Disability Insurance and Gender Differences: Evidence from Merged Survey-Administrative Data[*]

Hamish Low[†] and Luigi Pistaferri[‡]

February 6, 2019

Abstract

Using HRS data matched with Social Security administrative data, we document large gender differences in disability insurance programs admission rates and type I error rates. In particular, women who apply for DI/SSI are 13 percentage point less likely to be awarded benefits than men, controlling for health, occupation and a host of demographic characteristics. Moreover, women who self-report to be disabled are 20 percentage points more likely to be rejected than observationally similar men. We investigate whether these gender differences can be explained by heterogeneity in underlying unobserved health, differences in disability perceptions, higher noise-to-signal ratios, or SSA evaluators' assessment bias. We find little support for the first three explanations, and some indirect support for the latter.

Keywords: Disability Insurance, Gender Differences.
JEL Classification Codes: I380, J16.

PRELIMINARY AND INCOMPLETE. PLEASE DO NOT CITE.

[*] Thanks to Maxwell Rong and Sarah Eichmeyer for invaluable research assistance, and seminar participants at Berkeley and Bonn for comments. This paper uses SSA confidential data made available to Pistaferri through contract RDA # 2015-028.

[†] Oxford University and Institute for Fiscal Studies.

[‡] Stanford University, SIEPR, NBER and CEPR.

1 Introduction

This paper studies gender differences in the award rates of disability insurance applications in the US. We focus on the two major programs paying benefits against disability risk: the Social Security Disability Insurance (DI) program and the Supplemental Security Income (SSI) program. The DI program is financed through payroll taxation and pays cash and health care benefits to covered workers who have become disabled. The SSI program is financed through general taxation and pays cash benefits to low-income and low-resources individuals who are disabled or senior (aged more than 65). In this paper we focus on working age individuals.

DI and SSI have attracted a lot of attention in recent years for at least two reasons (see Duggan and Imberman, 2009, for a survey). First, they are large. In 2017 the DI program was paying cash benefits of around \$134 billions (in comparison, the Unemployment Insurance (UI) program was paying benefits worth only \$28 billions). In the same year, total SSI expenditure was \$59 billion, absorbing 16% of federal non-Medicaid welfare spending. Second, the two programs have grown at very high rates over the last 30-35 years. Between 1984 and 2017 the share of disabled workers receiving DI benefits out of all workers who are insured increased from 2.4% to 5.6%. Moreover, during the same period the share of total social security expenditure (OASI plus DI) accounted for by DI benefits paid to disabled workers increased from 8.7% to 14.1%. As for SSI, we calculated that the fraction of 18-64 years old in the US population who are receiving SSI benefits has doubled from 1.2% in 1984 to 2.4% in 2016 (this fraction includes people who are receiving both SSI and DI).

The growth of these programs has renewed concerns that some working-able individuals may quit into unemployment in order to apply for benefits or exaggerate their disability in order to become beneficiaries; or that beneficiaries may have little incentives to go back to work even when their health conditions improve. In the case of SSI, the additional concern is that applicants may be discouraged from saving in order to meet the asset test. There exists now a sizable literature that has looked at these so-called "moral hazard" aspects of disability insurance. Some of the earlier empirical literature is surveyed in Bound and Burkhauser (1999) and Haveman and Wolfe (1999). More recent contributions are surveyed in Low and Pistaferri (2015). Less attention has been paid in the literature to the "coverage" aspect of disability insurance, i.e., how many potentially eligible applicants are turned down given the inherent noise of the evaluation process.

These inefficiency aspects of disability insurance may in principle depend on the applicants' characteristics. For example, error rates may be higher for conditions that are harder to verify, such as muscolosketal or mental disorders. Moreover, certain demographic groups (such as women or minorities) may face explicit discrimination when assessed by evaluators.

To address these complex issues we merge survey data from the Health and Retirement Survey with Social Security administrative data. Survey data give us a measure of the "true" disability status of an individual (as self-reported by the individual); administrative data provide detailed information on the outcome of a DI/SSI application.[1] Jointly, the two sources of data give us measures of Type I and Type II error. Armed with these measures, we can study whether these errors differ by observable characteristics of applicants.

We document significant gender differences in both award rates and Type I error rates. Women with a severe, work-related, permanent impairment are more likely to have their disability insurance application turned down (i.e., suffer a type I error) than men with observationally equivalent characteristics. This main finding is robust to various robustness checks. In the second part of the paper we propose a simple model that tries to rationalize the main findings. Men and women could differ in terms of severity of actual or perceived impairments, or in terms of opportunity costs of applying. On the supply side, men and women could face different admission standards (for example because evaluators use a "unisex" approach that disadvantage women), or SSA may receive much less precise signal from women applicants. We show that supply considerations are more plausible explanations than demand-side channels.

The rest of the paper proceeds as follows. In Section 2 we provide a review of the existing literature. In Section 3 we provide institutional details on the programs that insure against disability shocks. In Section 4 we describe the data. Section 5 discusses our main findings. Section 6 presents a simple structural model that tries to distinguish between various explanations for our findings. Section 7 concludes.

