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*[*]

December, 2005

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 from the Disability Insurance (DI) and Supplemental Security Income (SSI) programs. We study a subset of individuals from the Health and Retirement Study (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 appeals) to the applicant's self-reported disability status $\tilde{d}$. We use these data to estimate classification error rates under the hypothesis that applicants' self-reported disability status $\tilde{d}$ and the SSA's ultimate award decision $\tilde{a}$ are noisy but unbiased indicators of $\tilde{\tau}$, a latent "true disability status" indicator. We find that approximately 20% of SSI/DI applicants who are ultimately awarded benefits are not disabled, and that 60% of applicants who were denied benefits are disabled. Our analysis also yields insights into the patterns of self-selection induced by varying delays and award probabilities at various levels of the application and appeal process. We construct an optimal statistical 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.

Keywords: Social Security Disability Insurance, Supplemental Security Income, Health and Retirement Study, Classification Errors.
JEL classification: H5

[*] Corresponding author: University of Maryland, Department of Economics, 4115 Tydings Hall, College Park, Maryland 20742. `jrust@gemini.econ.umd.edu`. This work was supported by NIH grant AG12985-02. Benítez-Silva is also grateful for the financial support of the "la Caixa Fellowship Program" in the early stages of this research. Buchinsky is grateful for the support from the Alfred P. Sloan Research Fellowship. 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, the Maryland Population Research Center seminar series, the SUNY-Stony Brook Labor and Health Workshop, and the North American Meetings of the Econometric Society in Evanston. 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 1974, 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 the SSA's "Blue Book", that are regarded 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 indeed 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.

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 conclu-

[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. 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).

sion 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 under 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.

Initial disability determinations, made by state-level bureaucracies known as “Disability Determination Services” (DDS), 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, in 2000, DDS award rates for DI applicants 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 Study (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 closest HRS interview to the time of application).[5] 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 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

[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”.

[5] The only exception is for people who applied right after an interview at which they had reported not being disabled. In that case we assigned the data provided at the subsequent interview, even if this was relatively long after the application date.

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.[6]

We estimate the 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 22% and the rejection error to be 58%.[7]

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 section 4), 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 variables $\tilde{\tau}$, $\tilde{d}$, and $\tilde{a}$ form a trivariate probit system with a correlation structure that matches 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. Our Bayes estimates of the award error rate is 21.7%, and the rejection error rate is 59.9%, which hardly differ from the award and rejection errors, 21.9% and 58.6%

[6] 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. 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. When the self-reported disability is taken from the wave of the HRS *after* the person reported applying for disability benefits it is possible that they could have recovered from their disability since the date they applied. While the timing difference and the differences in definition can be a source of some discrepancies between $\tilde{a}$ and $\tilde{d}$, we believe they have only "second-order" effects and cannot be responsible for the very large discrepancies that we find in this analysis.

[7] 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., $\Pr\{\tilde{a} = 0|\tilde{\tau} = 1\}$, while Type II error rate is the probability of not rejecting H_0 when it is false, i.e., $\Pr\{\tilde{a} = 1|\tilde{\tau} = 0\}$. Our point estimates of the Type I and II error rates of the SSA award process are 22% and 62%, respectively.

respectively, obtained in the case where we assume that $\tilde{d} = \tilde{\tau}$ with probability 1.

Our Bayes estimates are based on an "equicorrelation assumption" that implies that $\tilde{a}$ and $\tilde{d}$ are equally accurate signals of true disability status $\tilde{\tau}$. However, in order to assess the robustness of our conclusions we also compute classification error rates under the assumption that unobservable factors affect the Social Security award decision $\tilde{a}$ are much more highly correlated with the unobservable factors affecting true disability status $\tilde{\tau}$, than are the unobservable factors affecting individuals' self-reported disability status $\tilde{d}$. This implies that the Social Security award decision $\tilde{a}$ is a significantly more accurate signal of true disability than is an individual's self-report $\tilde{d}$. However, even in this case the classification error rates are substantial, and in our "best case" (for SSA) results, the award error rate is 16% and the rejection error rate is 52%.

We propose an alternative disability determination procedure based on an optimal statistical screening rule that awards disability benefits when the predicted probability that an applicant is truly disabled is sufficiently high. We show that this statistical screening rule results in significantly fewer classification errors than the multi-stage system currently used by the SSA. This suggests that there may be cheaper, faster, and more accurate ways of making disability determinations than the SSA's current award process. The SSA has considered similar types of statistical screening rules as part of a "disability process redesign" initiative introduced under the Clinton administration, but its current status is unclear. The SSA has recently implemented a more limited set of changes in the award process that may facilitate the transition to statistical screening rules as a replacement for initial determination decisions by the state level DDSs.

Section 2 reviews estimates of classification error rates from previous studies, including SSA's own internal audits of its disability award process. Section 3 describes the DI and SSI programs and the disability award process. Section 4 provides empirical evidence that survey respondents who are given strong, credible assurances of confidentiality provide truthful, unbiased, and "accurate" reports of their disability status. The accuracy of self-reports are critical to our ability to evaluate the classification errors of the disability award process. Section 5 describes the HRS data used in this analysis and demonstrates that the self-reported disability $\tilde{d}$ 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 6 presents our Bayesian analysis of the classification errors in the disability award process. Section 7 analyzes the sources of these classification errors at different stages of the disability award and appeal process. Section 8 describes a computerized disability screening procedure that outperforms the multi-stage system currently used at SSA. Section 9 summarizes various disability redesign initiatives that SSA has considered over the past decade, including the use of statistical screening rules similar to the one proposed in section 8. Section 10 offers some conclusions and qualifications of our research findings, and suggestions for further research.

2 Previous Estimates of SSDI Classification Error Rates

Our estimated classification error rates are higher than those estimated in previous internal and external audits of the SSA's disability award process. Some of these studies relied on decisions of independent "experts" who attempted to directly measure true disability status $\tilde{\tau}$. Smith and Lilienfeld (1971) reported the results of an internal audit of DI awards done by the SSA's own 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. However, the objectivity of SSA's "self-audits" may be open to question.

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 almost three times 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.[8] 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 decisions by the ALJs. Nevertheless, this report provides no explicit information on whether the appeal option increases or reduces the award and rejection errors.

[8] 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).

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. The programs provided benefits to nearly 12 million individuals in 2001, at a cost of $55 billion for DI, and $32 billion for SSI.[9] 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.

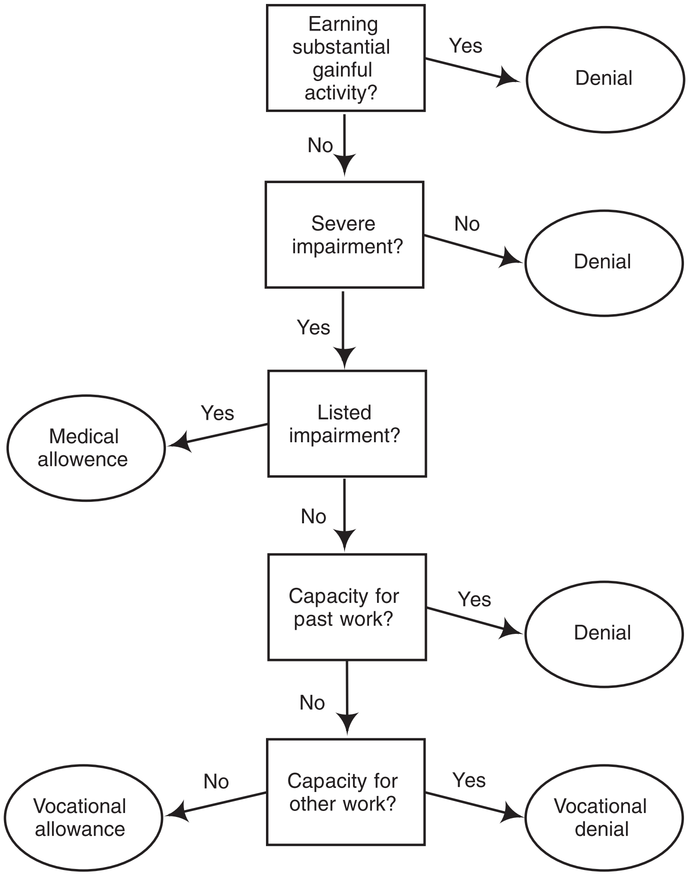
The volume of applications combined with the complexities involved in screening and adjudicating applications and appeals makes administration of these programs expensive and time consuming. For example, in 1998, the SSA processed more than two million applications for DI and SSI benefits, and over 500,000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

The process by which SSA makes disability determinations 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 an initial or "first-stage" award 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

[9] 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.

psychological problems is less clear-cut. In 2001, the mean time for an initial decision by the DDS was about 90 days. Each of the five stages is handled by separate specialists, and this is the reason why a large number of different people are typically involved in the DDS's award decision for a single applicant.

Figure 1: Five Stage DDS Disability Determination Procedure



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.[10] 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.

