.2	59.2	19.2
	(1.4)	(0.2)	(2.7)	(1.7)	(1.6)	(0.2)
Climb Stairs?	46.7	5.5	35.6	51.4	38.5	4.3
	(1.4)	(0.1)	(2.6)	(1.7)	(1.6)	(0.1)
Take a Bath?	15.1	0.8	9.2	17.6	11.6	0.4
	(1.0)	(0.1)	(1.5)	(1.3)	(1.0)	(0.1)
Reading a Map?	32.6	14.6	33.5	32.2	33.1	13.9
	(1.3)	(0.2)	(2.6	(1.6)	(1.6)	(0.2)
Pick up a Dime?	15.8	2.3	13.0	16.9	13.3	1.9
	(1.1)	(0.1)	(1.8	(1.3)	(1.1)	(0.1)
Have you ever had:						
Cancer?	4.8	0.8	4.3	5.0	1.78	0.8
	(0.73)	(0.10)	(1.29)	(0.89)	(0.65)	(0.10)
Lung Disease?	18.2	4.6	16.4	19.0	15.1	4.2
	(1.1)	(0.1)	(2.0)	(1.4)	(1.2)	(0.1)
Stroke?	7.4	0.4	3.4	9.0	4.7	0.3
	(0.8)	(0.04)	(1.0)	(1.0)	(0.7)	(0.03)
Heart Problems?	23.8	6.7	18.1	26.1	22.1	6.1
	(1.2)	(0.1)	(2.0)	(1.5)	(1.3)	(0.1)
Psych. Problems?	25.3	6.9	21.0	27.1	20.4	6.4
	(1.3)	(0.1)	(2.2)	(1.5)	(1.3)	(0.1)
Nursing home stay?	3.6	0.2	1.7	4.4	1.6	0.1
	(0.6)	(0.03)	(0.7)	(0.7)	(0.4)	(0.02)

Chi-Square Tests of Equality of Means

Group 1 (# Obs.)	Group 2 (# Obs.)	χ^2	df	*p*-value
Disabled (1,596)	Non-Disabled (25,927)	2,932	15	0.000
Applicants (1,087)	Non-Applicants (26,436)	1,975	15	0.000
Non-Disabled Applicants (333)	Disabled Applicants(754)	115	15	0.000
Non-Disabled Non-Applicants (25,594)	Disabled Non-Applicants (842)	1,294	15	0.000
Disabled Applicants (754)	Disabled Non-Applicants (842)	131	15	0.000
Non-Disabled Applicants (333)	Non-Disabled Non-Applicants (25,594)	378	15	0.000

We now turn to Tables 3 and 4, which summarize the characteristics of our sample, presenting means and standard deviations of various economic and health status measures. In Table 3 we compare observed characteristics for DI/SSI applicants and non-applicants, while in Table 4 we compare the characteristics of awardees and rejectees for a subset of DI and SSI applicants for which we have uncensored observations on their application (including appeals, if any) and their award or denial of benefits, following the application or appeal. To obtain uncensored observations, we excluded applicants whose cases are still pending (either via the initial DDS decision or via an appeal to an ALJ). This provided us with a subset of DI applicants for whom we could determine the "ultimate award" decision by the SSA. This is the subsample that we use for testing the accuracy of self-reported disability status (Benítez-Silva et al. 2001), and for assessing the magnitude of classification errors in the DI award process. In each of these tables we further divide the groups into disabled and non-disabled individuals according to the value of the self-reported disability indicator $\tilde{d}$.

Comparing columns (1) and (2) of Table 3, we see that DI applicants are significantly worse off than non-applicants in terms of both their physical health and their economic "health." The income of non-applicants is more than four times higher and their net worth is nearly three times higher than applicants. The non-applicants also worked substantially more hours in the year prior to their interview, 1,458 hours compared to 982 hours for applicants. Also, DI applicants are significantly more likely to be female and non-white than non-applicants, and they appear less likely to have a family support network. For example, only 66% of DI applicants are married, compared to 82% for non-applicants. Applicants also have significantly less education: only 7% of applicants have a BA degree, compared to 23% for non-applicants. Although applicants do have earned income, the average amount earned is very close to the $6,000 SGA threshold prevailing during the 1992-1996 period covered by our sample.

The remaining rows of Table 3 show that applicants are significantly less healthy than non-applicants according to virtually all subjective and "objective" measures of health and ADLs. In the year prior to their interview, applicants had made 9 more visits to a doctor than non-applicants, were hospitalized nearly six times more often, and were 18 times more likely to have had an overnight stay in a nursing home. As for ADLs, a high percentage of DI applicants report difficulty doing various simple tasks than non-applicants: walking across a room (15% vs. 0.8%), getting up from a chair (59% vs. 21%), sitting for a long time (46% vs. 14%), or climbing stairs (47% vs. 5%). These high percentages suggest that many DI applicants do have difficulty performing common physical tasks that are part of most jobs. The "objective" health measures indicate, for example, that 24% of applicants had heart problems compared to only 7% of non-applicants and that 18% of applicants have lung disease, compared to only 5% of non-applicants. The

patterns for cancer, strokes, psychological problems and many other health problems not listed in Table 3 show a similar pattern, something that should not come as a surprise.

Focusing on columns 3 and 4 of Table 3, we see that all of the objective health status indicators and ADLs are worse for the subsample of SSDI and SSI applicants who were "disabled", i.e. those who reported that they had a health problem that prevented them from working entirely (i.e. $\tilde{a} = 1$). For example 9% of disabled applicants reported having a stroke compared to 3% of non-disabled applicants. Similarly the rate of heart problems and psychological problems is three times higher among disabled applicants.

At the bottom of Table 3 we provide χ^2 tests for the equality of means between the different subsamples given in the columns of Table 3.[19] In almost all cases we overwhelmingly reject the null hypothesis that the compared sub-populations are the same. The fact that most of the populations are different is not surprising, since the χ^2 statistics merely provide a convenient metric for summarizing the overall distance between health and functional status indicators for the various subgroups presented.

Although DI applicants are clearly in much poorer health than non-applicants, only 70% of the DI applicants reported that their health condition prevented them from working entirely. The 348 "non-disabled" applicants could represent the "imposters" who are attempting to "game the system" hoping that the SSA will make an award error and accept their application. We see from Table 3 that the non-disabled applicants are in significantly better health compared to the disabled applicants, at least in terms of all of the observable health indicators and ADLs presented in Table 3. The χ^2 statistics at the bottom of Table 3 indicates that we can decisively reject the hypothesis that the means of these health indicators and ADLs are the same for the disabled and non-disabled subpopulations of applicants.

It is also interesting to compare "disabled applicants" with "disabled non-applicants", in columns (4) and (5) of Table 3. We see that for most of the health indicators and ADLs, the disabled applicants are about as close to the disabled non-applicants than they are to non-disabled applicants. Indeed, these two groups are fairly similar in terms of their χ^2 statistics reported at the bottom of Table 3. The main difference between the two groups is that disabled non-applicants are significantly older and more likely to be white, female, and married. Many of these disabled non-applicants are married women who have not accumulated the 20 quarters of coverage necessary to be covered by DI and whose spouses' income or assets exceeds the means-test threshold for eligibility for SSI benefits. This hypothesis is confirmed by the

[19] All tests compare the means of the following variables for the different subgroups of the population: number of hospitalizations, nursing home stays, and doctor visits in the previous year, and the dummy variables poor health, stroke, cancer, heart problems, psychological problems, difficulty reading a map, picking up a dime, taking a bath, sitting for a long time, getting up from a chair, walking across a room, and climbing stairs.

fact that 59% of the disabled non-applicants are ineligible for SSI, whereas only 20% of the applicants are ineligible.[20]

It is worth emphasizing that the distance between populations is much larger when we compare disabled and non-disabled individuals than when we compare applicants to non-applicants, recipients to rejectees, or awardees to rejectees. We conclude that self-reported disability $\tilde{d}$ is a more powerful predictor of more objective health and functional status measures than other indicators, such as the indicators for having applied for DI, or being an SSI or DI recipient. In other words, the self-reported disability measure is superior to knowledge of the SSA's award decision as a determinant of the respondents' other health status measures. These findings are consistent with our hypothesis that $\tilde{d}$ is an indicator of "true disability" and that the award decision $\tilde{a}$ is a noisy indicator of $\tilde{d}$. For example, the χ^2 statistic measuring the distance between the disabled and non-disabled populations is 2,932, is 50% larger than the χ^2 statistic measuring the distance between applicants and non-applicants.

Table 3 also provides evidence of self-selection in the DI application decision. The mere act of applying for DI reveals a great deal of information about the applicant, since very few healthy, wealthy, or high income individuals apply for DI benefits. It is likely that this self-selection is largely due to rational behavior on the part of applicants. That is, healthy, well educated individuals know they are likely to be denied, and given the progressive structure of Social Security benefits, high income individuals have less of a financial incentive to apply for DI since they receive a much lower replacement rate than do low income individuals.

Table 4 compares the characteristics of awardees and rejectees for the subsample of 387 DI and SSI applicants for whom we can observe uncensored observations on SSA's ultimate award decision over the period 1992 to 1996 from the first 3 waves of the HRS. The SSA's ultimate award decision $\tilde{a}$ clearly enables us to discriminate among the applicants in terms of the severity of their health conditions: Nearly all of the "objective" health indicators and ADLs for the awardees are significantly worse than for rejectees. However, we find much larger differences in the objective health characteristics when we separate individuals according to self-reported disability status $\tilde{d}$ than when we separate individuals according to SSA's award decision $\tilde{a}$. We can see this in Table 4 by observing that nearly every health indicator or ADL is significantly worse for disabled awardees than for non-disabled awardees. For example 5.0% of disabled awardees report having had cancer compared to 1.6% of non-disabled awardees, and 7.7% of disabled

[20] See Benítez-Silva et al. (1999) A for an explanation of the construction of the eligibility variable. Another reason why disabled non-applicants may not be applying for benefits is that their mean age is 61, only slightly more than a year away from eligibility for early retirement benefits from Social Security at age 62. Benítez- Silva et al. (1999) show that the propensity to apply for DI benefits declines sharply near the eligibility age for Social Security retirement benefits.

awardees report that they had a stroke compared to 1.6% for non-disabled awardees. We also find that almost all of the observed health indicators for disabled rejectees are significantly worse than the observed health indicators of non-disabled rejectees.