2 Literature Review

2.1 Classification errors

How large are the DI or SSI classification errors? Are they a function of observable applicant characteristics? Do they vary with the degree of reliability of the disability "signal“? Since the true disability status of an applicant is unknown, the examiners are prone to making two type of errors: Type I errors (rejecting a truly disabled applicant) and Type II errors (awarding benefits

[1] While DI and SSI have mostly been studied in isolation, there are a number of reasons why it may be valuable studying them jointly. First, the formal definition of disability is the same in both programs. Second, the disability determination process is done by the same agencies and officers (local Social Security field offices). Finally, in survey data the number of applicants to either program is typically small; merging application data from the two programs makes inference more reliable.

to applicants who are not truly disabled). We know of at least three paper that have estimated classification errors.

An early study is Nagi (1969). In this paper, a sample of 2,454 DI applicants who had received a final decision from the SSA were followed by a team of doctors, psychologists, and social workers. Through home visits and various psychological and physical tests, the team was tasked with making an alternative deliberation regarding the true disability status of an applicant. Nagi (1969) concluded that, at the time of the award, about 19% of those initially awarded benefits were undeserving, and 48% of those denied were truly disabled. To the extent that individuals recover but do not flow off DI, we would expect the fraction falsely claiming to be higher in the stock than at admission. This is the finding of Benitez-Silva et al. (2006) who use self-reported disability data on the over 50s from the Health and Retirement Study (HRS). Benitez-Silva et al. (2006) assume that the self-reported binary indicator of work limitations is a classical error-ridden measure of the "true" disability status. In the HRS, respondents report also the outcome of a DI/SSI application. The combination of these two pieces of information allow them to compute classification errors for the two programs combined. They find that over 40% of recipients of DI/SSI are not truly work limited. Low and Pistaferri (2015) follow a similar strategy of using self-reported health status alongside details of receipt of DI, but use data from the Panel Study of Income Dynamics (PSID) and focus on the DI program. Unlike the HRS, the PSID also includes younger workers (aged less than 50). Low and Pistaferri (2015) distinguish between severe and moderate disability (instead of using a binary indicator), and estimate classification errors using a structural model (since they only observe DI receipts but application information are missing). Low and Pistaferri (2015) find that the Type I error is large (approximately 2/3 for younger workers and 1/3 for older workers), while the Type II error is concentrated among those with moderate disabilities (18%) with the error being only 1% among those who apply while reporting no disabilities.

There are several issues that make estimates of classification errors from the studies above problematic. First, how strong is the "signal" embedded in the self-reported disability measures? Both Benitez-Silva et al. (2006) and Low and Pistaferri (2015) rely on it, although the latter study tries to go beyond the simplest binary disability indicator. Second, the Nagi (1969) study refers to a period in which the disability programs were fundamentally different (indeed, the SSI started only in 1974). The most dramatic difference is that in the pre-1984 period admission into disability programs because of muscolosketal or mental disorders was rare. The 1984 Social Security Disability Benefits Reform Act liberalized admission criteria for DI and SSI, resulting in a large increase in applicants and people awarded benefits with such conditions.[2] Since these are hard-to-verify conditions, it is likely that classication errors in the post-1984 era are much different than in the pre-1984 period studied by Nagi (1969). Third, the Benitez-Silva et al. (2006) study relies on survey-data information on the application process, which may be subject to measurement (recall) errors. These are particularly relevant in cases in which a disability improves or worsens, since one needs to "pin" the disability status at the time of the application in order to assess the extent of classification errors. Moreover, in some years disability application questions are only asked to those who report having a disability, which induces a mechanical understatement of Type II errors.

It is also worth stressing that none of these papers try to understand how classification errors vary by how verifiable the disabling condition is and whether they vary by demographic characteristics.[3] One might expect classification errors to be higher for hard-to-verify conditions (such as back pain) than for conditions for which clinical evidence is more easily obtained. Indeed, the disability determination process distinguishes between individuals with a so-called "listed impairment" (with essentially a certain award) and those without (where health conditions and work adaptability criteria are taken into account). Second, the reliance on the so-called disability matrix studied by Chen and van der Klaauw (2008) implies that classification errors may be higher for younger workers or those with high levels of education or experience (conditioning on a disability being present).[4]

In this paper we merge HRS information on self-reported disability with administrative information on DI and SSI applications and social security earnings. To our knowledge, this is the first paper that attempts to use these linked data for studying this efficiency aspect of the DI/SSI programs. Benitez-Silva et al. (2006) and Low and Pistaferri (2015) use survey data. Other papers use only administrative data and hence cannot not study classification errors (since these study

[2] Among the provisions of the Act there were at least three that may have increased the probability of admission because of such conditions: (1) the requirement that SSA obtain evidence from the applicant's treating physician instead of hired consultants, "since the treating physician is likely to be the medical professional most able to provide a detailed, longitudinal picture of the individual's medical condition"; (2) updating the criteria for evaluating mental impairments to "make them consistent with present-day diagnosis, treatment, and evaluation"; (3) the requirement that the SSA, in determining the severity of a person's disability, "consider the combined effect of all impairments without regard to whether any one impairment, if considered separately, would be severe" (Collins and Erfle, 1985).

[3] In fact, we are not aware of any academic paper studying differences in the screening process by observable demographic characteristics (such as gender or race). A study by the US General Accounting Office (1994) investigated the reasons for the higher disability insurance denial rates among women. The conclusion was that a significant fraction of the difference could be explained by occupation dummies and the fact that SSA evaluators assessed that women apply with lower impairments than men (a somewhat tautological statement given the absence of information on the "true" disability status of applicants).