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 determinations. 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.

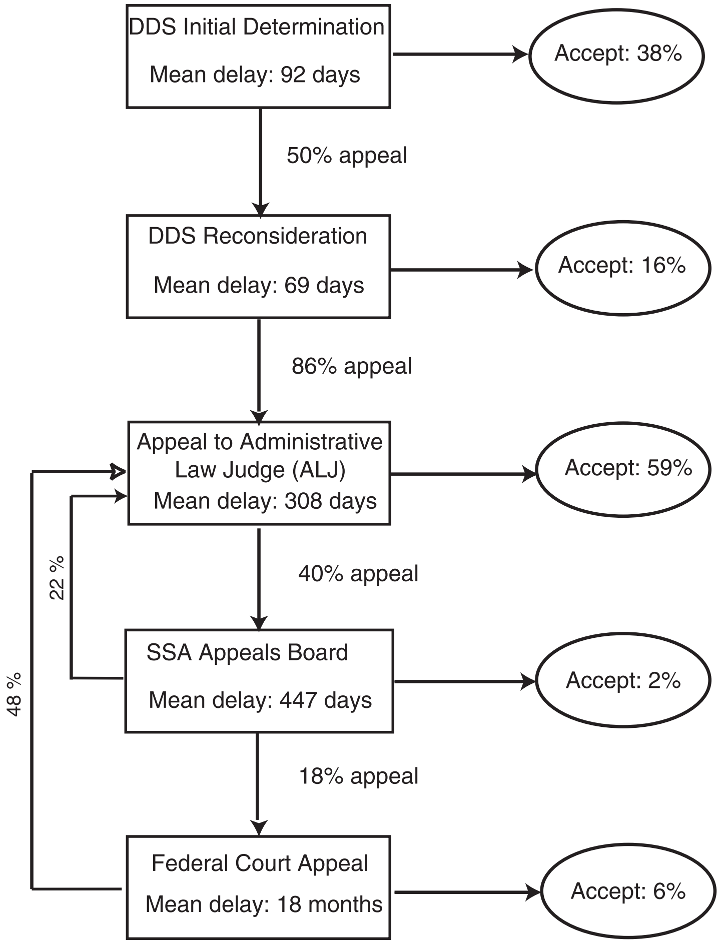
The other Federal program providing disability benefits is the SSI, a means-tested cash assistance program enacted in 1974. Unlike the DI, there is no work requirement for SSI benefits. However, SSI applications are evaluated according to the same process as DI benefits and satisfy the same basic definition 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.[11] As a result of different eligibility requirements,

[10] 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.

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

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.

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



4 Do People Truthfully Report Their Disability Status?

Our analysis of classification errors in the SSA's award process depends critically on the use of self-reported disability status. The credibility of our conclusions depend on the assumption that individuals' self-reports are truthful and accurate. This section provides evidence supporting this hypothesis.

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 the 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.[12]

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% 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 confidentiality 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 rolls.[13]

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,

[12] 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 the 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.

[13] 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 rolls 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 rolls 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 6 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 rolls. For this reason there appears to be a clear incentive for individuals with low wage prospects to remain on the rolls even if they are capable of earning in excess of the SSA's SGA threshold.

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, we have shown in Benítez-Silva, Buchinsky, Chan, Cheidvasser, and Rust (2003), (hereafter BBCCR), 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).[14] We interpret this “unbiased reporting” result as providing strong empirical support for the “accurate reporting hypothesis”. If the HRS respondents had a significantly 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 BBCCR 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 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 assumption that ε_a and ε_d are normally distributed random variables, and given identifying normalizations on the variances of ε_a and ε_d and/or the coefficient vectors, 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 $\tilde{a}$

[14] 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).

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 BBCCR 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 independently of the particular regime and overall degree of leniency in the SSA's award process. Additional evidence supporting this hypothesis is provided by Bound and Waidmann (1992), who compared time series data on self-reported disability to DI awards and rolls. 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 the 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 the

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 make 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, as indicated above, 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 $\tilde{\tau} = I(x\beta_\tau + \varepsilon_\tau > 0)$.

Hereafter 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 $\beta \equiv \beta_a = \beta_d = \beta_\tau/\sigma_\tau$, where σ_τ is the standard deviation of unobservable variables that affect true disability status $\tilde{\tau}$. In this case both $\tilde{a}$ and $\tilde{d}$ are noisy but unbiased indicators of the 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 assume that ε_τ, ε_a and ε_d have a multivariate normal distribution and we will impose restrictions on the correlations between these error terms that will enable us to compute classification error rates using Bayes rule.

5 Measuring Disability and Health Status in the HRS

This section provides a brief description of the Health and Retirement Study (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, $\tilde{d}$, is a very powerful predictor of application, appeal, and award decisions, and better predicts 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 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 8 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 being disabled, $P\{\tilde{d} = 1|X\}$, in terms of a large number of objectively verifiable health conditions and activities of daily living, X.

5.1 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.[15] 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 somewhat 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.[16] Finally, the HRS did not 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.[17]

Individual decisions as to when to apply or appeal for disability benefits are made in continuous time.[18] 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.[19]

[15] 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.

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

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

[18] 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.

[19] We have experimented with windows of different length, and although this affects the number of observations, it does not

5.2 Data Analysis

Benítez-Silva et al. (1999) and BBCCR conducted an internal validation of the quality of our "constructed" disability histories and the accuracy of individuals' responses to the HRS questions. In those papers 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%. The conclusions are robust to screening out the 52% of the sample for which imputations on the dates of disability onset, application, or award were made.

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 in BBCCR for testing the accuracy of self-reported disability status, and here 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 physical health and economic status. 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 versus 982 hours for applicants. Also, DI applicants are significantly more likely to be female and non-white, 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 over the 1992-1996 sample period.

significantly alter our results.

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 higher 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 sub-samples given in the columns of Table 3.[20] In all cases we clearly 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

[20] 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, pick up a dime from a table, taking a bath, sitting for a long time, getting up from a chair, walking across a room, and climbing a flight of stairs.

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 fact that 59% of the disabled non-applicants are ineligible for SSI, whereas only 20% of the applicants are ineligible.[21]

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, which 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.

[21] See Benítez-Silva et al. (1999) 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 60.9, 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. The latter effect might seem rather puzzling, after all, there is nothing preventing individuals from applying for disability at any age before the NRA, especially if access to SSDI will guarantee a higher lifetime benefit than retiring under the Old Age benefits system before the NRA. However, we have to take into account the costs imposed on applicants in the form of not being able to earn above the SGA threshold, having to hire a lawyer if they need to appeal their cases beyond the DDS stage, and also the reduction in the level of disability benefits resulting from having received retirement benefits up to the time of starting to receive SSDI benefits.

Table 3: Characteristics of DI Applicants and Non-Applicants

			SSI/DI Applicants		SSI/DI Non-Applicants	
Variable	SSI/DI	SSI/DI				
	Applicants	Non-Applicants	Non-Disabled	Disabled	Disabled	Non-Disabled
	(1)	(2)	(3)	(4)	(5)	(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	70.5	3.5	0.0	100.0	100.0	0.0
Prevents Work	(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 (1.0)	0.8 (0.1)	10.6 (1.6)	16.5 (1.3)	11.9 (1.0)	0.4 (0.1)
Sit for long time?	45.7 (1.4)	13.7 (0.2)	36.2 (2.6)	49.8 (1.7)	45.0 (1.6)	12.6 (0.2)
Get up from a chair?	58.6 (1.4)	20.6 (0.2)	47.4 (2.7)	63.2 (1.7)	59.2 (1.6)	19.2 (0.2)
Climb Stairs?	46.7 (1.4)	5.5 (0.1)	35.6 (2.6)	51.4 (1.7)	38.5 (1.6)	4.3 (0.1)
Take a Bath?	15.1 (1.0)	0.8 (0.1)	9.2 (1.5)	17.6 (1.3)	11.6 (1.0)	0.4 (0.1)
Reading a Map?	32.6 (1.3)	14.6 (0.2)	33.5 (2.6	32.2 (1.6)	33.1 (1.6)	13.9 (0.2)
Pick up a Dime?	15.8 (1.1)	2.3 (0.1)	13.0 (1.8	16.9 (1.3)	13.3 (1.1)	1.9 (0.1)
Have you ever had:						
Cancer?	4.8 (0.73)	0.8 (0.10)	4.3 (1.29)	5.0 (0.89)	1.78 (0.65)	0.8 (0.10)
Lung Disease?	18.2 (1.1)	4.6 (0.1)	16.4 (2.0)	19.0 (1.4)	15.1 (1.2)	4.2 (0.1)
Stroke?	7.4 (0.8)	0.4 (0.04)	3.4 (1.0)	9.0 (1.0)	4.7 (0.7)	0.3 (0.03)
Heart Problems?	23.8 (1.2)	6.7 (0.1)	18.1 (2.0)	26.1 (1.5)	22.1 (1.3)	6.1 (0.1)
Psych. Problems?	25.3 (1.3)	6.9 (0.1)	21.0 (2.2)	27.1 (1.5)	20.4 (1.3)	6.4 (0.1)
Nursing home stay?	3.6 (0.6)	0.2 (0.03)	1.7 (0.7)	4.4 (0.7)	1.6 (0.4)	0.1 (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

Table 4: Characteristics of Subset of DI Applicants

Variable	DI Awardees	DI Rejectees	Rejectees		Awardees	
			Non-Disabled	Disabled	Disabled	Non-Disabled
	(1)	(2)	(3)	(4)	(5)	(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	78.1	58.6	0.00	100.00	100.00	0.00
Prevents Work	(2.5)	(4.8)	(0.00)	(0.00)	(0.00)	(0.00)

Table 4: (Continued)

Variable	DI	DI	Rejectees		Awardees	
	Awardees (1)	Rejectees (2)	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?	35.9	36.9	32.6	40.0	35.2	36.1
	(2.9)	(4.7)	(7.1)	(6.3)	3.2	6.1
Pick up a Dime?	14.2	20.6	16.3	23.7	15.4	9.9
	(2.1)	(4.0)	(5.6)	(5.5)	2.4	3.8
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

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 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 8, 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. This will enable us to formalize a sense in which the SSA's current disability award process is "informationally inefficient."