At the bottom of Table 4 we report the χ^2 test statistics for the equality of the means of the various health indicators and ADLs listed in the table for the different subgroups. While we can reject the hypothesis that the observed health characteristics of disabled awardees and non-disabled awardees are the same, we are unable to reject the hypothesis that the health characteristics of disabled awardees and disabled rejectees are the same. In other words, the data suggests that in terms of observed health characteristics disabled awardees are much closer to disabled rejectees than to non-disabled awardees. Similarly, non-disabled awardees are more similar to non-disabled rejectees than they are to disabled awardees.

These findings are summarized in Figure 3. We see that the "χ^2 distance" between awardees and rejectees is 28 and statistically significant, confirming our earlier observation that the SSA's ultimate award decision $\tilde{a}$ does discriminate applicants in terms of the severity of objective health indicators. When we classify awardees based on their self-reported disability status $\tilde{d}$, we see that the χ^2 distance in the observable health characteristics of "disabled awardees" and "non-disabled awardees", 31, is *larger* than the χ^2 distance between awardees and rejectees. On the other hand, the χ^2 distance between "disabled awardees" and "disabled rejectees" is only 19 and is statistically insignificant. Similarly, the χ^2 distance between non-disabled awardees and non-disabled rejectees is also 19 and is statistically insignificant. This clearly suggests that self-reported disability $\tilde{d}$ provides a much better means of discriminating among our sample of DI and SSI applicants in terms of the severity of observable health conditions than the SSA's ultimate award decision $\tilde{a}$.

Indeed, the data suggest that if we were interested in awarding benefits to the least healthy individuals in this sample of applicants, the SSA should have awarded benefits to the 258 applicants who reported that their health impairment was sufficiently severe to prevent them from working entirely instead of the 266 people to whom the SSA actually awarded benefits. Of course, even if we believe that individuals provide truthful and accurate self-reports of their disability status in an anonymous interview such as the HRS, there is little reason to believe that they would truthfully report their disability in an application for DI or SSI benefits to the SSA. Thus, the SSA is at an inherent disadvantage since it must rely on an array of "noisy signals" such as the objective health indicators and ADLs shown in Tables 3 and 4. We return to this issue in Section 6, where we show that it is possible to construct a "statistical discriminant function" that uses a subset of the objective health indicators and ADLs that the SSA has access to, but results in significantly lower classification error rates than the SSA's current disability award process.

Table 4: Characteristics of Subset of DI Applicants

Variable	DI Awardees (1)	DI Rejectees (2)	Rejectees		Awardees	
			Non-Disabled (3)	Disabled (4)	Disabled (5)	Non-Disabled (6)
No. of Observations	283	104	43	61	221	62
Age	56.1	55.0	54.6	55.2	56.4	55.2
	(4.3)	(5.2)	(5.8)	(4.6)	(3.8)	(5.6)
White	59.0	44.2	51.2	39.3	59.7	56.4
	(2.9)	(4.9)	(7.6)	(6.2)	(3.3)	(6.3)
Male	39.2	45.2	48.8	42.6	39.4	38.7
	(2.9)	(4.9)	(7.6)	(6.3)	(3.3)	(6.2)
Married	58.0	63.5	55.8	68.9	57.9	58.1
	(2.9)	(4.7)	(7.6)	(5.9)	(3.3)	(6.3)
BA	6.7	9.6	11.6	8.2	6.8	6.4
	(1.5)	(2.9)	(4.9)	(3.5)	(1.7)	(3.1)
Prof. Degree	1.4	2.9	4.6	1.6	1.4	1.6
	(0.7)	(1.6)	(3.2)	(1.6)	(0.8)	(1.6)
Respondent Income	6,318	5,013	7,521	3,252	5,419	9,435
	(10,271)	(9,488)	(11,819)	(6,902)	(8,933)	(13,491)
Net Worth	76,583	81,847	41,847	114,220	73,911	87,017
	(121,890)	(244,341)	(73,049)	(318,387)	(106,429)	(168,833)
Annual Hours Worked	843	571	836	414	752	1,167
	(984)	(848)	(913)	(764)	(899)	(1,184)
Ineligible for SSI	22.6	32.8	25.8	38.9	25.5	28.3
	(3.0)	(5.7)	(7.9)	(8.1)	(3.4)	(6.6)
Hospital stays	1.0	0.7	0.5	0.8	1.0	0.9
	(1.6)	(1.6)	(0.8)	(2.0)	(1.4)	(1.9)
Doctor Visits	12.9	12.5	12.3	12.7	13.2	11.9
	(12.8)	(14.2)	(17.0)	(11.8)	(13.1)	(11.7)
Poor Health	46.9	39.8	23.8	51.8	49.3	38.7
	(3.0)	(4.9)	(6.6)	(6.7)	(3.4)	(6.2)
Health Limitation Prevents Work	78.1	58.6	0.00	100.00	100.00	0.00
	(2.5)	(4.8)	(0.00)	(0.00)	(0.00)	(0.00)

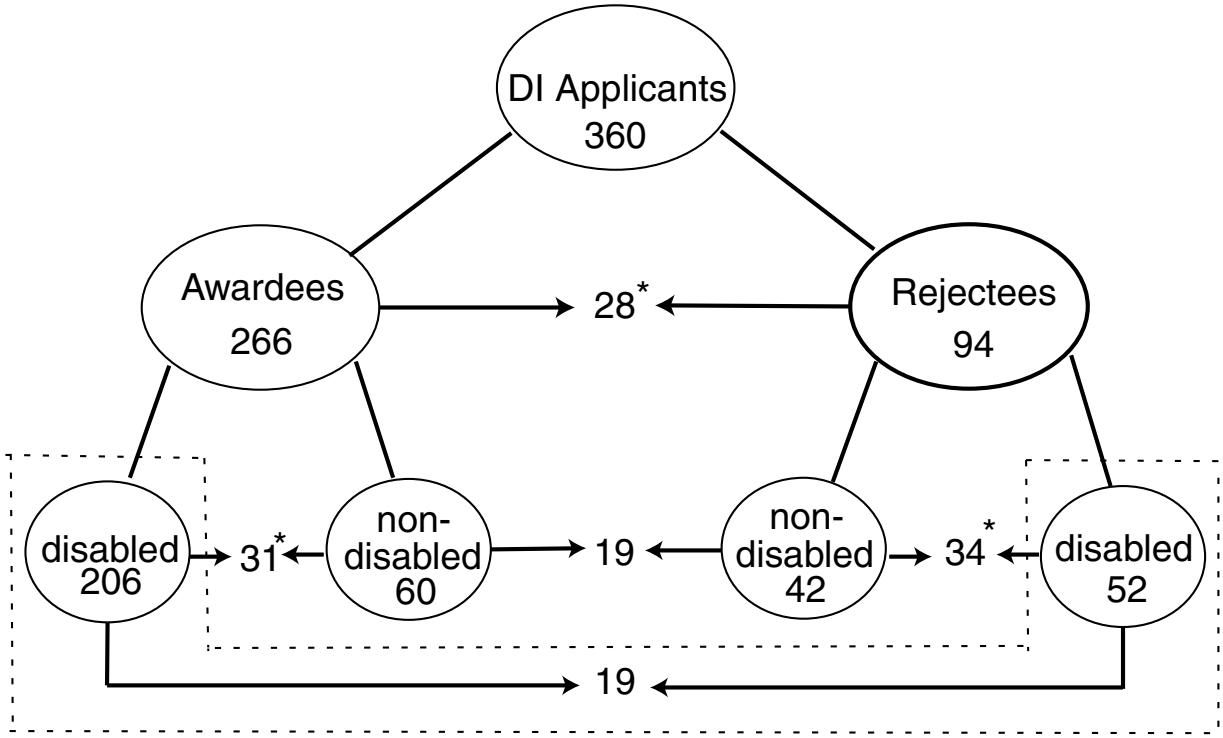
Table 4: (Continued)

Variable	DI Awardees (1)	DI Rejectees (2)	Rejectees		Awardees	
			Non-Disabled (3)	Disabled (4)	Disabled (5)	Non-Disabled (6)
Is it difficult for you to:						
Walk across a room?	16.6	8.6	2.3	13.1	17.2	14.5
	(2.2)	(2.8)	(2.3)	(4.3)	(2.5)	(4.5)
Sit for long time?	51.9	42.7	30.2	51.7	53.8	45.2
	(3.0)	(4.9)	(7.0)	(6.4)	(3.3)	(6.3)
Get up from a chair?	64.5	49.5	30.2	51.6	53.8	45.2
	(3.0)	(4.9)	(7.0)	(6.4)	(3.3)	(6.3)
Climb stairs?	48.8	38.5	34.9	41.0	51.6	38.7
	(3.0)	(4.8)	(7.3)	(6.3)	(3.4)	(6.2)
Take a Bath?	13.4	12.5	7.0	16.4	14.0	11.3
	(2.0)	(3.2)	(3.9)	(4.7)	(2.3)	(4.0)
Reading a Map?	32.6	14.6	33.5	32.2	33.1	13.9
	(1.3)	(0.2)	(2.6)	(1.6)	(1.6)	(0.2)
Pick up a Dime?	15.8	2.3	13.0	16.9	13.3	1.9
	(1.1)	(0.1)	(1.8)	(1.3)	(1.1)	(0.1)
Have you ever had:						
Cancer?	4.2	5.8	7.0	4.9	5.0	1.6
	1.2)	(2.3)	(3.9)	(2.8)	(1.5)	(1.6)
Lung Disease?	16.2	14.4	20.9	9.8	17.6	11.3
	(2.2)	(3.4)	(6.2)	(3.8)	(2.6)	(4.0)
Stroke?	6.4	10.6	2.3	16.4	7.7	1.6
	(1.4)	(3.0)	(2.3)	(4.7)	(1.8)	(1.6)
Heart Problem?	21.2	15.4	18.6	13.1	21.7	19.4
	(2.4)	(3.5)	(5.9)	(4.3)	(2.8)	(5.0)
Psychological Problems?	26.15	23.1	20.9	24.6	25.8	27.4
	(2.6)	(4.1)	(6.2)	(5.5)	(2.9)	(5.7)
Nursing home stay?	4.2	1.9	0.00	3.3	5.0	1.6
	(1.2)	(1.3)	(0.00)	(2.3)	(1.5)	(1.6)