[4] The extent of errors in the process may also differ at different points of the application process. Given the long waiting periods and different appeal processes that applicants go through to get onto DI, we might expect different stages of application to be subject to different sorts of error. Further, there is an interaction between effects: work limitations caused by muscoloskeletal illnesses or mental health tend only to lead to the award of DI at the later stages of the appeal process.

lack an independent measure of the true disability status of an individual). With caveats discussed later, we assume in this paper that self-reported disability is an error-ridden measure of the true disability status of an applicant.

3 Institutional Details

3.1 The DI program

The DI program is a social insurance program that provides cash and health care benefits for covered workers, their spouses, and dependents. Cash benefits are computed using the same formulae used to compute Social Security retirement benefits. In particular, DI beneficiaries receive indexed monthly payments corresponding to their Primary Insurance Amount (PIA), which is based on taxable earnings averaged over the number of years worked (known as Average Indexed Monthly Earnings, or AIME). While benefits are independent of the extent of the work limitation, caps on the payroll tax financing the DI program as well as the nature of the formula determining benefits make the system progressive. Because of the progressivity of the benefits and because individuals receiving DI also receive Medicare benefits after two years, the replacement rates are substantially higher for workers with low earnings and those without employer-provided health insurance.

The purpose of the DI program is to provide insurance against persistent health shocks that impair substantially the ability to work.[5] The difficulty with providing insurance is, of course, that health status and the impact of health on the ability to work are imperfectly observed.

The award of DI benefits depends on the following conditions: (1) An individual must file an application; (2) There is a work requirement on the number of quarters of prior employment: Workers over the age of 31 are disability-insured if they have 20 quarters of coverage during the previous 40 quarters; (3) There is a statutory five-month waiting period out of the labor force from the onset of disability before an application will be processed; (4) individuals who work must earn no more than a so-called "substantial gainful amount" (SGA, $1,170 a month for non-blind individuals as of 2017); and (5) the individual must meet a medical requirement, i.e. the presence of a disability. Since this last requirement is the same as in the SSI program, we discuss it below after a short description of SSI.

[5] The emphasis on the severity and persistence of the health shock distinguishes the DI program from the Workers Compensation program, which pays cash and health care benefits for temporary health shocks that are work-related, or private medical leave programs.

3.2 The SSI Program

Working-age individuals who are deemed to be disabled and have limited income and limited resources are eligible to receive supplemental security income (SSI). [6] The definition of disability in the SSI program is identical to the one for the DI program, while the definitions of low income and low resources is similar to the one used for the Food Stamps program.[7] SSI benefits are adjusted annually. In 2017, an individual (couple) would receive $735 ($1,103) in cash benefits.

3.3 The Disability Determination Process

The disability determination process is common to both DI and SSI applicants and consists of sequential steps. Applicants submit their application to a local field office. The case is evaluated by a Disability Determination Service (DDS) officer. The first step for DDS is to determine whether the applicant has a medical disability that is severe and persistent. This is defined as: "*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.* If such disability is a "listed impairment", the individual is awarded benefits without further review (step 2).[8] If the applicant's disability does not match a listed impairment, the DDS evaluators try to determine the applicant's residual functional capacity. In the last two stages of the disability determination process the pathological criterion is paired with an economic opportunity criterion. In particular, the third stage tries to verify if the individual retains functional capacity for his/her *past* work; in the fourth and last stage, if there is functional capacity for *any* work. Two individuals with identical work limitation disabilities may receive different disability determination decisions depending on their age, education, general skills, and even economic conditions faced at the time the determination is made.

The disability insurance application that is unsuccessful at the initial DDS stage (a not-so-rare event, given that only about 37% of applicants are awarded benefits at this stage) can be appealed. About 1/3 of denied applicants do so. The case, which is not updated with new information, is transferred to a different officer within DDS, a stage that is called "reconsideration". The success rate at this stage is even lower than at the initial stage (14%). Those denied at the reconsideration

[6] The SSI program serves also children with disabilities and seniors with limited-means (with or without a disability), two groups that are not our focus.

[7] In particular, individuals must have income below a "countable income limit", which typically is slightly below the official poverty line (Daly and Burkhauser, 2003). SSI eligibility also has an asset limit ($2,000 for individuals and $3,000 for couples.).

[8] The listed impairments are described in a blue-book published and updated periodically by the SSA ("Disability Evaluation under Social Security"). They are physical and mental conditions for which specific disability approval criteria has been set forth or listed (for example, Amputation of both hands, Heart transplant, or Leukemia).

stage can further appeal (and 3/4 of denied applicants do so). Their case leaves DDS and is decided by Administrative Law Judges (ALJ), where applications tend to have a much larger success rate (63%).[9] Rarely do cases go beyond the ALJ stage, and if they do the award rates are significantly lower.

4 Data

4.1 The Health and Retirement Study (HRS)

The Health and Retirement Study (HRS) is a panel data set administered by the Institute for Social Research at the University of Michigan. It is the main data source for researchers interested in investigating questions related to population aging in the US. Its population target consists of individuals aged 50 and more. We merge a harmonized version of the HRS that has been assembled by the RAND Center for the Study of Aging, containing biannual waves 1992 through 2014, with other HRS data from the raw files. The most relevant variables in this dataset are: (a) the self-reported presence of a work limitation, defined as "an impairment or health problem that limits the kind or amount of paid work" that a respondent can do, together with information about whether the condition is temporary, and whether it prevents work altogether; (b) indicators for the presence of specific health conditions (high blood pressure, diabetes, cancer, lung disease, heart disease, stroke, psychiatric problems, and arthritis), as reported to the respondent by his/her own physician, as well as a variety of other health indicators; (c) Disability Vignette data.