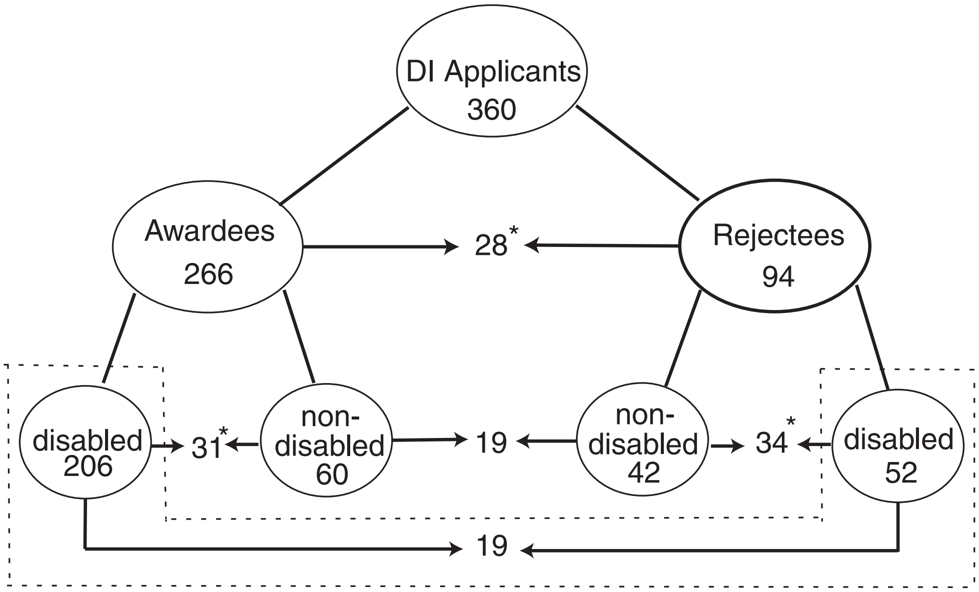


Figure 3: Summary of Classification Errors in the DI Award Process

6 Bayesian Estimates of Classification Error Rates

The previous section presented strong empirical support for the hypothesis that the Social Security Award decision $\tilde{a}$ and self-reported disability status $\tilde{d}$ are unbiased accurate indicators of "true disability" status. Indeed, Figure 3 from the preceding section suggests that self-reported disability $\tilde{a}$ is a relatively *more accurate* indicator of disability than the SSA's award decision $\tilde{a}$, insofar as self-reported disability does a better job of differentiating individuals according to the severity of their disability as measured by a vector X of more objectively verifiable health problems and activities of daily living. In this section we show how this information can be used to measure the magnitude of classification errors in the DI award process. To give every benefit of a doubt to the SSA, we will estimate classification error rates under the assumption that the SSA's ultimate award decision $\tilde{a}$ is a more accurate indicator of "true disability" $\tilde{\tau}$ (i.e. $\tilde{a}$ is more highly correlated with τ) than self-reported disability $\tilde{d}$.

As we discussed in the introduction, our analysis of classification errors avoids any judgement as to whether the SSA's ultimate award decision $\tilde{a}$ is too strict or lenient relative to some objective, absolute definition of disability. Instead we assume that the SSA's administration of the DI award process sets a *social standard* for disability that becomes common knowledge for all individuals applying for disability benefits. This social standard can, and most likely does, change over time. The SSA implements its definition of disability via its award decisions, and disability applicants are aware of the standard in effect at any given time and adapt their self-reported disability relative to this social standard. The (RUR) hypothesis, which we are unable to reject econometrically, states that except for random "noise" representing differences in information, random "mistakes" and idiosyncratic judgement errors, individuals and the SSA use the same decision rule to determine $\tilde{d}$ and $\tilde{a}$. Econometrically the RUR hypothesis can be stated in terms of the following *conditional moment* (CM) restriction:

$$E\left[\tilde{a}-\tilde{d}\middle|x\right]=0, \tag{1}$$

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, equation (1) is equivalent to $\Pr(\tilde{a}|x)=\Pr(\tilde{d}|x)$. In BBCCR (2003) we tested the CM restriction using a battery of different CM tests, but were unable to reject the null hypothesis that (1) holds.

If we further assume that $\tilde{a}$ and $\tilde{d}$ can be represented as *index rules* with unobservable (to the econometrician) factors in these rules entering additively separably as a pair of variables $(\varepsilon_a,\varepsilon_d)$ that have a bivariate normal distribution, then the conditional probabilities governing $\tilde{a}$ and $\tilde{d}$ are represented by the following bivariate probit model

$$\begin{aligned}\Pr(\tilde{a}|x) &= E\left[I\left(x'\beta_a+\varepsilon_a\geq 0\right)\right]=\Phi(-x'\beta_a) \quad \text{and}\\ \Pr(\tilde{d}|x) &= E\left[I\left(x'\beta_d+\varepsilon_d\geq 0\right)\right]=\Phi(-x'\beta_d),\end{aligned} \tag{2}$$

where Φ is the standard normal cumulative distribution function. For this parametric model, the RUR hypothesis implies the restriction that $\beta_a=\beta_d$. As is commonly done in the literature on discrete choice models, we normalize the variances of $(\varepsilon_a,\varepsilon_d)$ to 1 but allow them to have an unrestricted correlation coefficient $\lambda\in(-1,1)$. BBCCR was unable to reject this parametric form of the RUR hypothesis at conventional significance levels.

Now suppose there is a third binary indicator, representing the individual's *true disability status* $\tilde{\tau}$, which is not observed by either the SSA or the individuals. We assume that $\tilde{\tau}$ can also be represented by an index rule

$$\tilde{\tau} = I\left(x'\beta_\tau + \varepsilon_\tau \geq 0\right). \tag{3}$$

The quantities $\tilde{a}$ and $\tilde{d}$ can be considered as noisy indicators of true disability $\tilde{\tau}$.

Since true disability $\tilde{\tau}$ is unobserved, we need to impose assumptions about the index coefficients β_τ and the distribution of the unobserved error terms $(\varepsilon_a, \varepsilon_d, \varepsilon_\tau)$ entering the index representations for (a, d, τ) in order to compute classification errors. The RUR hypothesis implies the restriction that $\beta_a = \beta_d$. We now impose additional assumptions that will enable us to compute classification error rates.

Assumption 1: (Joint Normality) *The error terms $(\varepsilon_a, \varepsilon_d, \varepsilon_\tau)$ are independent of the vector of observed characteristics X and have a joint trivariate normal distribution, with variances normalized to* 1.[22]

Assumption 2: (Unbiasedness) *The indicators $\tilde{a}$ and $\tilde{d}$ are conditionally unbiased indicators of true disability status $\tilde{\tau}$. Given Assumption 1, this implies that*

$$\beta_\tau = \beta_a = \beta_d. \tag{4}$$

Assumption 3: (Projection of $(\varepsilon_a, \varepsilon_d)$ on ε_τ) *The error terms ε_a and ε_d can be represented as projections on ε_τ as*

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

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

where ν_a and ν_d are independent of ε_τ and each other.