Chi-Square Tests of Equality of Means

Group 1 (# Obs.)	Group 2 (# Obs.)	χ^2	df	*p*-value
DI Awardees (266)	DI Rejectees (94)	27.5	15	0.024
Non-Disabled Rejectees (42)	Disabled Rejectees (52)	34.2	14	0.001
Disabled Awardees (206)	Non-Disabled Awardees (60)	30.9	15	0.008
Disabled Rejectees (52)	Disabled Awardees (206)	18.8	15	0.222
Non-Disabled Rejectees (42)	Non-Disabled Awardees (60)	18.8	13	0.173
Non-Disabled Awardees (60)	Disabled Rejectees (52)	29.3	15	0.014
Non-Disabled Rejectees (42)	Disabled Awardees (206)	75.7	14	0.000

Figure 3: Summary of Classification Errors in the DI Award Process



5 Evaluation of the Disability Award Process

The previous sections have provided empirical support for the hypothesis that self-reported disability status $\tilde{d}$ is a valid measure of the "true disability" status. This is also supported by the findings of Benítez-Silva et al. (2001). Hence, $\tilde{d}$ can be used to measure the magnitude of classification errors in the DI award process. Although DI applicants presumably have strong financial incentive to misreport their disability status to the SSA, our results suggest that they accurately report private information regarding health status to the HRS interviewers, perhaps due in part to the survey's strong guarantee of confidentiality. These results justify our reliance on the HRS responses in measuring the rates of classification error at the various stages of the SSA's disability award process. Nevertheless, we also provide a sensitivity analysis in which we assume that neither measure, i.e., $\widetilde{d}$ or $\widetilde{a}$, are true measures of disability. We term the resulting estimated classification errors based on the latter calculations the *Bayes classification errors*. The results of this exercise do not change by much, validating again our assumptions about $\widetilde{d}$.

For clarity of exposition we first summarize the RUR hypothesis and its implications. We then provide the additional structure, which, in turn, allows us to compute the Bayes classification errors.

We assume that the SSA sets a "*social standard*" for disability that becomes common knowledge for all individuals applying for disability benefits. However, this social standard can, and most likely does, change over time. The SSA implements its definition via its award decisions. The RUR hypothesis is a claim that individuals and the SSA do not differ systematically in their valuation of the individuals' health conditions. Statistically, we formulate the RUR hypothesis in terms of the *conditional moment (CM)* restriction given by

$$E\left[\tilde{a} - \tilde{d} \middle| x\right] = 0, \tag{2}$$

where x denotes a vector of observed health and socioeconomic characteristics that are observed by both the individuals and the SSA. Since $\tilde{a}$ and $\tilde{d}$ are Bernoulli random variables, (2) is equivalent to $\Pr(\tilde{a}|x) = \Pr(\tilde{d}|x)$. In BBCRS (2003) we tested the CM restriction using several CM tests, but were unable to reject the null hypothesis that (2) holds.

We further focused on a parametric version of the RUR hypothesis, where the conditional probabilities are derived from a bivariate probit function given by

$$\begin{aligned} \Pr(\tilde{a}|x) &= E\left[I\left(x'\beta_a + \varepsilon_a \geq 0\right)\right], \quad \text{and} \\ \Pr(\tilde{d}|x) &= E\left[I\left(x'\beta_d + \varepsilon_d \geq 0\right)\right]. \end{aligned} \tag{3}$$

For this parametric model, the RUR hypothesis amounts to the restriction that $\beta_a = \beta_d$. As is commonly done in the literature on discrete choice models, we assume that $(\varepsilon_a, \varepsilon_d)$ have a bivariate normal distribution with correlation coefficient $\rho \in (-1, 1)$ and variances standardized to 1. Again, we were unable to reject the RUR hypothesis at conventional significance levels.

To motivate the structure in (3), consider the following *true disability indicator* $\tilde{\tau}$, which is not observed by either the SSA or the individuals. That is,

$$\tilde{\tau} = I\left(x'\beta_\tau + \varepsilon_\tau \geq 0\right). \tag{4}$$

The quantities $\tilde{a}$ and $\tilde{d}$ can be considered as noisy indicators of true disability $\tilde{\tau}$, held by the SSA and by the applicant, respectively. Furthermore, for this formulation to make sense, we also have

$$\varepsilon_a = \rho_a \varepsilon_\tau + \nu_a, \tag{5}$$

$$\varepsilon_d = \rho_d \varepsilon_\tau + \nu_d, \tag{6}$$

where ν_a and ν_d are independent of ε_τ and each other. The RUR hypothesis amounts then to the restrictions $\beta \equiv \beta_a = \beta_d = \beta_\tau$, with probability 1.

If we further normalize the variance of ε_τ, ε_a, and ε_d to be 1, then it follows that the correlation between ε_a and ε_d is given by $\rho \equiv \text{Cov}(\varepsilon_a, \varepsilon_d) = \rho_a \rho_d$. We also assume that $\rho \equiv \rho_a = \rho_d$. With these assumptions it follows that

$$\varepsilon = (\varepsilon_\tau, \varepsilon_a, \varepsilon_d) \sim \text{N}(0, \Sigma_\varepsilon), \tag{7}$$

where

$$\Sigma_\varepsilon = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho^2 \\ \rho & \rho^2 & 1 \end{pmatrix}.$$

The information in (3)-(7) is sufficient for computing the Bayes classification errors.

In these calculations the objective is to compute the classification errors relative to the true measure of disability, that is, our goal is to compute the: (a) *award error*, i.e., $\Pr(\widetilde{\tau} = 0 | \widetilde{a} = 1)$; (b) *rejection error*, i.e., $\Pr(\widetilde{\tau} = 1 | \widetilde{a} = 0)$; (c) *type I error*, i.e., $\Pr(\widetilde{a} = 0 | \widetilde{\tau} = 1)$; and (d) *type II error*, i.e., $\Pr(\widetilde{a} = 1 | \widetilde{\tau} = 0)$.

We demonstrate here how to compute the award error, the computation of the other classification errors are done similarly. The award error can be written as

$$\Pr(\widetilde{\tau} = 0 | \widetilde{a} = 1) = \int \Pr(\widetilde{\tau} = 0 | \widetilde{a} = 1, x) \, f_x(x) dx, \tag{8}$$

where $f_x(x)$ is the density of the observed characteristics. Note that the probability inside the integral in (8) can also be written as,

$$\begin{aligned} \Pr(\widetilde{\tau} = 0 | \widetilde{a} = 1, x) &= \Pr\left(\widetilde{d} = 0\right) \Pr\left(\widetilde{\tau} = 0 | \widetilde{a} = 1, \widetilde{d} = 0, x\right) \\ &\quad + \Pr\left(\widetilde{d} = 1\right) \Pr\left(\widetilde{\tau} = 0 | \widetilde{a} = 1, \widetilde{d} = 1, x\right). \end{aligned} \tag{9}$$

Further note that

$$\Pr\left(\widetilde{\tau} = 0 | \widetilde{a} = 1, \widetilde{d} = 0, x\right) = \frac{\Pr\left(\widetilde{\tau} = 0, \widetilde{a} = 1, \widetilde{d} = 0 | x\right)}{\Pr\left(\widetilde{a} = 1, \widetilde{d} = 0 | x\right)}, \tag{10}$$

and similarly for $\Pr\left(\widetilde{\tau} = 0 | \widetilde{a} = 1, \widetilde{d} = 1, x\right)$.

The probabilities in the numerator and the denominator in (10) can be easily computed, given the distribution of ε in (7), using, for example, the GHK algorithm and the coefficient estimate for β from Benítez-Silva et al. (2001). In the example of the probability in the numerator of (10),

$$\Pr(\widetilde{\tau} = 0, \widetilde{a} = 1, \widetilde{d} = 0 | x) = \Pr(\varepsilon_\tau < -x\beta, \varepsilon_a \geq -x\beta, \varepsilon_d < -x\beta | x).$$

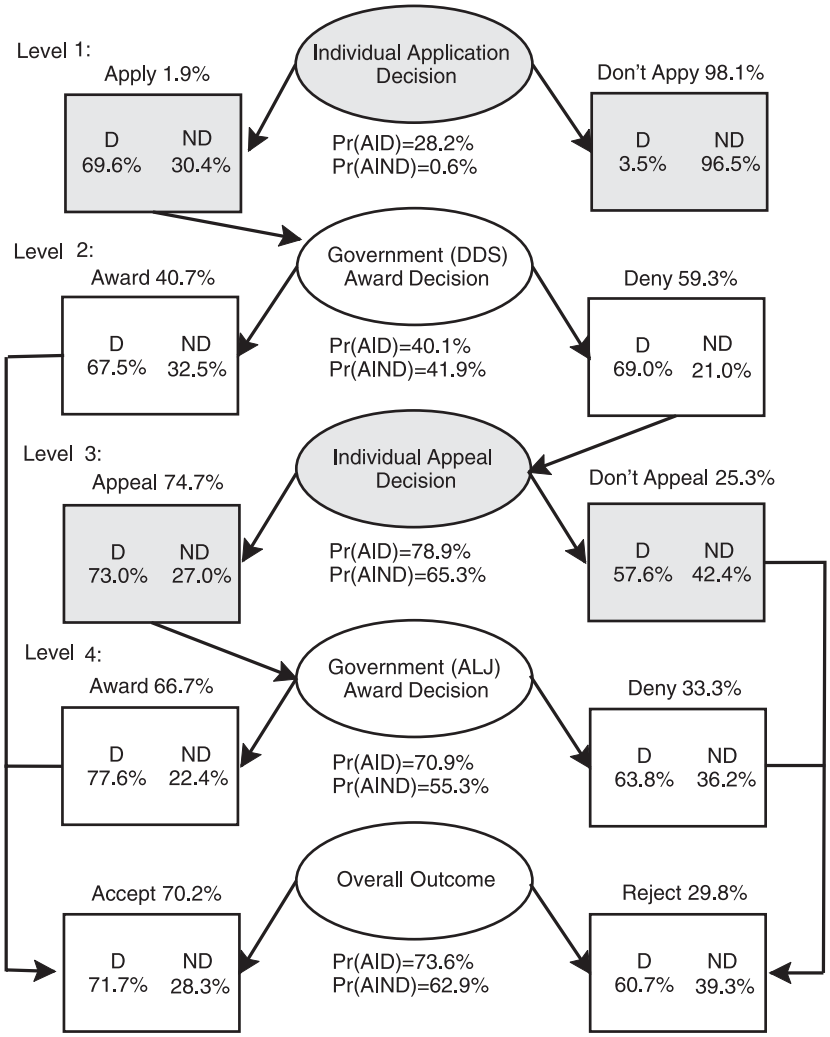
The computation of the *rejection probability* can be done in a similar manner.