4.2 Social Security Administrative Data

For consenting respondents (approximately 3/4 of the entire sample), HRS data can be linked to administrative data on earnings and benefits available from the Social Security Administration (the Master Earnings File (MEF), and the Master Beneficiary Record (MBR) file), and to Form 831 Disability Records (F831), which contain information on the initial medical determination (i.e., the outcome of the initial review and of the reconsideration, both done at the SSA level) of an applications to DI and/or SSI. The F831 database does not contain information on decision made at the ALJ level and beyond.[10] However, from the Master Beneficiary Record (MBR) file, one can verify whether a DI application was eventually successful by checking whether an individual is receiving social security benefits classified as: "Benefits to a disabled worker". Unfortunately, we cannot reconstruct whether SSI applicants were eventually successful.

[9] The higher success rate at this stage partly reflects applicants' selection, partly the possibility of integrating the file with new information, and partly the possibility to advocate one's case in court.

[10] Also, no F831 case is open if the applicant receives a "technical denial" (i.e., people with earnings above SGA).

The F831 database includes multiple records per individual. We distinguish between application cycles and application rounds.[11] For each cycle we observe four key variables: (a) the exact application date of any round; (b) the outcome of each application round, together with the exact decision date; (c) the primary impairment (body system) code;[12] and (d) whether it is a DI, an SSI, or a concurrent DI/SSI application.

5 Results

5.1 Self-reported disability indicators vs. clinical health measures

To estimate Type I and Type II errors, we need a measure of the "true" disability status of an individual. The SSA defines disability as: "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." We replicate this definition using three survey questions from the HRS: (1) "Do you have any impairment or health problem that limits the kind or amount of paid work you could do?"; (2) "Is this a temporary condition that will last for less than three months?"; and (3) "Does this limitation keep you from working altogether?". We classify as disabled people who answer "Yes" to the first and third question and that report that the condition is not temporary. This way, we replicate very closely the three criteria set forth by the SSA definition: the presence of a work-related impairment, its severity, and its expected duration.

In thinking about estimates of Type I and Type II errors, the key question is how reliable this variable is in measuring true disability. The literature has not reached a full consensus on this issue. Benitez-Silva et al. (2004) argue that "...self-reported measures give individuals latitude to summarize a much greater amount of information about [the applicant's] health and disabilities than can be captured in the more objective, but very specific indices used in previous studies". As discussed in Bound and Burkhauser (2000), the use of self-reported disability measure raises two basic issues: (a) inter-personal comparability, and (b) endogeneity with respect to labor market outcomes (i.e., those who apply for DI or are out-of-work are more likely to self-report a disability as a way of rationalizing their decisions). Disability vignettes are one way of tackling the first

[11] An application cycle may include up to two rounds: the initial DDS assessment, and the DDS reconsideration (if there is one).

[12] These are: Musculoskeletal system, Special senses and speech, Respiratory system, Cardiovascular system, Digestive system, Genito-urinary system, Hemic and lymphatic system, Skin, Endocrine system, Multiple body systems, Neurological, Mental disorders, Neoplastic diseases, Immune deficiency, Growth impairment, Other. We also observe a more detailed sub-categorization (impairment codes) (i.e., for those applying with a Musculoskeletal system body system code, we observe whether it is Disorders of Back (discogenic and degenerative), Osteoarthrosis and Allied Disorders, and so forth).

criticism, as we argue below. As for the second issue, we can verify whether self-reported disability is associated with more objective or clinical indicators of disability for which there is less scope for rationalization. This is what we do in Table 1.

The HRS contains additional information on the health of respondent which are of a more objective or diagnostic nature. First, respondents are asked whether they have difficulties with basic activities in their daily living (ADL's), such as dressing, preparing meals, etc., because of a health condition.[13] Second, we observe some objective indicators of poor health, such as whether a person has spent some time in hospital and for how long, BMI data (so we can determine obesity or being underweight), and whether people leave the sample because of death, clearly the most objective indicator of poor health. Finally, we have information on whether a doctor has told the respondents that they have some specific condition, like high blood pressure, cancer, etc.

Table 1 compares average values of these various health indicators for the "disabled" and "not disabled" groups. Clearly, people who self-report a disability are much more likely to have a clinical or diagnostic health condition, and more likely to encounter difficulty in ADL's. For example, only about 2% of not disabled peiople have trouble walking across a room, as opposed to 20% in the disabled group. There are similarly large differences for other ADLs. Mortality is 25% vs 45%. Hospital stays are almost three times more likely and five times longer among the disabled group. Finally, the disabled are much more likely to have been diagnosed with a serious health condition. If we stratify by gender (results not shown), there is similar qualitative evidence.