Assumption 4: (Equicorrelation of ε_a and ε_d with ε_τ) *The correlation of ε_a with ε_τ is the same as the correlation of ε_d with ε_τ, i.e.*

$$\rho_a = \rho_d. \tag{7}$$

[22] This normalization, $\sigma_a^2 = \sigma_d^2 = 1$, is the most commonly used in the literature, and allows us to estimate β_a and β_d as described in BBCCR. Alternatively, one can set $\sigma_a^2 = 1$ and set the two constant coefficients in β_a and β_d to be the same. The main difference between these two normalizations is that the first model assumes that the variances of the two latent indices for the SSA and the individuals are the same, and allow the location of the indices, i.e., the constant coefficients, to differ. In contrast, the second model assumes that the locations for the SSA and the individuals are the same, but that the SSA and the individuals may have different variances for the error terms ε_a and ε_d, respectively, reflecting their differences in the subjective knowledge about the individuals' disability. Notice that in both cases σ_τ^2 is not identified, and one can find reasons to believe it can be larger or smaller than the variances of the two other errors. We present results that assume various values for this element. In this version of the paper we only report results under the first normalization given that our estimates of σ_d, when using the second normalization, where not statistically different from 1 at any traditional level of significance. The Likelihood Ratio statistic of the test of the model with equal variances against one that allows them to be different is 1.3279, which follows a χ_1^2. The p-value is equal to 0.249, so we cannot reject the hypothesis that both variances are equal to 1.

Let λ denote the correlation of ε_a and ε_d. Since $\tilde{a}$ and $\tilde{d}$ are observed, we can estimate λ via a bivariate probit model for $(\tilde{a},\tilde{d})$ as in BBCCR. Assumption 3 implies that $\lambda = \rho_a \rho_d$, so the equicorrelation assumption implies that

$$\rho_a = \rho_d = \sqrt{\lambda}. \tag{8}$$

With these assumptions it follows that

$$\varepsilon = (\varepsilon_\tau, \varepsilon_a, \varepsilon_d) \sim \mathrm{N}(0, \Sigma_\varepsilon), \tag{9}$$

where

$$\Sigma_\varepsilon = \begin{pmatrix} 1 & \sqrt{\lambda} & \sqrt{\lambda} \\ \sqrt{\lambda} & 1 & \lambda \\ \sqrt{\lambda} & \lambda & 1 \end{pmatrix}.$$

The information in (2)-(9) is sufficient for computing the Bayes classification errors for this case. We believe assumptions 1-3 are relatively innocuous. The key assumption underlying our estimates of the classification errors is the equicorrelation assumption. Intuitively, this assumption means that both $\tilde{a}$ and $\tilde{d}$ are equally accurate indicators of $\tilde{\tau}$. This seems to be a reasonable "compromise" between the extremes of assuming that $\tilde{\tau} = \tilde{d}$ with probability 1 (i.e. that individual self-reports are always "right") and $\tilde{\tau} = \tilde{a}$ with probability 1 (i.e. that the SSA's award decision is always right).

We also calculate classification errors under an alternative assumption that implies that the SSA's award decision is a substantially more accurate indicator of true disability $\tilde{\tau}$ than an individual's self-report $\tilde{d}$. To do this we relax the equicorrelation structure (Assumption 4) and the structure imposed by Assumption 3, and assume instead that the observables terms $(\varepsilon_a, \varepsilon_d, \varepsilon_\tau)$ have an unrestricted covariance matrix, apart from the normalization that the variances of ε_a and ε_d are equal to 1. That is, we assume

$$\varepsilon = (\varepsilon_\tau, \varepsilon_a, \varepsilon_d) \sim \mathrm{N}(0, \Sigma_\varepsilon), \tag{10}$$

where the covariance matrix Σ_ε is given by

$$\Sigma_\varepsilon = \begin{pmatrix} \sigma_\tau^2 & \sigma_\tau \rho_{a\tau} & \sigma_\tau \rho_{d\tau} \\ \sigma_\tau \rho_{a\tau} & 1 & \lambda \\ \sigma_\tau \rho_{d\tau} & \lambda & 1 \end{pmatrix},$$

where we maintain our normalization that $\sigma_a = \sigma_d = 1$ but allow σ_τ to be bigger or smaller than 1 and allow the correlation coefficients $\rho_{a\tau}$ and $\rho_{d\tau}$ to be unrestricted.

It is not possible to econometrically identify these additional parameters under this weaker assumption, but we can vary σ_τ, $\rho_{a\tau}$, and $\rho_{d\tau}$ from the values they take under our equicorrelated "base case", since we can identify all the parameters of the covariance matrix when we impose the equicorrelation assumption. So to test the robustness of our conclusions, we recompute the classification errors for two alternative cases where we vary σ_τ, $\rho_{a\tau}$ and $\rho_{d\tau}$ away from the values in the base case, and in a direction that is "favorable" to the SSA, i.e. where $\rho_{a\tau}$ is significantly larger than $\rho_{d\tau}$ which implies that $\tilde{a}$ is a significantly more accurate indicator of $\tilde{\tau}$ than $\tilde{d}$.

We are now ready 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 computations of the other classification errors have similar derivations. The award error can be written as

$$\Pr(\widetilde{\tau}=0|\widetilde{a}=1)=\int \Pr(\widetilde{\tau}=0|\widetilde{a}=1,x)\, f_x(x)dx, \tag{11}$$

where $f_x(x)$ is the density of the observed characteristics. Note that the probability inside the integral in (11) can also be written as,

$$\begin{aligned}\Pr(\widetilde{\tau}=0|\widetilde{a}=1,x) &= \Pr\left(\widetilde{d}=0\right)\Pr\left(\widetilde{\tau}=0|\widetilde{a}=1,\widetilde{d}=0,x\right)\\ &\quad+\Pr\left(\widetilde{d}=1\right)\Pr\left(\widetilde{\tau}=0|\widetilde{a}=1,\widetilde{d}=1,x\right).\end{aligned} \tag{12}$$

Further note that

$$\Pr\left(\widetilde{\tau}=0|\widetilde{a}=1,\widetilde{d}=0,x\right)=\frac{\Pr\left(\widetilde{\tau}=0,\widetilde{a}=1,\widetilde{d}=0|x\right)}{\Pr\left(\widetilde{a}=1,\widetilde{d}=0|x\right)}, \tag{13}$$

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 (13) can be easily computed, given the distribution of ε in (9) and (10), using, for example, the GHK algorithm and the coefficient estimate for β from BBCCR. In the example of the probability in the numerator of (13),

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

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

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

Table 5 presents the Bayes estimates of the classification errors implied by Assumptions 1-2. The first row shows the classification errors in the base case, where we also impose Assumptions 3 and 4. Therefore ε_a and ε_d are equicorrelated with ε_τ. The common correlation parameter ρ equals the square root of λ, the correlation between ε_a and ε_d. In this case $\lambda = .12$ and $\rho = \sqrt{\lambda} = .34$. The award error rate is estimated to be 21.7%, and the rejection error rate is estimated to be 59.9%.

Table 5: Bayes Classification Errors using Alternative Models

Model	Error Type			
	Award	Rejection	Type I	Type II
a . Equicorrelation case: $\rho_{ad} = .12$; $\rho_{a\tau} = \rho_{d\tau} = \sqrt{\rho_{ad}}$				
$\sigma_\tau = 1$	21.71%	59.94%	23.71%	67.67%
b . Asymmetric correlation case with $\rho_{ad} = .12$; $\rho_{a\tau} = .4$; $\rho_{d\tau} = .2$				
$\sigma_\tau = .9$	19.34%	57.49%	21.44%	55.05%
c . Asymmetric correlation case with $\rho_{ad} = .12$; $\rho_{a\tau} = .5$; $\rho_{d\tau} = .1$				
$\sigma_\tau = .8$	16.10%	52.03%	19.27%	41.60%

Panels (b) and (c) show the effect of relaxing the equicorrelation assumption (and the structure of Assumption 3) in a direction that is favorable to the SSA. Panel (b) shows the results in the case where $\rho_{d\tau}$ is lowered to .2 and $\rho_{a\tau}$ is increased to .4 and σ_τ is decreased to .9. This is intended to reflect a situation where both ε_a and ε_d are "noisier" than ε_τ (reflecting the effect of bureaucratic errors in $\tilde{a}$ and idiosyncratic judgement errors in individuals' self-reports $\tilde{d}$), and the correlation of ε_a with ε_τ is assumed to be twice as high as the correlation between ε_d and ε_τ to reflect the assumption that the SSA's award decision is a more accurate indicator of true disability than the individual's self-report. The classification error rates decrease in this case, but they are still quite high: the award and rejection error rates fall by only 2 percentage points to 19.3% and 57.5% percent, respectively.

The final panel shows the classification error rates in an even more extreme case where we have reduced the variance of σ_τ to only 0.8, a full 20% reduction in the variance of ε_τ relative to the unit normalized variances for ε_a and ε_d in the base case, and the correlation between ε_a and ε_τ is increased to .5, which is 5 times higher than the correlation between ε_d and ε_τ. The award error rate falls by a little over 5 percentage points in this case, to 16.1%, and the rejection error rate falls by nearly 8 percentage points, to 52%.

Thus, we conclude that a) award and rejection error rates are high, and b) rejection error rates are more than twice as high as award error rates, and are robust to fairly big changes in the covariance matrix for $(\varepsilon_a, \varepsilon_d, \varepsilon_\tau)$ in directions that are favorable to SSA (i.e. which tend to reduce classification errors). Interestingly, finding b) is consistent with the Nagi study discussed in section 2: that study also found that rejection error rate was more than twice as high as the award error rate. The 19% award error rate in the Nagi study is consistent with panel b of Table 5, although the rejection error rates in Table 5 are uniformly higher than the 48% rejection error rate found in Nagi's study.