Note that for type I and type II errors we have, respectively

$$\begin{aligned}\Pr(\widetilde{a}=0|\widetilde{\tau}=1) &= \Pr(\widetilde{\tau}=1|\widetilde{a}=0)\frac{\Pr(\widetilde{a}=0|x)}{\Pr(\widetilde{\tau}=1|x)}, \quad \text{and}\\ \Pr(\widetilde{a}=1|\widetilde{\tau}=0) &= \Pr(\widetilde{\tau}=0|\widetilde{a}=1)\frac{\Pr(\widetilde{a}=1|x)}{\Pr(\widetilde{\tau}=0|x)}.\end{aligned}$$

We turn now to the presentation of the main results. Figure 4 presents a simplified schematic diagram of the SSA's complete disability award process, with estimated classification error rates presented at each stage of the process. The first level of the figure represents the individual's decision whether or not to apply for benefits. Over a two year time period, 1.9% of individuals in the HRS applied for DI and the remaining 98.1% did not apply. The final node presents the outcome of the overall award process, including all possible appeals. The ultimate award rate is 70.2%, with an award error rate of 28.3% and a rejection error rate of 60.7%; that is, $\Pr(\tilde{a}=1)=.702$, $\Pr(\tilde{d}=0|\tilde{a}=1)=.283$, and $\Pr(\tilde{d}=1|\tilde{a}=0)=.607$.

Figure 4: Analysis of Classification Errors in SSA's Disability Award Process



Although the magnitude of these classification error rates provides evidence of considerable noise in the DI award process, it is clear that the SSA award decision $\tilde{a}$ is not arbitrary. Figure 4 indicates that $\tilde{a}$ is an informative signal, positively correlated with $\tilde{d}$, that succeeds in partially differentiating disabled and non-disabled applicants. In particular, a disabled applicant has a 73.6% probability of ultimately being awarded benefits compared to the lower award rate of 62.9% for non-disabled applicants. Still, 62.9% is a surprisingly high award rate for non-disabled applicants, suggesting that there could be relatively high returns for certain "imposters" to apply for DI benefits, perhaps explaining why 30.4% of DI applicants report that they are not disabled (see Level 1 of Figure 4). This leads to an obvious question: why does only 1.9% of the non-disabled population choose to apply for DI benefits during a two-year window? We will return to this issue shortly when we consider the factors responsible for the clear patterns of self-selection that we observe in the application and appeal decisions.

We now analyze the overall outcome, considering each stage of the determination process. Level 2 of Figure 4 shows the results of the SSA's first-stage award decision made by one of the 54 DDS centers, as discussed in Section 2. For our HRS sample, the first-stage award rate is only 40.7%, far smaller than the 70.2% ultimate award rate. This suggests that the eligibility threshold is significantly higher at the DDS stage than at the ALJ appeal level, depicted in level 4 of Figure 4. In particular, although the rate of award error is the same for the DDS stage as for the overall award process, the DDSs have a significantly higher rate of rejection errors, 69.0% (vs. 60.7% for the overall process with appeals), suggesting that the DDSs are erring on the side of rejecting applicants. This might be a reasonable strategy, since a rejected applicant has the option to appeal, whereas SSDI awardees are unlikely to leave the roles except via death or conversion to Old Age benefits at age 65. It seems plausible that the SSA perceives higher political and financial costs to making an award error relative to a rejection error, since the former can be more visible (e.g., via media exposés) and perhaps more difficult for the SSA to ferret out via Continuing Disability Reviews (CDRs). It is quite likely that the SSA may perceive a lower cost of rejecting applications at the first stage, especially since an individual is entitled to appeal an early rejection.

However, the conclusion that the DDS stage is more stringent and more inaccurate than later stages is not warranted given the self-selected nature of the pool of applicants choosing to appeal DDS rejections. This can be seen from level 3 of Figure 4. Although nearly 75% of rejected applicants choose to appeal, disabled candidates are much more likely to appeal than non-disabled candidates (73% vs. 27%). This implies that the self-selected pool of appealed cases considered by the ALJs has a higher proportion of "truly disabled" applicants than does the initial pool considered by the DDS (73% vs. 70%). Of the 25% of the initially rejected applicants who chose not to appeal, 42% are non-disabled.

The fourth level of Figure 4 represents the ALJ decision. The award rate at this stage, 66.7%, is substantially higher than at the DDS level. Despite this, the ALJ award error rate is slightly lower than that incurred by the DDSs (22.4% vs. 32.5%). This provides counterevidence to the claim implicit in the GAO study of the appeal process discussed in Section 2, namely, that the ALJs are too lenient and increase award errors through judicial reversals of poorly documented (but presumably valid) rejections at the DDS stage. Our results suggest that the ALJ contribution to the award process is beneficial, decreasing the high award error rate incurred by the DDS. Interestingly, we see that the rate of rejection errors among ALJs, 63.8%, is not much lower than the 69.0% rejection error rate at the DDS level.

In terms of the type-dependent success rates, both disabled and non-disabled applicants have a higher chance of being awarded benefits at the ALJ stage than at the DDS stage. However, the ALJ does improve a disabled applicant's odds of being awarded compared to those of a non-disabled applicant. At the DDS stage, disabled applicants have a slightly lower chance of being awarded benefits than non-disabled applicants (40.1% vs. 40.9The difference in award rates is much higher at the ALJ level (70.9% vs. 55.3%). The reduction in award and rejection errors at the ALJ level may not be a result of superior ability to discriminate: as we noted, there is significant self-selection in the appeal decision and this, combined with the uniformly higher acceptance rates at the ALJ stage, succeed in reducing the overall rate of rejection errors without increasing the overall rate of award errors. This result is contrary to the suggestions implicit in the GAO report and the recent literature on the disability process reform reviewed in Section 2.

Although it is difficult to quantify how much of the screening is accomplished by the applicants themselves, through self-selection, and how much is achieved by the SSA, through its "monitoring technology", it is important to note that the ALJs may have a significant advantage over the DDS due to the self-selected nature of those choosing to appeal. That is, 73% of rejected applicants who appeal are disabled compared to 58% of those who choose not to appeal. The initially denied applicants who appeal are a subset of an already highly self-selected applicant population, 70% of whom are disabled. The nature of the self-screening of applicants is closely related to the structure of the disability award process, most importantly the delays at the various stages. In previous work (Benítez-Silva et al. (1999), we estimated the delay distributions at each stage of the award process and showed that the "truly disabled" individuals (i.e., those for whom $\tilde{d} = 1$) are more likely to persist at each stage, and ultimately be awarded benefits.

One might be concerned that the estimated classification errors might be biased, since, after all, $\tilde{d}$ cannot be literally taken to be the true measure of disability. As explained at the beginning of this section we also calculated the Bayes classification errors and the results were very similar. The overall award error, i.e., the probability that an awardee is, in fact, non-disabled is 23%, slightly lower than the 28% award rate

obtained under the assumption that $\tilde{d}$ is the true disability measure (*the base estimates*). That is, if we do not take $\tilde{d}$ to be the true measure of disability the estimated award error is basically unchanged. However, our estimate of the rejection error rate is unchanged when we allow for the possiblity that self-reported disability status $\tilde{d}$ is a noisy indicator of true disability status $\tilde{\tau}$: the rejection error rate is 61% which is the same as the 61% rejection error rate we estimated under the assumption that $\tilde{d} = \tilde{\tau}$ with probability 1.

Our Bayes estimates of the type I and type II errors ere 23% and 68%, respectively, compared to 26% and 68% under the assumption that $\tilde{d}\tilde{\tau}$ with probability 1. Thus, allowing for the possibility that $\tilde{d}$ is a noisy but unbiased indicator of true disability $\tilde{\tau}$ *worsens* our estimates of the rate of classification errors in SSA's disability award process. Overall, the Bayes and base results are very similar, strengthening the claim that $\tilde{d}$ seems to be a very accurate measure of disability.

As noted above, to the extent that SSA does make accurate classifications, most of the credit appears to be responsible to the applicants themselves, due to self-screening in the application and appeal process. Part of the self-screening is due to the non-disabled individuals' perception of the odds for being awarded benefits. Overall, a disabled applicant has a higher expected success rate (73.6%) than does a non-disabled applicant (62.9%). Nevertheless, this difference does not seem large enough to explain the magnitude of the observed self-selection in the application and appeal decisions. Another key explanation for the self-selectivity is *processing delays*. While some the delays are unintentional, in that they resulted from the rapid recent increases in application rates, delays also have important strategic consequences for applicants. Specifically, delays tend to act as an "application fee" that assists the SSA in distinguishing between disabled and non-disabled candidates. The SSA is able to use this "price discrimination", because non-disabled applicants incur a greater opportunity cost than "truly disabled" individuals, for whom the opportunity cost is, essentially, zero. However, to the extent that there are liquidity constraints, preventing an applicant from borrowing to finance consumption during the long period that an application is pending and the applicant is out of work, delays do impose deadweight welfare costs on all applicants. A more structural approach would be required to incorporate these real welfare costs as an important component of the overall costs and benefits of the current DI process.