5.2 Descriptive statistics for the matched sample

Our main estimation sample consists of HRS respondents who apply for disability insurance and whose disability status is observed around the time of the application.[14] In principle, one would like to observe the disability status exactly at the time of the application. Unfortunately, if we were to match only those whose interview date coincides with the date of disability insurance application, we would be left with an extremely reduced sample (especially because HRS is conducted every other year). Instead, we use all applications that we can match with an HRS interview that is no more than 12 months away from the application date. To make sure that this criterion is not responsible for our results, we perform several robustness checks.

In Table 2 we report descriptive statistics for the matched sample, comprising 944 first-round

[13] The presence of ADL difficulties plays an important role in the official determination of disability. For example, DI/SSI applicants are required to fill in an "Activities of Daily Living Form" report (known as the Function Report, SSA-3373). Moreover, long-term care insurance policies require that an applicant needs help with two or more ADL before triggering benefits.

[14] We drop proxy respondents.

applications (374 from men and 575 from women). This sample reproduces almost identically the award rate observed in the population of all applicants at the initial consideration stage (37%). The award rate is slightly higher if we also consider the reconsideration stage (42%). There are large differences in award rate between men and women (a 12-13 percentage point difference), and especially in Type I error (a 22-23 percentage point difference), despite the fact that a larger fraction of women than men report to be disabled (54% vs. 47%).

Finally, it is worth noting that almost half of the individuals in the applicant sample reports not to have a disability. Two comments are in order. First, this is hard to reconcile with a "rationalization" story (and more likely to be consistent with the idea that people report truthfully their health conditions to HRS interviewers). Second, our definition of disability (which requires people to be altogether unable to work) is possibly more stringent than the SSA definition (where people can actually work up to the SGA amount). It is therefore even more surprising to find the high rejection rates and type I error rates we do find.

5.3 Award Rates and Classification Errors

Some of the differences in award rates and Type I error between men and women could be generated by differences in observable characteristics, something that our formal regressions is designed to account for. Indeed, as Table 2 shows, men and women differ in many important dimensions. Male applicants are older and with more labor market experience, they are more likely to be college-educated, and less likely to be black, unmarried or widowed. Summary statistics for the primary disability condition reported on the F831 application form are in the lower part of Table 2. Men are more likely to apply for disability insurance because of a cardiovascular condition, while women are slightly more likely to apply because of a muscolosketal or mental disorder.

While denial rates at initial consideration are high, they differ substantially by primary disability code, as shown in Figure 1. Denial rates are higher for conditions that are harder to verify, such as muscolosketal disorders and mental disorders. Awards are more likely for applicants with cancers or disorders of the genito-urinary system (such as kidney failures). This means that when we try to explain differences in award rates, it is key to account for the underlying disability conditions one is applying for (i.e., higher denial rates among women could be due to the fact that women are more likely to apply with high denial-rate conditions, such as muscolosketal disorders).

In Table 3 we present results for the probability of being awarded disability insurance benefits at the initial consideration stage. This is the least problematic stage, since award at reconsideration and further stages are affected by various forms of selection. We consider different specifications varying in terms of richness of controls. The first column reproduces the unconditional difference noticed in Table 2. In column (2) we control for self-reported disability. In column (3) we add key demographic controls. In columns (4)-(6) we add increasingly richer controls for diagnostic or objective health indicators. In the final column (7) we include controls for occupation dummies, since some of the rejections could come from being in an occupation where it is more likely to retain some functional capacity (i.e., a sedentary job). The results are surprisingly stable, regardless of how rich the set of controls is. Women are 12-13 percentage point less likely to be awarded benefits than men with observationally similar characteristics. Unsurprisingly, self-reported disabled and older people are more likely to have their application approved, as well as exclusive SSI applicants. Occupation dummies are jointly statistically significant.

To estimate the relationship between classification errors and gender, we run the following probit model for applicants who report to be disabled ($D_{ij} = 1$):

$$\Pr(Reject_{ij}|D_{ij} = 1) = \Phi(X'_{ij}\alpha_0 + \alpha_1 Female_i) \tag{1}$$

For Type II errors, we run the following probit model for applicants who do not report a disability ($D_{ij} = 0$):

$$\Pr(Award_{ij}|D_{ij} = 0) = \Phi(X'_{ij}\beta_0 + \beta_1 Female_i) \tag{2}$$

where i is individual, j is application, and X_{ij} includes individual- and application-level controls. As said above, our primary focus is on at outcomes at the initial consideration stage.

Table 4 reports results for Type I errors; Table 5 for Type II errors. Controlling for a vast variety of characteristics, we find statistically significant higher type I error rates for women (a 20 percentage point difference in the richest specification of column (7)). Apart from gender, the only significant characteristic is age: older applicants are less likely to be turned down if truly disabled. The key of these large differences is that, as shown in Figure 2, denial rate differences are much larger for women relative to men for high-award rate conditions such as cancer, respiratory conditions, neurological conditions, or genito-urinary conditions. For these conditions, rejection rates for women are 1.5 to 3 times larger than for men.

Table 5 shows that the effect of gender on Type II error is consistent with the idea that women applicants are "less believed", both when they are truly disabled and when they are not severely disabled. However, the estimate is smaller, poorly measured, and insignificant in the richest specification where we also control for occupation dummies. For these reasons, from now on we will focus on the evidence for Type I error. The focus on Type I error is also because of our interest on the insurance aspects (as opposed to the moral hazard aspects) of disability insurance.