Note that the Bayes estimates of classification error rates are quite close to the classification error rates that result from the assumption that $\tilde{d} = \tilde{\tau}$ with probability 1. This suggests that self-reported disability $\tilde{d}$ is a very accurate measure of true disability status $\tilde{\tau}$, and is consistent with the results of the previous section that suggest that self-reported disability is actually a more accurate indicator of true disability status than the Social Security's award decision.

7 Analysis of the sources of errors in the disability award process

This section attempts to provide more insight into the source of the high classification errors in the SSA's disability award process. We analyze the multi-stage award and appeal process that was summarized in section 5, assigning classification errors to each stage of the process. In order to carry out this analysis, we will need to invoke the assumption that $\tilde{d} = \tilde{\tau}$ with probability 1. As we noted in the previous section, our estimates of classification error rates are not substantially affected by this assumption relative to the assumption that both $\tilde{a}$ and $\tilde{d}$ are noisy indicators of true disability status $\tilde{\tau}$.

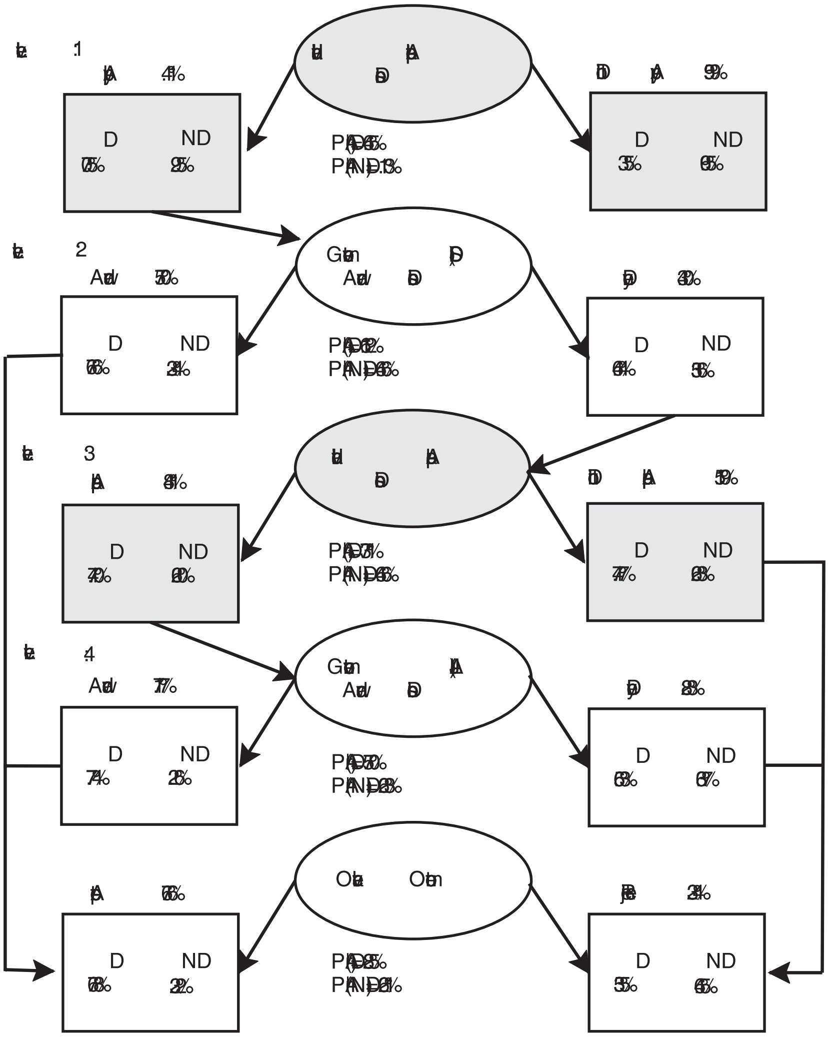
Figure 4 presents a simplified schematic diagram of the SSA's complete disability award process, with estimated classification error rates presented at each stage of the process. The first level of the figure represents the individual's decision whether or not to apply for benefits. Over the period of analysis, 4.1% of individuals in the HRS applied for DI and the remaining 95.9% did not apply. The final node presents the outcome of the overall award process, including all possible appeals. The ultimate award rate is 76.6%, with an award error rate of 23.2% and a rejection error rate of 53.5%; that is, $\Pr(\tilde{a} = 1) = .766$, $\Pr(\tilde{d} = 0|\tilde{a} = 1) = .232$, and $\Pr(\tilde{d} = 1|\tilde{a} = 0) = .535$.[23]

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

[23] Notice that these numbers are slightly different from the ones in Table 4 due to the fact that in this figure we are not restricting the sample of applications to be in a one year window around an interview date.

non-disabled applicants. In particular, a disabled applicant has an 82.5% probability of ultimately being awarded benefits compared to the much lower award rate of 62.1% for non-disabled applicants. Still, 62.1% is a surprisingly high award rate for non-disabled applicants, suggesting that there could be relatively high returns for "imposters" to apply for DI benefits. This may be the reason was why 29.5% of DI applicants report that they are not disabled (see Level 1 of Figure 4).

Figure 4: Analysis of Classification Errors in SSA's Disability Award 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 4. For our HRS sample, the first-stage award rate is only 57%, far smaller than the 76.6% 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 incur a significantly higher rate of rejection errors, 64.4% vs. 57.7%, respectively, suggesting that the DDS centers are erring on the side of rejecting applicants. This might be a reasonable strategy, since a rejected applicant has the option to appeal, whereas DI awardees are unlikely to leave the rolls 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 about two-thirds of rejected applicants choose to appeal, disabled candidates are much more likely to appeal than non-disabled candidates (73.1% vs. 46.6%). 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 (74% vs. 70%). Of the 52% of the initially rejected applicants who chose not to appeal, 62% are non-disabled.

The fourth level of Figure 4 represents the ALJ decision. The award rate at this stage, 71.7%, is substantially higher than at the DDS level. Despite this, the ALJ award error rate is slightly lower than that incurred by the DDSs (22.6% vs. 23.4%). 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, 65.3%, is actually slightly *higher* than the 64.4% 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 31% higher chance (14.6 percentage points) of being awarded benefits than non-disabled applicants. This difference decreases to 20% (or 12.7 percentage points) at the ALJ level. Despite the fact that the ALJs offer no major improvement in screening quality, the fact that there is significant self-selection in the appeal decision 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 9.

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 have a significant advantage over the DDS due to the self-selected nature of those choosing to appeal. That is, 74% of rejected applicants who appeal are disabled compared to 48% 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.

As noted above, to the extent that SSA does make fairly 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 (82.5%) than does a non-disabled applicant (62.1%). 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.

Many of the conclusions from our analysis are consistent with problems that a previous SSA commissioner, Kenneth Apfel, noted in the disability award process, particularly the problem of excess stringency on the part of the DDS which are counteracted by 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.

8 Analysis of a statistical screening rule for disability determinations

This section shows that it is possible to outperform SSA's disability award process by using an optimal statistical screening rule to determine award decisions. The optimal procedure takes the form of an "index rule" that accepts applicants when the calculated disability index $x'\beta$ is sufficiently large. We compare the error rates incurred under the current disability process to those implied by the optimal screening procedure, and show that this screening procedure significantly reduces the award and rejection errors in the disability award process. We then 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 in the next section.

The idea behind the statistical screening rule is simple. The RUR hypothesis implies that self-reported disability $\tilde{d}$ is an unbiased "signal" of true disability status $\tilde{\tau}$. If SSA had the luxury of observing each applicant's 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. Suppose, however, that SSA has access to either 1) a "training sample" of applicants who did truthfully reveal their disability status $\tilde{d}$ to a neutral third party (such as the self-reports in an anonymous survey such as the HRS), or 2) unbiased measurements of $\tilde{\tau}$ for a set of applicants that have been obtained via independent determinations of a group of experts such as the moderated team decision in the Nagi study discussed in section 2. In either of these two cases, the unbiasedness of the measurements implies that the training sample can be used to estimate the probability of being truly disabled as a function of the objectively measurable characteristics x. That is, these training samples can be used to estimate the probability an applicant is truly disabled, $P\{\tau = 1|x\}$, as a function of their characteristics x.