A former SSA's commissioner, Kenneth Apfel, seems to have been aware of some of our conclusions about the sources of classification errors in SSA's disability award process, particularly the problem of excess stringency on the part of the DDS, and the need for frequent reversals at the ALJ stage:

> "The SSA strives to deliver the highest levels of service by making fair, consistent and timely decisions at all adjudicative levels. However, applicants and beneficiaries sometimes find the current process complex, confusing and impersonal. Some also perceive the process as one in which different decisions are reached on similar cases at different levels of the administrative

> review process, thus requiring applicants to maneuver through multiple appeals steps before they receive benefits. Furthermore, denial cases are more error prone than are allowance cases at the initial claims level while the opposite is true at the hearing level." (Apfel 1999, p.11).

Apfel's assertion that denial cases are more error prone than allowance cases is consistent with our finding: The DDS rejection error rate is 64%, while the award error rate is only 23%. In contrast, our results do not accord with Apfel's statement that award cases are more error prone than denial cases at the hearing (ALJ) level. Our results indicate that the ALJs have virtually the same award and rejection error rates as the DDS, i.e. a 23% rate of award errors and a 65% rate of rejection errors.

6 An Analysis of a Computerized Disability Screening Process

This section considers whether it is possible to outperform the current disability award process in the U.S. using a simple computerized screening rule for determining awards. We compare the error rates incurred under the current disability process to those implied by a feasible computerized disability screening procedure that uses an alternative index rule for determining disability status. We show that this alternative screening rule significantly reduces the award and rejection errors in the disability award process. We present the computerized screening rule and discuss some caveats and problems that might arise in its implementation. In particular, the use of computerized screening procedures does not obviate the need for human input. We conclude with a discussion of some of the higher level bureaucratic incentive problems that are involved in implementing alternative screening procedures, including the proposed disability process reforms discussed above.

The computerized rule is motivated by discussions within the SSA regarding the need for increased standardization and internal consistency in the award process, as well as improved documentation of the grounds for award decisions. Additional motivation is provided by the validity of the *rational unbiased reporting* (RUR) hypothesis presented in BBCRS (2003).

The RUR hypothesis implies that if the SSA had the luxury of observing each applicant's "true disability" status $\tilde{d}$, it could simply use $\tilde{d}$ as the basis for its award decisions. But, of course, the SSA does not have this luxury. In Benítez-Silva et al. (2001) we argued that the DI applicants accurately and truthfully report their disability status in an anonymous survey, even though they have a strong incentive to misreport their status to the SSA. Although the SSA cannot observe $\tilde{d}$ for each applicant, it can *predict* the value of $\tilde{d}$ using a set of observed characteristics, say x. In particular, if it believes that the RUR hypothesis holds for respondents in an anonymous survey, it can use data such as that provided by the HRS to regress observed values of $\tilde{d}$ for individuals in the survey on their observed x values. This results in an estimated

conditional probability $\hat{P}(\tilde{d}|)$, and estimate for $P(\tilde{d}|x) \equiv \Pr(\tilde{d}=1|x)$, that can be regarded as the SSA's "posterior belief" that an applicant, with observed verifiable characteristics x, is "truly disabled." The computerized rule is then to simply accept all applicants whose predicted probability of being disabled is sufficiently large. That is, the computerized award decision a_c is defined by

$$a_c = I\left\{\hat{P}(\tilde{d}|x) \geq \rho_c\right\}, \tag{11}$$

where $\rho_c \in [0,1]$. Adjusting the threshold value ρ_c produces different award rates, with a decrease in award rate achieved by an increase in ρ_c. If the SSA wanted to target a fixed award rate p_c, it would determine the threshold value ρ_c as the smallest solution to

$$p_c = \int I\left\{P(\tilde{d}|x) \geq \rho_c\right\} f(x)dx. \tag{12}$$

We computed values of the threshold ρ_c to match the 56% acceptance rates at the DDS stage and the 75% ultimate award rate in the current DI award process, analyzed in Section 4. This was done by specify6ng an initial guess for ρ_c, calculating the sample analog of (12), using the empirical distribution of x in the HRS data, and increasing or decreasing ρ_c until the implied award rate matched the desired award rate p_c.

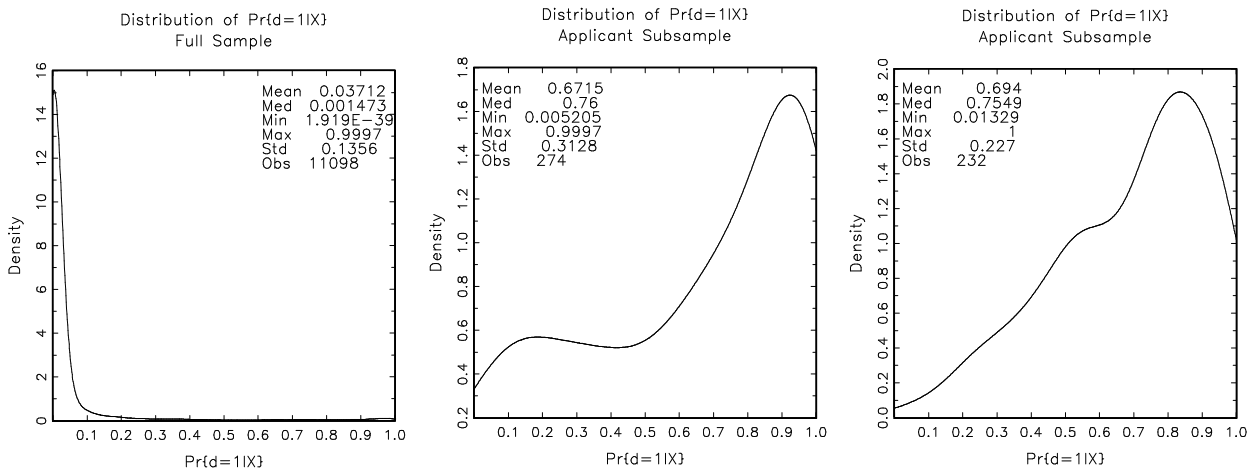
Table 5 presents estimates of a logit specification for $P(\tilde{d}|x)$ using the HRS data. We present two sets of results, one for the full sample of applicants and non-applicants, and another for the subsample of DI applicants used in the previous section. The results are highly significant and generally of the expected signs. Specifically, the coefficients for diabetes, stroke, number of hospitalizations and doctor visits, and other indicators of poor health are all positive, and are all important predictors of $\tilde{d}=1$ (i.e., the event that an individual reports having a health problem preventing all work). The variables "applied for DI" and "proportion of months worked in the past year" are the two most important predictors of disability status in the full sample results. Although there are far fewer observations in the applicant subsample, we feel these results are the more relevant for this paper since it is for this subsample that we have verified the RUR hypothesis. Because we cannot compare a non-applicant's reported value of $\tilde{d}$ with $\tilde{a}$, we are unable to test the RUR hypothesis for the larger sample of non-applicants. However, we have no strong *a priori* reason for believing that non-applicants are any more accurate or truthful than applicants. Although non-applicants should have less of an incentive than applicants to misreport their disability status, non-applicants are likely to be more poorly informed about how the DI award process works, unless they are contemplating applying in the near future. Thus, it is possible for a non-applicant's report of $\tilde{d}$ to be noisier and less accurate than that of an applicant.

Table 5: Logit Estimates Predicting Self-Reported Disability Status $\tilde{d}$

		Full Sample		Applicants sample	
No.	Variable	Estimate	St. Error	Estimate	St. Error
1	Constant	-6.15	1.17	1.53	3.03
2	Have applied	3.15	0.28	—	—
3	Non-eligible for SSI/SSDI	0.46	0.18	-0.41	0.47
4	White	-0.26	0.19	0.19	0.41
5	Vocational training	0.10	0.18	0.00	0.44
6	Bachelor Degree	-0.08	0.25	0.33	0.60
7	Male	0.69	0.20	0.20	0.45
8	Married	-0.18	0.55	-0.20	0.87
9	Divorced	-0.33	0.59	-0.47	0.91
10	Application Age	0.03	0.016	-0.01	0.05
11	Previously applied for DI	-0.44	0.68	0.47	0.55
12	No. of hospitalizations in past year	0.12	0.11	0.14	0.21
13	No. of doctor visits in past year	0.03	0.009	0.03	0.01
14	Had High Blood Pressure	-0.39	0.28	0.12	0.42
15	Had Diabetes	0.23	0.30	-0.45	0.46
16	Had Cancer	0.05	0.54	0.64	0.80
17	Had Lung disease	0.01	0.33	-0.14	0.56
18	Had Coronary Problems	0.68	0.23	-0.06	0.44
20	Had Heart Surgery	0.21	0.64	-0.94	0.95
21	Previous stroke	0.71	0.86	0.29	1.07
22	Had Arthritis	0.09	0.19	-0.16	0.45
23	Back problems	0.42	0.16	-0.38	0.41
24	Feet problems	0.48	0.17	-0.07	0.43
25	Memory Test	-0.09	0.03	-0.04	0.07
26	Cognitive Test	-0.02	0.03	-0.07	0.08
27	Difficulty jogging	0.70	0.23	1.21	0.57
28	Difficulty walking across a room	1.00	0.49	0.66	0.83
29	Difficulty sitting for a long time	0.29	0.18	0.25	0.46
30	Difficulty getting up from a chair	0.26	0.18	0.71	0.48
31	Difficulty using the stairs	0.55	0.22	0.59	0.45
32	Difficulty carrying objects	0.32	0.20	-0.71	0.54
33	Difficulty stooping or crouching	0.29	0.18	-0.84	0.56
34	Difficulty bathing	-0.14	0.58	-0.89	0.70
35	Difficulty reaching objects	0.17	0.21	0.59	0.45
36	Difficulty pushing objects	1.17	0.20	1.12	0.51
37	Difficulty getting dressed	1.13	1.86	14.4	0.87
38	Difficulty eating	1.14	1.58	-2.09	1.93
39	Difficulty Reading a map	0.18	0.18	-0.30	0.40
40	Current Smoker	0.39	0.18	-0.07	0.43
41	Current Drinker	-0.28	0.16	-0.14	0.39
42	Mother Alive	0.03	0.04	-0.05	0.10
43	Father Alive	-0.05	0.06	-0.16	0.13
44	Proportion of months worked in past year	-3.71	0.41	-0.72	0.69
45	Total Family Income ($1000) in past year	0.00	0.00	0.00	0.00
46	Respondent's earnings ($1000) in past year	-0.00	-0.00	0.00	0.00
47	Total Hours Worked (in 100) in past year	0.05	0.02	0.00	0.04
	Avg. Log L/Obs.	-0.0655	11,098	-0.4852	232