5.4 Robustness

We perform various robustness checks. First, we focus on the DI sample, which is the program more traditionally studied in the literature. Second, we experiment with different "timing" assumptions. Third, we change the definition of disability. All the results are contained in Table 6. For comparison, column (1) reproduces the results of the baseline specification (from Table 4, column (7)). All regressions include the same controls used in the baseline.

The results are confirmed, and if anything display a slightly larger gender differences, if we focus on the DI applicants sample. Our baseline sample includes individuals who are interviewd within 12 months from the date of their disability insurance application. Since the "timing" of the match is arbitrary, in columns (3)-(5) we experiment with different assumptions. In column (3) we use those interviewed 3 months before to 9 months after the application date. In column (4) we focus on those interviewed up to 9 months following the date of application. Finally, in column (5) we use the same criterion of the baseline, but weight more those interviewed closer to the application date (we use as weight $1/\sqrt{d}$, where d is the distance between the date of the interview and the date of the disability insurance application). If people recover from a disability, this criterion is the closest we can get to the "true" disability status at the point of application. While the results change quantitatively (ranging from 0.11 to 0.19), they are qualitatively similar to the baseline: Female applicants experience higher type I error than observationally equivalent male applicants.

In the final columns (6) and (7) we adopt different disability definitions. Our definition (that a person has a non-temporary impairment that prevents work altogether) may be even stricter than the one adopted by SSA (where applicants and recipients are permitted to do some work as long as pay remains below the SGA). In column (6) we classify as disabled those who report to have an impairment or health problem that limits the kind or amount of paid work they can do. This is the standard binary definition of disability used in many papers in the literature. In column (7) we assume that an individual is disabled if he/she reports difficulties with two or more activities of daily living. We adopt this definition because it is the one used by LTCI policies for triggering payment of benefits. The samples are clearly different than the baseline, and yet the qualitatively estimates are very similar, confirming the presence of significant gender differences in Type I errors.

5.5 Type I errors at later evaluation stages

So far we have focused on outcomes at the initial consideration stage. In Table 7 we consider Type I error at the overall DDS level (i.e., initial consideration *and* reconsideration), and Type I error over the entire sample period where an individual is observed (i.e., if the applicant is ever awarded disability insurance).

Starting with overall DDS experience, we notice that adding a reconsideration stage (where Type I errors may in principle be reduced) does not change the results - women are statistically significantly more likely to be turned down if disabled than observationally equivalent men whether or not we add an appeal stage. In contrast, the results from column (3) are different: we no longer find a significant effect.[15] This partly reflects the fact that in the long run average Type I error declines from 54% to only 18%. However, while the elimination of the gender differences in Type I error is comforting (in the sense that the appeal process remedies some of the errors of the earlier stages of the disability determination process), there are a number of caveats to this conclusion.

First, even if errors are corrected at later stages of the evaluation process, there are still welfare implications from going uninsured during the time it takes to process applications and appeals. The evaluation process can be fairly long. A study by the Office of the Inspector General for fiscal year 2006 estimated that the average (cumulative) processing times for a disability insurance application were 131 days for the initial DDS decision, 279 days for the DDS reconsideration decision, 811 days for the ALJ decision, and 1,720 days for a Federal Court decision. Another implication is the effect of waiting on human capital. A recent paper by Autor et al. (2015) argues that longer processing times reduce the employment and earnings of DI applicants for multiple years following application, with the effects concentrated among applicants denied benefits at the initial stage.

Second, the sample used in column (3) does not include those who move into OASI by having reached age of retirement while the DI application was in process, and those who have an application still in process when the HRS end, two forms of censoring we do not tackle directly. Moreover, at stages above DDS people can update the health information, boosting the chances of an award, something we do not observe. As well, and perhaps more importantly, appeals are obviously higher among the sicker cases. Finally, gender differences may still exist despite no differences in eventual awards if women appeal at higher rates than men.

6 Explanations

In this section we propose a simple structural framework that distinguishes between a number of different explanations for the differences across gender in Type 1 errors. First, men and women may have a different distribution of the underlying (poor) health. Second, they may have different "pain threshold" perceptions, and apply to disability insurance accordingly. Third, they may have

[15] Given data limitations, we can only measure (from the SSA Master Beneficiary Record database) if a person has ever received DI, but cannot verify if a person has ever received SSI, hence in this regression we have a reduced sample.

different perceptions of the norms that are set by SSA in regards to what it means to be disabled. Fourth, from the supply side the SSA may set different admission thresholds for men and women, or observe male and female applicants' signals with different precision.

6.1 A simple structural framework

Consider a simple statistical framework that tries to capture these different channels. Suppose that the true, latent disability status of an individual i is given by:

$$D_i^* = \alpha_0 + X_i'\alpha_x + \alpha_D Female_i + \varepsilon_i \tag{3}$$

where F is a female dummy and $\varepsilon_i \sim N(0,1)$. The female dummy captures potential shifts in the underlying health distribution – women may have less severe ($\alpha_D < 0$) or more severe ($\alpha_D > 0$) underlying health impairments than men. *Ceteris paribus*, women with less severe impairment than men are more likely to experience Type I errors.