Using this estimated probability, consider a test of the "hypothesis" that a given applicant is disabled. By the Neyman Pearson Lemma the optimal statistical procedure (i.e. the uniformly most powerful hypothesis test of fixed size) is a likelihood ratio test: *reject H_0 (the null hypothesis that the applicant is disabled) if and only if the likelihood ratio exceeds a cutoff level k:*

$$\frac{1-P\{\tau=1|x\}}{P\{\tau=1|x\}} > k. \tag{14}$$

Note that rejecting H_0 is equivalent to the decision to award disability benefits, and this is equivalent to the following rule: *reject the applicant if and only if the predicted probability of being disabled is sufficiently small:*

$$P\{\tau=1|x\} < \frac{1}{1+k} \tag{15}$$

Neyman and Pearson proved that such a rule is optimal in the sense that no other statistical screening rule results in a smaller probability that a truly disabled applicant would be rejected given any fixed *size.* In this case the size of this "test" is equivalent to the probability of rejecting an applicant who is really disabled. *Thus, this rule is optimal in the sense that it minimizes the probability of accepting applicants who are not disabled, given any fixed probability of rejecting applicants who are disabled.*

We implement this rule by using a consistent estimate of $P\{\tau=1|x\}$. In our case, using self-reported disability data $\tilde{d}$ from the HRS we estimate the probability $\hat{P}\{\tilde{d}=1|x\}$ of being disabled. This is a consistent estimator of $P\{\tau=1|x\}$ when the rational unbiased reporting hypothesis holds. Similar to hypothesis testing, we set the rejection threshold k (which in our case is the award threshold) so that the test has a pre-specified size. In our case, we propose a statistical screening rule of the form

$$a_o = I\left\{\hat{P}(\tilde{d}=1|x) \geq k_o\right\}, \tag{16}$$

where a_o denotes the optimal award decision and $k_o \in [0,1]$ denotes a "critical value" or award threshold. By varying the award threshold k_o we can specify different award probabilities. If the SSA wanted to target a fixed award rate p, it would determine the threshold value k_o as the smallest solution to

$$p = \int I\left\{\hat{P}(\tilde{d}=1|x) \geq k_o\right\} f(x)dx, \tag{17}$$

where $f(x)$ is the distribution of the observed health characteristics in the applicant population.

We computed values of the threshold k_o to match the 57% acceptance rates at the DDS stage and the 76% ultimate award rate in the current DI award process, analyzed in Section 7. This was done by specifying an initial guess for k_o, calculating the sample analog of (17), using the empirical distribution of x in the HRS data, and increasing or decreasing k_o until the implied award rate matched the desired award rate p (.57 or .76 in this case).

Table 6 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.

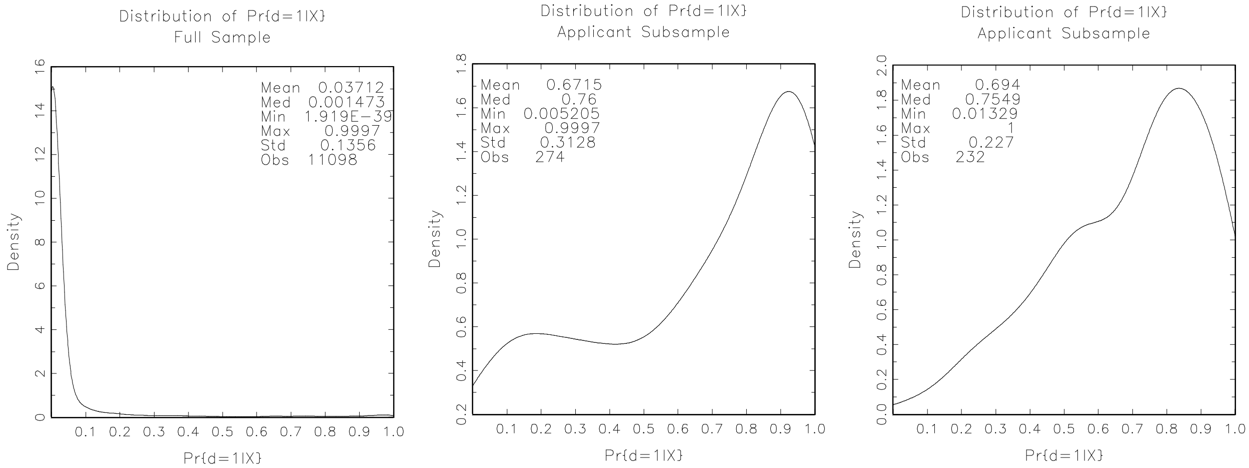
Although there are far fewer observations in the applicant subsample, we feel these results are the more relevant for this paper since it is for this subsample that we have verified the RUR hypothesis. Because we cannot compare a non-applicant's reported value of $\tilde{d}$ with $\tilde{a}$, we are 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.

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 6. 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 6, 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 6. 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 corresponding 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.

Table 6: 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: Distributions of $\Pr\{\tilde{d} = 1|x\}$ for Different Subsamples



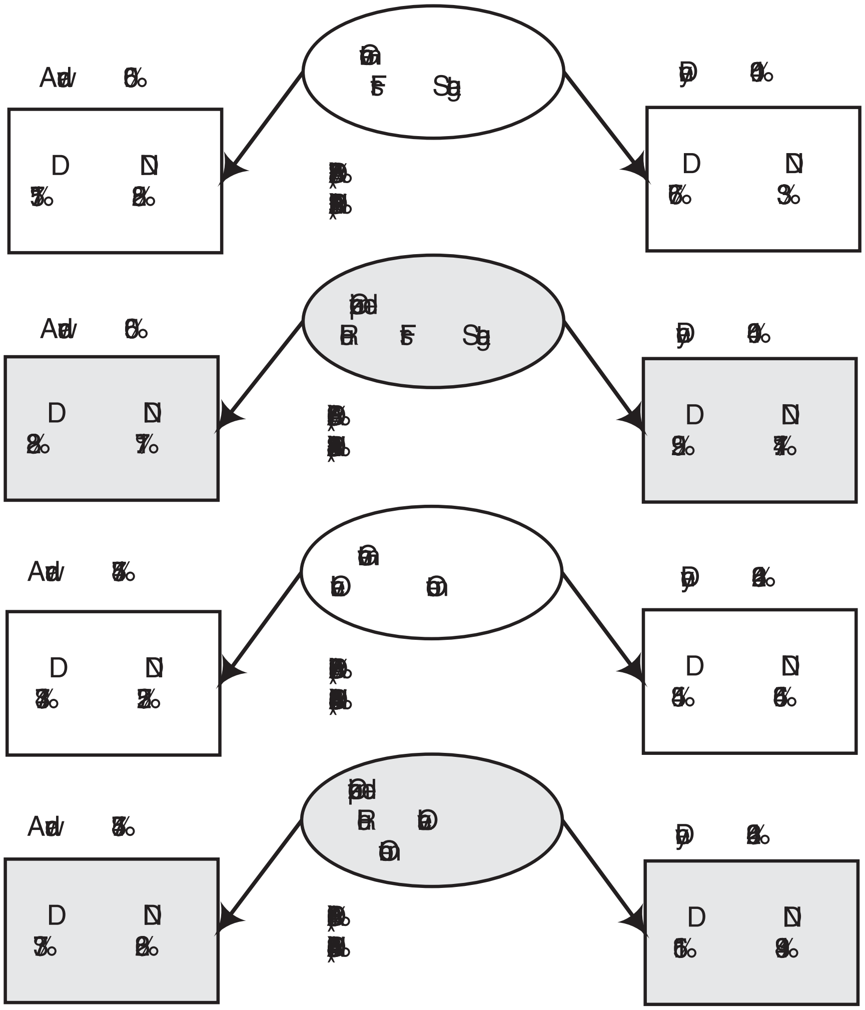
The third panel of Figure 5 plots the distribution of $\Pr\{\tilde{d} = 1|x\}$ that emerges when we re-estimate the model using only the subsample of SSI and SSDI applicants. This panel corresponds to the 232 applicants in the second column of Table 6. This distribution is even more skewed to the right 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 $\Pr\{\tilde{d} = 1|x\}$ since the individuals who are most likely to be aware of the SSA's definition of disability (and thus satisfy the RUR hypothesis) are the subsample of SSI and SSDI applicants. Furthermore, as we have seen in Table 3, the applicant population is very different from the full HRS sample. Indeed, we see significant differences in the logit coefficient estimates in Table 6 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 6 and in the two right hand panels of Figure 5 confirm this.

Figures 6 and 7 compares the classification errors resulting from using the optimal statistical screening rule a_o versus the SSA's actual award decision $\tilde{a}$, for the full sample and applicant subsample, respectively. While it might seem that by construction the optimal screening rule will necessarily outperform any other screening rule, including the SSA's award rule, this will only be the case when the SSA's decisions are based on the same health characteristics x that we used to construct the optimal screening rule. If the SSA has access to more information on the applicant's health status than we were able to observe in the HRS, then the "optimal" rule based on limited information will not necessarily outperform the SSA.

We calculated two thresholds k_o: the first enabled us to match the lower 57% first stage award rate of the

DDSs, and the second allowed us to match the higher 76% ulitmate award rate for our sample of applicants from the HRS. We compared the classification errors from the optimal screening rule to those from SSA's actual award decisions using the same approach as in the previous section (i.e. where we continue to assume that $\tilde{d} = \tilde{\tau}$ with probability 1).