Figure 5 plots the distribution of estimated probabilities $\Pr\{\tilde{d}=1|x\}$ for different subsamples. The far left panel of figure 5 plots the distribution of $\Pr\{\tilde{d}=1|x\}$ for the full sample of $N=11,098$ individuals corresponding to the left hand columns of Table 5. We see that in the full sample, most individuals have a predicted probability of being disabled that is very close to zero. There is, however, a long thin tail corresponding to the small number of individuals who have high predicted probabilities of being disabled. From the leftmost column of Table 5, we see that by the most powerful predictor of being disabled, $\tilde{d}=1$, is the dummy variable for applying for SSDI or SSI benefits. The middle panel of figure 5 plots the predicted probabilities of $\Pr\{\tilde{d}=1|x\}$ for the subsample of 274 SSI and SSDI applicants for which we have complete information on all the x variables in Table 5. We see that for this subsample, the distribution of predicted probabilities of $\Pr\{\tilde{d}=1|x\}$ is skewed to the right, although there is evidence of a secondary mode in the distribution correponding to individuals with low values of $\Pr\{\tilde{d}=1|x\}$. The secondary mode corresponds to the 30% of DI applicants who are not disabled.

Figure 5: Distributions of $\Pr\{\tilde{d}=1|x\}$ for Different Subsamples



The third panel of figure 5 plots the distribution of $\Pr\{\tilde{d}=1|x\}$ that emerges when we re-estimate the model using only the subsample of SSI and SSDI applicants. This panel corresponds to the 232 applicants in the second column of Table 5. This distribution is even more skewed to the right than the middle panel, and the "secondary mode" in the distribution with values of $\Pr\{\tilde{d}=1|x\}$ near zero has disappeared. We feel this distribution is the most relevant one to use for determining the cutoff levels for $\Pr\{\tilde{d}=1|x\}$ in our "computerized screening rule." The reason is that the individuals who are most likely to be aware of the SSA's definition of disability are the subsample of SSI and SSDI applicants. Furthermore, as we have seen in Table 3, the applicant population is very different than the full HRS sample. Indeed, we see significant differences in the logit coefficient estimates in Table 5 between the full sample and the subsample of SSI

and SSDI applicants. For these reasons, there is no reason to believe that the relationship between various x variables and self-reported disability status $\tilde{d}$ should be the same for the applicant population and the full sample. The significant differences we observe in the two columns in Table 5 and in the two right hand panels of figure 5 confirm this.

Figures 6 and 7 present the results of our comparison of the classification errors resulting from the computerized screening rule a_c and the SSA actual award decision $\tilde{a}$, for the full sample and applicant subsample, respectively. After calculating the threshold ρ_c, the computerized screening yields an accept/reject decision a_c for each person in the sample. Using these computer-generated award decisions, we evaluated the classification errors using $\tilde{d}$ as a measure of "true disability" status, just as we did in our evaluation of the actual DI award process in Section 4. Each figure presents two comparisons. The top part of each figure compares the classification errors implied by the computerized rule to the classification errors of the actual process in the first stage, accomplished by setting the acceptance rate p_c equal to the first-stage DI award rate of approximately $p_c = .56$.[21] The bottom part of Figures 6 and 7 compare the computerized screening rule to the overall outcome of the award process when the option to appeal is allowed. Here, p_c is set equal to the ultimate award rate rather than the first-stage award rate set by the DDS.

For the full sample, we find that the computerized screening rule substantially outperforms the DDS. As seen in the top part of Figure 6, the computerized screening rule results in an award error rate of 17.7% compared to the actual DDS rate of 28.5%. Similarly, the computerized screening rule results in a large reduction of the rejection error: 52.9% of those rejected by the computerized screening rule are disabled compared to 66.7% of the DDS rejectees. The computerized rule achieves better discrimination between disabled and non-disabled applicants, yielding a first stage award rate for disabled applicants of 66.5% and a 32.4% first stage award rate for non-disabled applicants. In comparison, the DDSs appear to have great difficulty distinguishing between disabled and non-disabled applicants. The success rate for disabled applicants, 57.8%, is only slightly higher than the success rate of non-disabled applicants, 52.1%. In contrast, the bottom part of Figure 5 shows that the computerized screening rule does not improvement on the overall DI award process when appeals to ALJs are included. Although there are still numerous discrepancies between the SSA's award decision and the computerized award decision, these discrepancies tend to be offsetting so that the two procedures yield approximately the same rates of classification errors.

[21] The actual award rate differs slightly from the $p_c = .57$ target in Figure 4 since we had to condition on a subsample of applicants for whom all 47 covariates had no missing values. For this subsample, the first-stage award rate happened to be slightly lower, 56%. We used this actual award rate as the basis for our comparison.

Figure 6: Computerized Screening Rule: Full Sample

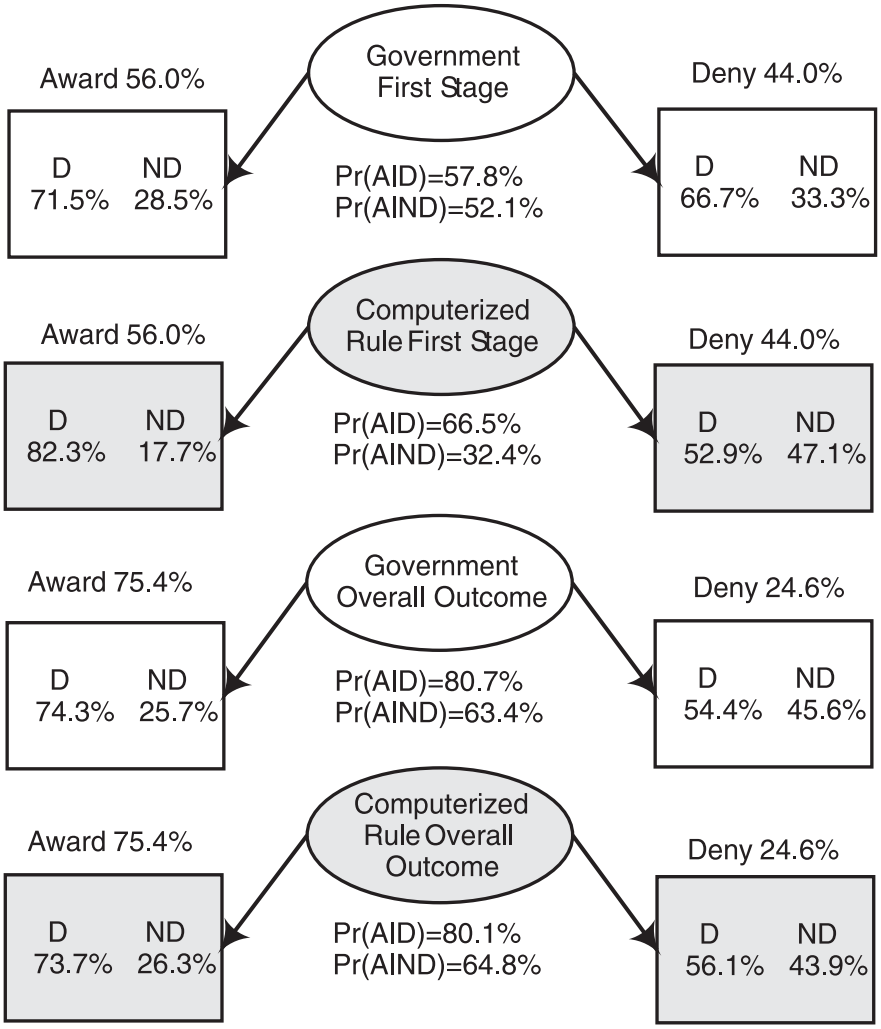
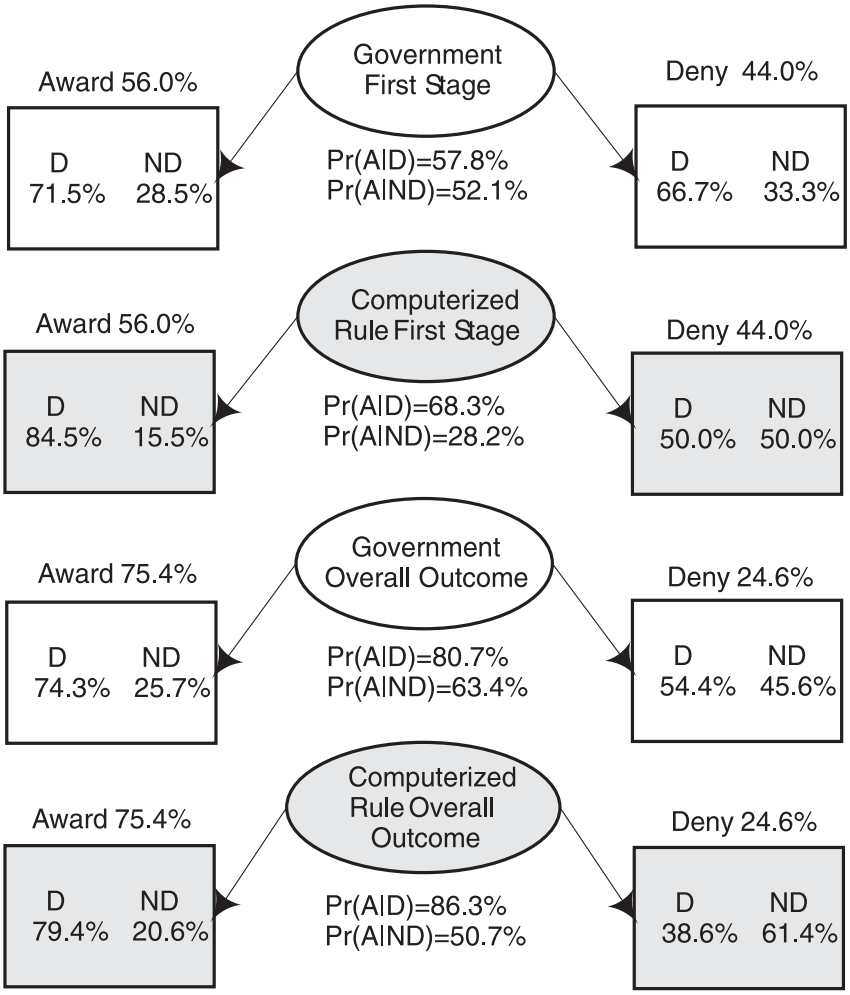