A second possibility is that men and women differ in their "pain threshold" and so differ in their self-reports of disability accordingly. Assume that individuals report to be disabled if their latent disability status is above a certain threshold, $\bar{D}_i$:

$$D_i = \mathbf{1}\{D_i^* > \bar{D}_i\} \tag{4}$$

We allow this threshold to vary by gender:

$$\bar{D}_i = \gamma_0 + \gamma_{\bar{D}} Female_i + u_i \tag{5}$$

where $u_i \sim N(0,1)$. If women have a lower pain threshold, $\gamma_{\bar{D}} < 0$ and more women than men will classify themselves as disabled despite their underlying health being the same (this is the problem of interpersonal comparison of self-reports of disability). There is a vast medical literature that attempts to examine the relationship between pain tolerance/sensitivity and gender. In a review of the experimental literature, Racine et al. (2012) conclude: "10 years of laboratory research have not been successful in producing a clear and consistent pattern of sex differences in human pain sensitivity, even with the use of deep, tonic, long-lasting stimuli, which are known to better mimic clinical pain". In a review of both clinical and experimental studies, Fillingim et al. (2009) conclude that "recent clinical and epidemiologic findings generally indicates that women are at increased risk for many chronic pain conditions, and women tend to report higher levels of acute procedural pain" (which of course would make our findings even more puzzling), while "findings regarding sex differences in experimental pain indicate greater pain sensitivity among females compared with males for most pain modalities [...]. The evidence regarding sex differences in laboratory measures of endogenous pain modulation is mixed, as are findings from studies using functional brain imaging to ascertain sex differences in pain-related cerebral activation".

The identification problem is evident once we replace (3) in (5) and compute the probability of self-reporting a disability (which, unlike latent disability, is something we observe in the data). This is:

$$\Pr(D_i = 1) = \Phi((\alpha_0 - \gamma_0) + X_i'\alpha_x + (\alpha_D - \gamma_{\bar{D}})Female_i) \tag{6}$$

Clearly, it is impossible to assess whether women are more likely to report a disability because they have a lower pain threshold ($\gamma_{\bar{D}} < 0$) or because their underlying health is worse ($\alpha_D > 0$).

Another source of potentially useful information is the decision to apply for disability insurance. Can application decisions help identifying the parameters of interest? Assume that people apply for disability insurance if their latent, true disability status exceeds a given threshold:

$$D_i^* > \bar{A}_i \tag{7}$$

The application threshold $\bar{A}_i$ may differ from the "perceived disability" threshold $\bar{D}_i$ for a number of reasons (although we generally expect them to be positively correlated). For example, applicants may "cheat", i.e., apply even when they are not truly disabled. Another possibility is that applicants have imperfect information about SSA norms or face different transaction costs of applying. Finally, some individuals may have a very high application threshold if they continue to be productive in the labor market despite the presence of a genuine disability.

We assume that the application threshold differ from the disability threshold by a linear function of characteristics, including gender:

$$\bar{A}_i = \bar{D}_i + \rho_0 + X_i'\rho_x + \rho_{\bar{A}}Female_i \tag{8}$$

This captures the possibility that men and women differ in their cost of applying as well as in their knowledge of SSA norms (or cheating attitudes).

Combining (3) and (8) we can compute the probability of applying for disability insurance (which we observe in the data):

$$\Pr(A_i = 1) = \Pr((\alpha_0 - \gamma_0 - \rho_0) + X_i'(\alpha_x - \rho_x) + (\alpha_D - \gamma_{\bar{D}} - \rho_{\bar{A}})Female_i) \tag{9}$$

Equation (9) shows that we can estimate the shift in application decision due to gender (for example, if women have worse knowledge of the SSA norms or lower costs of applying, $\rho_{\bar{A}} < 0$).

However, we are still unable to separate α_D from $\gamma_{\bar{D}}$. To solve the identification problem, we use disability vignette data available in the 2007 wave of the HRS. Disability vignettes, which have been pioneered in the disability literature by Kapteyn et al. (2007) are used to separate shifts in the underlying health distribution from subjective evaluation of thresholds, precisely the identification problem faced in our context.

In the disability vignette literature, respondents are asked to assess, on the same scale on which they asses themselves, the extent of disability in hypothetical situations and for hypothetical individuals. The 2007 Disability Vignette is a special, mail-only supplement of the HRS.[16] Respondents are first asked if they have a health limiting condition (”Do you have any impairment or health problem that limits the kind or amount of work you can do?”), and to rank it in terms of severity (possible responses are ”None”, ”Mild”, ”Moderate”, ”Severe” and ”Extreme”). To keep up with the analysis from the first part of the paper we convert this into a binary indicator, and assume a person is disabled if he/she answers that the limitation is ”Severe” or ”Extreme”. Next, each respondent is asked nine vignette questions in total, with three questions per each health condition (”Affect”, ”Pain”, and ”Cardiovascular disease”), describing the situation of individuals with different degrees of work limitations. As an example, one of the vignettes reads: “X has pain in [his/her] back and legs, and the pain is present almost all the time. It gets worse while [he/she] is working. Although medication helps, [he/she] feels uncomfortable when moving around, holding and lifting things at work. How much is X limited in the kind or amount of work [he/she] could do?”. Possible responses are ”None”, ”Mild”, ”Moderate”, ”Severe” and ”Extreme” (which we again convert into a binary disability indicator if the response is ”Severe” or ”Extreme”). Hence, people are asked to rank the vignettes using the same severity scale that was used to rank their own health. The key aspect of the vignette questions is that respondents are randomly assigned to one of two versions, which differ in the order of questions and in the gender assigned to the hypothetical person described in the vignettes. Hence for some people the X person above is a “Mark” and for other respondent the same description refers to a “Tamara”.