Figure 6: Computerized Screening Rule: Full Sample



The top panels of Figures 6 and 7 compares the classification errors for the optimal screening rule to classification errors of SSA's actual first stage award decisions that are made by the DDSs where we adjusted the threshold k_o so that the optimal screening rule resulted in an award rate of approximately $p_c = .56$.[24] The bottom part of Figures 6 and 7 compare the classification errors for the optimal screening rule to the classification errors in SSA's ultimate award decision, i.e. allowing for the option to appeal. In this case p is

[24] The actual award rate differs slightly from the $p_c = .57$ target in Figure 6 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.

set equal to the ultimate award rate .76 rather than the first-stage award rate set by the DDS.

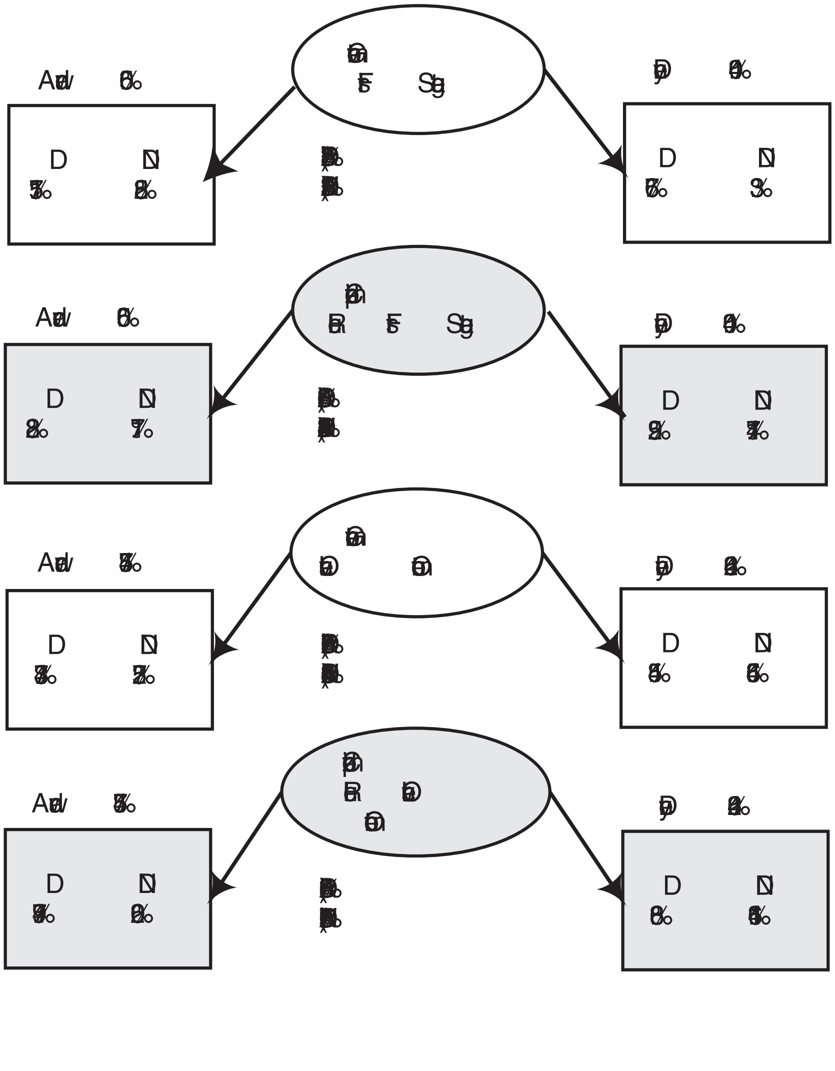
For both the full and restricted sample, we find that the optimal screening rule substantially outperforms the first stage decisions by the DDS in terms of classification errors. For example, the top part of Figure 6, the optimal screening rule results in an award error rate of 17.7% compared to the actual DDS rate of 28.5%. Similarly, the optimal screening rule results in a large reduction of the rejection error: 52.9% of those rejected by the optimal screening rule are disabled compared to 66.7% of the DDS rejectees. The optimal 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 6 shows that the optimal screening rule does not outperform the SSA when we use predicted probabilities $\hat{P}\{\tilde{d} = 1|x\}$ based on the full sample, and where we consider the ultimate award rate, i.e. after allowing for appeals to ALJs. Although there are still numerous discrepancies between the SSA's award decision and the optimal screening rule, these discrepancies tend to be offsetting so that the two procedures yield approximately the same rates of classification errors.

Figure 7 shows that the optimal screening rule does uniformly dominate the SSA's disability award process when we compute $\hat{P}\{\tilde{d} = 1|x\}$ using only data from the subsample of SSDI and SSI applicants. Figure 7 shows that at both stages of the disability award process the optimal rule dominates the SSA's award process in terms of award and rejection error rates. In the first stage, the optimal rule results in a 17.7% award error rate and a 52.9% 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 optimal 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. This could be due to our conjecture that self-reported disability is more accurately reported by DI applicants, since they presumably are more informed about the SSA's standards for judging disability.

We conclude this section with a discussion of several caveats regarding the interpretations of these findings. First, there is an issue of how SSA could obtain a "training sample" of unbiased signals of true disability status. One possibility would be to rely on survey results from a trusted third party, such as the HRS, that is not associated with the SSA and can provide strong assurances of anonymity and confidentiality to motivate truthful and accurate reporting of disability status. However, a drawback of this approach is that due to survey time and budget constraints, the set of health characteristics x that were collected in the HRS may only be a small subset of the set of characteristics that the SSA would like to use to make disability determinations. To gather these data the SSA may want to consider running a large scale replication of the Nagi study described in section 2. This would involve having teams of doctors, psychologists, vocational rehabilitation experts etc. make independent determinations about whether a large sample of DI applicants is truly disabled or not, recording the results of an extensive set z of medical and health measures that the SSA is interested in using to make disability determinations. This study would then provide the relevant "training sample" from which the probability of disability $P\{\tilde{\tau}=1|z\}$ could be estimated.

Figure 7: Computerized Screening Rule: Sample of Applicants



Once this disability probability has been estimated from the training sample, subsequent disability determinations can be made by having a team of suitably qualified doctors examine the applicant and record all of the relevant z variables that SSA wishes to use to make a disability determination. We recommend that these variables be measured by a team of doctors who are agents of the SSA rather than continuing to follow current practice where, in effect, the SSA relies on the z variables recorded by the *applicant's doctor.* Presumably the applicant's doctor, being an agent of the applicant rather than of the SSA, could have incentive to misreport the z variables in an attempt to increase the chances that their client (the applicant) will be awarded disability benefits.

We would also recommend that instead of allowing the examining team to make the disability determination, the z variables should be transmitted to SSA and it would make the determination if the predicted probability of disability exceeds a preset acceptance threshold. The advantage of this procedure relative to allowing the examining team to make the disability determination "on the spot" after their examination, is that it reduces the possibility that "empathy" on the part of the examiners could affect award decisions. It also helps to promote a more uniform application of disability award standards throughout the country. As we noted in section 1, there are very wide unexplained state to state variations in award standards under the *status quo.*

Clearly, a statistical screening rule does not obviate the need for human beings and some type of disability determination bureaucracy. We envision that a statistical 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 and "intangible variables" to be considered rather than relying only on a predefined vector z of health variables, and a mechanical implementation of the optimal screening rule. Besides allowing hard to quantify information to be considered, an appeal would also allow the applicant to challenge the z variables recorded by the SSA's medical examiners. If the examining team was found to have mismeasured or misrecorded key variables, the relevant members of the team could be subjected to penalties, providing further incentive for the SSA examiners to be as accurate as possible in measuring the applicant's z characteristics.

9 Implications for SSA's "Disability Process Redesign" Initiative

The problems with the DI determination process discussed in the previous sections motivated the SSA to propose a comprehensive "disability process redesign" plan in 1994. The stated goal of the plan was 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 alternative approaches to disability evaluation using *functional impairment indices* which are 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 discussed in section 2—that the high reversal rates by the ALJs actually serves to *reduce* the 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.

Regarding the larger issue of redesigning the disability award process the SSA realized that any fundamental change in its disability award process would require a substantial research effort combined with the collection of new data on disabled individuals. In the mid 1990s the SSA funded contracts for the design of a survey called the National Study of Health and Activity (NSHA) and contracted with the Committee on National Statistics (CNS) to advise it on the design of this survey and the overall plan of research for its disability redesign initiative. SSA recognized that "unless SSA invests substantially more funds to research and development of the simplified disability determination methodology, the full benefits of the redesigned process … will not be possible." (SSA, 1994).

The key element of its redesign initiative was a new approach that would "focus directly, rather than indirectly, on the applicant's functional ability to work and would rely on standardized instruments for measuring functional capacity to reach decisions." (Wunderlich et al. 2002, p. 116). SSA's proposed new procedure for making disability determinations bears some resemblance to the "statistical screening" approach that we have described in the previous section, although the procedure that SSA envisioned did not necessarily take the form of a formal "index rule" such as we have proposed.