Figure 7 shows that the computerized screening rule does uniformly dominate the SSA's disability award process when we compute $\Pr(\tilde{d} = 1|x)$ for the subsample of SSDI and SSI applicants. Figure 7 shows that both at both stages of the disability award process the computerized rule dominates the SSA's award process in terms of award and rejection error rates. In the first stage, the computerized rule results in a 15.5% award error rate and a 50.0% rejection error rate, which are significantly lower than the SSA's error rates of 28.5% and 66.7%, respectively. For the overall process, the computerized rule reduces the award error from 25.7% to 20.6% and reduces the rejection error rate from 54.4% to 38.6%. These results seem to suggest that the smaller, more relevant subsample of DI applicants yields a better predictor of $\tilde{d}$ than does the full sample, even though the sample is quite small.

We conclude this section with a discussion of several caveats regarding the interpretations of these findings. First, note that the superiority of the computerized screening rule using only the 47 observed

covariates x is not unexpected: it is a simple consequence of the efficiency of maximum likelihood estimation, with $\hat{P}(\tilde{d}|x)$ a maximum likelihood estimator of "true disability" status $\tilde{d}$. But the SSA presumably has access to far more information about the applicant than the 47 or so variables we used in our computerized model of disability determinations. To the extent that the SSA does have more information, it should incur fewer classification errors. However, the value of this extra information could well be offset by the "bureaucratic noise" ε_a discussed in the presentation of the RUR model in Section 5. Because the SSA does not observe $\tilde{d}$, it does not have the luxury of systematically tweaking the parameters of its disability award process to best predict it. These factors can explain why classification errors are larger for the SSA's disability award process. The billion dollar question is whether a computerized screening rule similar to the one described here could actually be implemented, and if so, whether or not it could be easily exploited while avoiding certain serious problems of its own.

Figure 7: Computerized Screening Rule: Sample of Applicants



Clearly, the computerized screening rule does not obviate the need for human beings and some type of disability determination bureaucracy. Note that there would be enormous incentive compatibility problems if applicants ever became aware of the threshold rule in equation (11). In this case they would have a strong incentive to distort their x characteristics, reporting values to the SSA that would lead to large values of $\hat{P}(\tilde{d}|x)$. To avoid this problem, we suggest that the x measurements be made by a team of doctors and vocational counselors, similar to what was done in the Nagi study. This team would be paid by the SSA to visit the home of each DI applicant and conduct an examination to secure the necessary measurements. The team would simply report the recorded x information back to the SSA, where a disability determination would be made by applying the uniform computerized rule given in equation (11). Since the teams of doctors would be agents of the SSA rather than of the DI applicant, we could be confident that the problem of misreporting and mismeasuring of x variables in order to game the system would be minimized.

There is an additional potential problem that members of the disability examination team might have personal feelings about the disability status of a candidate and attempt to manipulate the award outcome by distorting the reported value of x. This could be minimized by pairing experts in the disability examination teams and having randomly scheduled audits or tests where another examiner would conduct an independent examination of an applicant, and the original examiner could be penalized if the auditor discovered errors in the original examiner's diagnoses.

We envision that a computerized screening rule would only be used as a replacement for initial determinations at the DDS level. The appeal process would remain fully human, allowing judgment to be added to the purely algorithmic approach of the decision procedure suggested above. If an appeal established an error in an applicant's medical exam, the examiner at fault could be subjected to penalties, providing further incentive for the SSA examiners to be as accurate as possible in measuring the applicant's x characteristics. Recall that the SSA bases its decisions on additional information about the applicant, information that is inaccessible through the HRS. If a computerized screening approach similar to the one discussed here were to be pursued seriously, it would be necessary to conduct a separate disability evaluation survey that followed a large sample of DI applicants and included all of the additional information that the SSA would like to use in a computerized award rule. It would be critical for the survey to provide strong guarantees of confidentiality and have no association with the SSA per se, in order to ensure honest reporting of disability status on the part of respondents.

More ambitiously, SSA might conduct a study similar to Nagi's (1969) study, with a team of medical expert interviewers recording all the information as well as their own collective judgment of the disability status of the applicant. Further econometric analysis of the reliability and accuracy of self-reported disabil-

ity could be performed by comparing the computerized results with those made by the collective judgment of a moderated team of medical experts. Although we stand by our finding that self-reported disability is an accurate report of "true disability," further studies might reveal that some combination of self-reported disability status and the collective judgment of a team of expert examiners could bring us even closer to a socially and politically acceptable measure of this concept.

7 Conclusions

This paper provides new insights into the operation of the SSA's disability award process, the large, costly bureaucracy that constitutes the "monitoring technology" and chief "gatekeeper" determining who receives disability benefits. Partly as a result of large backlogs, long delays in providing decisions, unexplained variation in state to state award rates, and the large number of initial denials that are reversed on appeal, the SSA has embarked on a major re-examination of its entire disability determination process to consider how it might be restructured to make it faster, fairer, and more consistent. This is an ambitious undertaking that could provide a better understanding of how the DI bureaucracy should efficiently collect and process information in order to best allocate its available resources to those applicants who are disabled according to the SSA's definition of "being unable to engage in substantial gainful activity." Ideally, this would allow the SSA to perform a cost-benefit calculation to determine whether the improvement in decision-making and resource allocation resulting from the operation of the monitoring technology outweighs the large costs of running it.

Economic theory, in general, has relatively little to say about such problems. In principle, it is possible to run a DI program without employing any monitoring technology: the SSA would simply set a sufficiently low benefit level to deter most non-disabled individuals from applying. In the simplest two-type models, only those who are "truly disabled" will consider applying for benefits (see Diamond and Mirrlees 1978). This saves the expense of running a DI application and appeal bureaucracy, but imposes high costs on "truly disabled" individuals who receive below poverty level benefits due to the informational problems in verifying their disability status. Akerlof (1978) and Parsons (1996) showed that the SSA can achieve more efficient outcomes (higher social welfare) if it has access to a monitoring technology, even if the signals it provides are very noisy. However, these analyses have ignored the costs of running a monitoring technology and have not considered the underlying problem of how best to use information from applicants in reaching accept/reject decisions. There is a wider unresolved question about whether disability is best viewed as a binary outcome, or whether it is better to think of it more along a continuum,

with a sliding scale of benefits depending on the level of "partial disability," such as is currently done in Germany, Spain, and The Netherlands.

Our paper attempts to make some steps towards an empirical framework that could enable us to model these complicated aspects of the disability award process. A first step is to understand how the process really works, and to attempt to measure its accuracy and efficiency. Unfortunately, we are not at the point of being able to provide a framework for evaluating the cost-benefit trade-offs of different ways of structuring the DI award process. We begin with a relatively limited evaluation of the accuracy of the process—an examination of its *classification errors*. These consist of award errors (awarding benefits to a non-disabled applicant) and rejection errors (denying benefits to a disabled applicant). Our analysis is simplified by the SSA's binary definition of disability as the "inability to engage in substantial gainful activity." While the definition seems unambiguous, the actual determination of disability on a case by case basis is a difficult process involving many complicated, often subjective judgments about whether a specific health limitation does in fact prevent an applicant from working altogether.

The only previous academic study of the classification errors in the DI award process was done more than 30 years ago in the seminal study by Nagi (1969). Nagi's investigation relied on independent audits of a set of intercepted DI applicants by teams of medical experts. It offered the closest attempt to providing formal, objective definition of "true disability". Unfortunately, this approach to program evaluation is extremely costly and time consuming, and nobody has attempted to replicate it. In the absence of a better alternative, we proposed a potentially controversial approach to measuring "true disability," namely, we identify self-reported disability as true disability. Specifically, we use the HRS respondents' answer to the question: "Do you have a health limitation that prevents you from working entirely?" ($\tilde{d}$) as an accurate measure of their "true disability" status. This puts us square in the middle of an empirical minefield, since there have been many conflicting empirical studies on the reliability of self-reported health measures. Some claim that such measures are noisy, biased, and endogenous, while others find that they are powerful, exogenous predictors of application, appeal, and labor supply decisions.

In an earlier study (Benítez-Silva et al. 2001) we have gone even further, demonstrating through a number of empirical tests that $\tilde{d}$ is an unbiased and accurate indicator of "true disability." The beauty of this hypothesis is that it can, and indeed does, hold even in the absence of an absolute, objective definition of this concept. Our framework allows us to treat the SSA's interpretation of disability as a slowly evolving social construct that reflects political, economic, and social conditions as much as it reflects the state of medical practices and objectively verifiable measures of health. We introduced the *Rational Unbiased Reporting hypothesis* that DI applicants are fully informed about the rules governing the disability award

process and criteria by which applicants with varying characteristics, are accepted or rejected. Their report of $\tilde{d}$ coincides with the award decision that the SSA would have made if it had the same information as the applicants. We developed a framework for testing this hypothesis and showed that we are unable to reject it at conventional significance levels. In view of this, we have argued that it is reasonable to use self-reported disability $\tilde{d}$ to assess the classification errors in the SSA's disability award process.