> *"SSA assumed that under this proposed decision process, the majority of disability claims would be evaluated using a standardized approach to measure functional ability to perform substantial gainful activity. Standardizing the approach to assessing individual functional ability would facilitate consistent decisions regardless of the professional training of the decision makers in the decision process. The new disability decision process, as envisioned by SSA, would assess a person's functional ability once, relying on objective, standardized, functional assessment instruments. SSA believed that focusing decisions on the functional consequences of a person's medical impairments would permit physicians and others who provide medical evidence, as well as decision makers, to use a consistent frame of references for determining disability, regardless of the diagnosis and would facilitate evidence collection by reducing the need for developing extensive medical records." (Wunderlich et al. 2002, p. 116).*

The CNS report that evaluated SSA's research plan advised that "SSA should a. establish evaluative criteria for measure the performance of the decision process; b. conduct research studies and analyses to determine how the current processes work relative to these preestablished criteria, and c. evaluate the extent to which change would lead to improvement." (Wunderlich et al. 2002, p. 131). We believe our study makes significant headway in all three of these recommendations. Unfortunately, the SSA subsequently decided to discontinue its research on the disability redesign initiative and the NSHA survey. In part due to concerns raised by CNS on the design of its research plan, in 1998 "SSA undertook an internal reevaluation of its disability decision process redesign initiatives. SSA concurred with several of the committee's conclusions and some of its recommendations. However, rather than undertaking the additional research and redirection of the research as recommended by the committee, SSA decided to no longer actively pursue the new decision-making process proposed in Disability Redesign, but to improve the current process, focusing at this time on updating the Listings." (Wunderlich et al. 2002, p. 123).

In September 2003, Jo Anne Barnhart, Commissioner of Social Security, introduced a new set of initiatives to change the current process to determine eligibility of disability benefits. The main objectives of the new proposal are to speed up the processing times, which the Commissioner acknowledges as excessive, create accountability along the process, and expand the employment opportunities for people with disabilities. The most radical changes suggested by the Commissioner include the elimination of the DDS reconsideration and Appeals Board stages of the disability appeal process. These stages would be replaced by a Federal Reviewing Official who will be in charge of appeals that were previously handled by these separate stages. The plan also establishes new "expert units" who are specialized in screening applications to make quick decisions in clear-cut cases.[25] The latter set of decisions will be made at the Regional Offices

[25] The elimination of the DDS reconsideration stage and the creation of the Federal Reviewing position is likely to reduce the large state variation we discussed earlier in section 1. Notice that the Commissioner does not directly mention classification errors, but does mention the importance of accountability and supports the creation of a centralized quality control system. More details on the Commissioner's proposal can be found at `www.ssa.gov`.

where the team of experts will be located, even before the applications are send to the DDSs. This set of experts would include doctors in different areas of clinical specialty, including orthopedics, psychiatry, etc. If the SSA were to record the disability determinations of these teams of experts, along with data on their measurements of physical conditions of the applicant and various functional impairment indices, this could be an important first step towards collection of data for Nagi-style analysis of classification errors, and possibly the first step towards a statistically based screening procedure similar to the one we have described in the previous section.

10 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 in the mid 1990s embarked on a major re-examination of its entire disability determination process, and introduced innovative new ideas about how it might be restructured to make it faster, fairer, and more consistent.

Unfortunately in response to critiques raised by a study of its redesign effort by the Committee of National Statistics, the SSA has appeared to have retreated from this ambitious undertaking, and has scrapped plans for the collection of data necessary to evaluate its existing disability award process and evaluate alternative designs. Thus, we are currently in a situation where the most recent available study evaluating the classification errors in SSA's disability award process is a study by the sociologist Nagi over 3 decades ago. In light of the underinvestment in data collection, this study has tried to fill this gap by using data from the Health and Retirement Study to assess the magnitude of classification errors in SSA's award process. Although our approach of relying on a self-reported disability indicator from the HRS as a key piece of information to assess the magnitude of classification errors is likely to be controversial, it represents the only available means we have of making such assessments, given the government's unwillingness to invest in the necessary research and data collection that would enable us to evaluate the DI award process via less controversial (but undoubtedly far more expensive and time consuming) methodologies.

Although there is a substantial theoretical literature on "mechanism design" and the use of "monitoring technologies" to solve adverse selection and moral hazard problems, the theory has few concrete implications for the design of the U.S. disability award process. In principle, it is possible to run a DI program without employing any monitoring technology: the SSA could 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 also 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 (BBCCR) we used a battery of powerful empirical tests to show that $\tilde{d}$ is an unbiased and accurate indicator of the Social Security's ultimate award decision $\tilde{a}$. This paper builds on this result by hypothesizing that both $\tilde{a}$ and $\tilde{d}$ are noisy but unbiased indicators of *true disability* $\tilde{\tau}$. By making some additional assumptions about the stochastic structure of these variables we have been able to use observations on $(\tilde{a}, \tilde{d})$ to make inferences about true disability $\tilde{\tau}$, and calculate classification errors via Bayes rule. Our calculations show that the classification errors in the existing award process are substantial: over 20% of DI awardees are not disabled and as many as 60% of DI rejectees are disabled. We find that these results are robust to variations in our assumptions about the stochastic structure governing $(\tilde{a}, \tilde{d}, \tilde{\tau})$. In particular, even though we provide strong evidence that self-reported disability is a more accurate indicator of true disability than the SSA's award decision, the estimated classification error rates remain very large even when we change the stochastic structure to make the SSA's award decision $\tilde{a}$ to be significantly more highly correlated with the true disability indicator $\tilde{\tau}$ than self-reported disability status $\tilde{d}$.

Critics might claim that the reason our previous study failed 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* (which 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, it seems unlikely that our conclusions are spurious, or result 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 55%, respectively). The magnitude of these errors is consistent with Nagi's findings, although an entirely different methodology was employed.

2. 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 recomputed the classification errors under the assumption that both the self reported disability $\tilde{d}$ and SSA's award decision $\tilde{a}$ are noisy but unbiased indicators of true disability status $\tilde{\tau}$, and obtained very similar estimates of the degree of classification and Type I and II errors, even when we assume that the the award decision is more correlated with the true measure than the self-reports are.

3. 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.

4. 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.

5. 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.

6. We compared the performance of the disability award process to an alternative statistical screening rule based on an estimate of the conditional probability that an individual is disabled given objectively verifiable characteristics x. The computerized rule accepts an applicant for whose predicted probability of being disabled is sufficiently high. We set this threshold so that the computerized screening rule would yield the same award rate as is currently generated by the SSA. We found that the statistical

screening rule substantially reduces the rate of classification errors. If we were to use our optimal screening rule to replace the overall DI award process, we predict that for our sample, the award error rate would fall from 26% to 21% and the rejection error rate would fall from 54% to 39%. If we use the optimal screening 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 optimal screening rule results in an award error rate of 17.7% (nearly 10 percentage points less than the 28.5% award error rate of the DDSs), and an rejection error rate of 52.9% (14 percentage points lower than the rejection error rate of the DDSs). The implementation of this computerized rule seems more feasible than ever with the proposed reforms to the determination process that Commissioner Barnhart has recently presented, which advocate for the establishment of a team of medical experts at each Regional Office.

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 statistical 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 would use its teams of medical experts who would collect accurate measures of x for each applicant. We presume that since these experts would be 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 (Benítez-Silva, Buchinsky, and Rust 2003).

We are also currently working on estimating 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, solving, and estimating 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

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., 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, S. Cheidvasser, and J. Rust (2003), "How Large is the Bias in Self-Reported Disability Status?" forthcoming, *Journal of Applied Econometrics.*

Benítez-Silva, H., M. Buchinsky, and J. Rust (2003), " Using a Life Cycle Model to Predict Induced Entry Effects of a $1 for $2 benefit offset in the SSDI Program," manuscript, University of Maryland.

Bound, J., R. Burkhauser (1999): "Economic Analysis of Transfer Programs Targeted on People with Disabilities," in O. Ashenfelter and D. Card (eds.), Handbook of Labor Economics, Volume **3C.** Elsevier Science: Amsterdam.

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

Haveman, R.H., and B. Wolfe (2000): "The Economics of Disability and Disability Policy," in in A.J. Culyer and J.P. Newhouse (eds.), Handbook of Health Economics, Volume 1. Elsevier Science: Amsterdam.

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.

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. (1996): “Imperfect ‘Tagging’ in Social Insurance Programs,” *Journal of Public Economics*, **62** 183–207.

Rust, J. and C. Phelan (1997): “How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Markets,” *Econometrica*, **65-4** 781–831.

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.

Social Security Administration (1994) “Plan for a Disability Claim Process” Baltimore, Maryland.

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.

Wunderlich, G. S. Dorothy P. Rice and N. L. Amado (eds.) (2002) *The Dynamics of Disability: Measuring and Monitoring Disability for Social Security Programs* National Academy Press, Washington, D.C.