Critics might claim that the reason why we fail to reject the RUR hypothesis is that our tests have low power, especially given the relatively few observations of DI applicants in the HRS. However, we showed that when we compared self-reported disability $\tilde{d}$ to a different definition of the SSA's award rate $\tilde{a}$, namely the initial award rate of the DDS instead of the *ultimate award rate* that allows for the possibility that initial rejections can be appealed, we showed that we can decisively reject that hypothesis that $\tilde{d}$ is an unbiased indicator of $\tilde{a}$. Thus,i it seems unlikely that are conslusions are spurious, resulting from a small number of observations and low power tests. Our experience with other data sets suggests that when it is possible to independently verify individuals' survey responses, the answers are surprisingly accurate. Rust and Phelan (1997) showed that the distribution of health care expenditures constructed from self-reported Medicare expenses in the RHS data set closely matched the true distribution constructed for equivalent age/sex groups using the Medicare Statistical System. Hu et al. (1997) compared self-reported health measures to the SSA disability records using a special data set that linked these records for a subset of SIPP participants. Our work will clearly not be the last word on this subject, and we hope it will encourage further theoretical and empirical studies in this important area.

Although there is little more we can say to convince a skeptic that self-reported disability status is a valid measure of "true disability," we conclude by briefly listing some of the insights into the operation of the disability award process that follows from it:

1. There is a substantial amount of noise in the DI award process leading to large rates of award and rejection errors (over 20% and 50%, respectively). The magnitude of these errors is consistent with Nagi's findings, although an entirely different methodology was employed.

2. Much of the screening occurring in the DI award process is achieved by the individuals themselves. We find that there is strong evidence of self-selection in which disabled individuals are substantially more likely than non-disabled individuals to apply for benefits and appeal if denied.

3. It is difficult to estimate the magnitude and the value of the "direct" screening that the DI award process provides. It begins with an applicants' pool that consists of approximately 70% disabled and 30% non- disabled. Of the 75% of these applicants who are ultimately accepted by the SSA,

approximately 77% are disabled and 23% are non-disabled. However, it appears that the substantial delays at various stages of the application and appeal process have strong indirect effects, serving as type-dependent "application fees" that discourage non-disabled individuals from applying for benefits and appealing denials.

4. The U.S. government GAO reports suggest that much of the noise in the disability award process results from reversals by ALJs in the appeal stage, after initial rejections by the DDS. Our results support the opposite conclusion, namely the DDS seem to be too stringent, causing high rates of rejection error through their willingness to err on the side of rejection. The ALJ reversals succeed in significantly reducing the rate of rejection errors (from 64% to 54%) without increasing the rate of award errors.

5. While most of the analysis is carried out under the assumption that the individual self-reported disability status is the true measure of disability, we provide strong evidence that this assumption is in fact valid. When we recomputed the classification errors under the assumption that both the self reported disability $\tilde{d}$ is a noisy but unbiased indicator of true disability status $\tilde{\tau}$, we obtained even higher estimates of the degree of classifcation and Type I and II errors.

6. We compared the performance of the disability award process to an alternative computerized award process based on an index rule involving an estimate of the conditional probability that an individual is disabled given objectively verifiable characteristics x. The computerized rule accepts an applicant for whom $P(\tilde{d} = 1|x)$ is above a certain threshold ρ_c. We set ρ_c so that the computerized screening rule would yield the same award rate as is currently generated by the SSA. We find that when P is estimated using a sample of DI applicants, the computerized screening rule substantially reduces the rate of classification errors. If we were to use the computerized rule to replace the overall DI award process, the computerized rule reduces the award error rate from 26% to 21% and reduces the rejection error rate from 54% to 39%. However if we use the computerized rule to replace only the "first stage" decision by the DDS bureaucracies (but retaining human judges at the appeal level), the gains in accuracy at the first stage level are even more impressive: the computerized rule results in an award error rate of 16% (nearly half of the 29% award error rate of the DDSs), and an rejection error rate of 50% (17% percentage points lower than the rejection error rate of the DDSs).

We believe that this paper's principal contribution is to illustrate a simple method for analyzing classification errors that may prove useful in helping to redesign the current DI award process. The section on

computerized screening rules specifically suggests the possibility that there could be more efficient ways to process disability information. However, we are not suggesting that our computerized rule necessarily dominates the current DI award process in practice. There are a number of practical obstacles to implementing a computerized screening rule similar to the one we describe here. Applicants would have a strong incentive to game the system by attempting to distort their reports of x, so as to maximize their chance of being awarded benefits. To guard against this problem, the SSA could hire teams of medical experts who would be paid to collect accurate measures of x for each applicant. We presume that if these experts were paid by the government, proper incentives can be developed to make the measure x as accurate as possible.

In any event, it seems clear that the DI award process can never be completely computerized. Human decision makers such as expert examiners will always play a key role. In addition to the outright wage costs of hiring these examiners, the designer of any collective decision process has to anticipate that its expert examiners and decision makers may not always act as perfect agents in implementing its preferred policy. This leads to a recursive problem of "monitoring the monitors". The improvements in classification error rates that we have found here represent a best-case scenario and ignore costs associated with practical implementation of a computerized rule. In fact, it would be naive to simply substitute a computerized screening for the current first-stage DDS determination without a more careful modeling of applicants' endogenous reactions to this change. In particular, if computerized screening methods significantly reduced delays involved in applying for benefits, they could encourage a large increase in applications and change the relative mix of disabled and non-disabled applicants. This issue is being addressed in the work we currently are undertaking.

We are currently working on developing an empirical dynamic programming model of the joint decision to work, retire, and apply/appeal for disability benefits. This will allow us to derive individuals' endogenously determined "best replies" to various policy changes including specific aspects of disability process reforms that are currently being considered by the SSA. Although building and solving such models requires substantial work, we think that these more formal approaches will provide additional insights that will be useful in improving the disability determination process in the U.S.

References

Aarts, L.J.M. and P.R. De Jong (1992): Economic Aspects of Disability Behavior Amsterdam, North Holland.

Akerlof, G. A., (1978): "The Economics of 'Tagging' as Applied to Optimal Income Tax, Welfare Programs, and Manpower Planning," *American Economic Review*, 68-1 8–19.

Apfel, K.S. (1999) "Social Security and Supplemental Security Income Disability Programs: Managing for Today, Planning for Tomorrow," unpublished manuscript, U.S. Social Security Administration available online at http://www.ssa.gov/policy/pubs/dibreport.html.

Benítez-Silva, H. (1999): "Micro Determinants of Labor Force Status Among Older Americans," unpublished manuscript, Yale University.

Benítez-Silva, H., M. Buchinsky, H-M Chan, J. Rust, and S. Sheidvasser (1999): "An Empirical Analysis of the Social Security Disability Application, Appeal and Award Process," *Labour Economics*, 6 147-178.

Benítez-Silva, H., M. Buchinsky, H-M. Chan, J. Rust, and S. Sheidvasser (2003), "How Large is the Bias in Self-Reported Disability Status?" forthcoming, *Journal of Applied Econometrics.*

Bound, J., R. Burkhauser (1999): "Economic Analysis of Transfer Programs Targeted on People with Disabilities," forthcoming in O. Ashenfelter and D. Card (eds.), Handbook of Labor Economics, Amsterdam, North Holland.

Bound, J. and T. Waidmann (1992) "Disability Transfers, Self-Reported Health, and the Labor Force Attachment of Older Men: Evidence from the Historical Record," *Quarterly Journal of Economics* **107-4** 1393–1419.

Diamond, P. and J. Mirrless (1978): "A Model of Social Insurance with Variable Retirement," *Journal of Public Economics*, 10 295–336.

Hu, J., K. Lahiri, D.R. Vaughan, and B. Wixon (1997): "A Structural Model of Social Security's Disability Determination Process," ORES Working Paper No. 72, Office of Research and Evaluation Statistics, Social Security Administration, 500 E Street SW, Washington, D.C.

Johnson, W. G. (1977): "The Effect of Disability on Labor Supply: Comments," *Industrial and Labor Relations Review*, 30 380-381.

Lahiri, K., D.R. Vaughan, and B. Wixon (1995): "Modeling SSA's Sequential Disability Determination Process Using Matched SIPP Data," *Social Security Bulletin*, 58-4 3–42.

Muller, L. S. (1992) "Disability Beneficiaries Who Work and Their Experience Under Program Work Incentives" *Social Security Bulletin* **55-2** 2–19.

Nagi, S.Z. (1969): Disability and Rehabilitation: Legal, Clinical, and Self-Concepts and Measurement, Ohio State University Press.

Parsons, D.O. (1991b): "Measuring and Deciding Disability," in C.L. Weaver (ed.), Disability and Work: Incentives, Rights, and Opportunities, American Enterprise Institute, Washington, D.C.

Parsons, D.O. (1996): "Imperfect 'Tagging' in Social Insurance Programs," *Journal of Public Economics*, 62 183–207.

Smith, R.T. and A.M. Lilienfeld (1971): "The Social Security Disability Program: An Evaluation Study," Research Report 39, Social Security Office of Research and Statistics.

Social Security Advisory Board (1998): "How SSA's Disability Programs Can be Improved," Report 6, Social Security Advisory Board, available on the web at http://www.ssab.gov/ Report6.html.

Stapleton, D., B. Barnow, K. Coleman, K. Dietrich, and G. Lo (1994): Labor Markets Conditions, Socioeconomic Factors and the Growth of Applications and Awards for SSDI and SSDI Disability Benefits: Final Report, Lewin-VHI, Inc. and the Department of Health and Human Services, The Office of the Assistant Secretary for Planning and Evaluation.

U.S. Department of Health and Human Services (1988): Social Security Handbook, Tenth Edition.

U.S. General Accounting Office (1997): "Social Security Disability: SSA Actions to Reduce Backlogs and Achieve More Consistent Actions Deserve High Priority," GAO/T-HEHS-97-118.