To see how vignettes can be used to tackle the identification issues discussed above, suppose that respondent i evaluates the latent disability status of vignette v according to the following equation:

$$D^*_{v,i} = \theta_{v,i} + \zeta_{v,i} \tag{10}$$

and classifies the vignette as disabled if the underlying latent variable crosses a threshold, i.e.:

[16] The HRS conducted a vignette survey also in 2004 (a “leave behind“ supplement), but there was no gender randomization involved, so we focus on the 2007 version.

$$D_{v,i} = \mathbf{1}\{D^*_{v,i} > \bar{D}_{v,i}\} \tag{11}$$

where $\zeta_i \sim N(0,1)$. As discussed by Kapteyn et al. (2007), the two key identification assumptions that are standard in this literature are: (1) Vignette Equivalence, and (2) Response Consistency. The first assumption is that the situation described in the vignette is perceived by respondents in the same way, that is, $\theta_{v,i} = \theta_v$ for all i. This is because all respondents are presented with the identical description of an hypothetical person, and the only differences in their perceptions should be random (i.e., misread the sentence). The second assumption is that respondents evaluate the health of the vignette characters in the same way that they evaluate their own health, i.e., the threshold they use for the vignette is the same as they would use for themselves: $\bar{D}_{v,i} = \bar{D}_i$ for all i.

Given these assumptions, the probability that respondent i classifies the vignette as disabled is given by:

$$\Pr(D_{v,i} = 1) = \Phi(\theta_v - \gamma_0 - \gamma_{\bar{D}} Female_i) \tag{12}$$

This shows that differences in "pain thresholds" by gender ($\gamma_{\bar{D}}$) can be pinned down by how men vs. women respondents differ in their evaluation of the vignette's disability – leaving reports of own disability to pin down shifts in the underlying true health status (α_D).

Another potential explanation for differences in type I error between men and women is the possibility that SSA sets higher standards for women than for men (a "supply" explanation). The ideal experiment would be to assign two identical applications (one by a man, one by a women) to a DDS evaluator and check if award rates differ. Unfortunately, this experiment is not feasible. However, we can test whether there is an inherent higher threshold set by evaluators for female applicants by using again the vignettes.

Rewrite slightly the perceived disability status of a vignette (equation (10)) in the following way:

$$D^*_{v,i} = \theta_v + \theta_F Female_v + \zeta_{v,i} \tag{13}$$

where F_v is a dummy for whether the hypothetical vignette describes a woman. If the respondent sets higher disability standards for women, he/she would be less likely to classify the vignette as disabled if that vignette describes a woman, i.e., $\theta_F < 0$. We can also test if women are "tougher" on women by adding the interaction $Female_v * Female_i$ to (13).

To summarize, the simple structural model described in this section can generate higher Type I errors for woman (consistently with the empirical analysis) in four different ways: (1) women may have a lower pain threshold, $\gamma_{\bar{D}} < 0$, which generates more application for less objectively disabled women (and hence larger type I errors); (2) women may have a lower cost of applying ($\rho_{\bar{A}} < 0$); (3) women's health may be objectively better ($\alpha_D < 0$); or (4) women may face tougher standards set by SSA ($\theta_F < 0$). To estimate these key parameters, we are going to use variation from three sources: (a) self-reports of disability; (b) disability insurance application decisions; and (c) the probability of classifying a vignette as disabled. Since we no longer need information from form F831, these regressions use the entire HRS sample.

6.2 Results

6.2.1 Disability self-reports and Application decisions

Table 8 reports the results of probit regressions for the decision to self-report a disability and the decision to apply for disability insurance (equations (6) and (9)). In the first case, our outcome variable is 1 if the individual reports to be disabled (defined as in the baseline regressions) and zero otherwise. In the second case, the outcome variable equals 1 if we observe an "open" first-round application to DI or SSI at any point in time in a given calendar year t for individual i, and 0 otherwise (i.e., an application that is either unadjudicated at the time of the HRS interview or was adjudicated in the same interview year). For the observations with $Applied_{it} = 0$, we focus on those person-years in which applying to DI or SSI is in a person's choice set. We implement this condition by dropping person-years in which $Applied_{it} = 0$ and the individual is a recipient of SSI or DI benefits in that year. The sample only includes people below age 65.

In both regressions, the key variable is the "female" dummy. We add the same controls used in the Type I regressions above (except the F831 disability code dummies which are not observed for the whole HRS sample). We find that women are, controlling for a wide variety of characteristics (especially, health variables), *less* likely to report a disability and less likely to be applicants, although the marginal effects are small.

6.2.2 Disability Vignettes

Table 9 reports summary statistics for the vignette data. We are going to present results for two samples. The first includes all the HRS participants who are in the HRS 2007 wave and respond to the vignette questions. The other sample includes only the respondents aged 70 or less, which is an attempt to mimic the age of potential DDS or ALJ evaluators. Table 9 (using only the whole sample) shows that men tend to be more "lenient" than women (they report higher disability rankings, independently of the gender of the person in the vignette).

Table 10 presents the results of controlled regressions. The dependent variable is whether the respondent classifies a given vignette as disabled. This is the equivalent of equation (12) above. We control for a female respondent dummy, a female vignette dummy, and dummies for the vignette domain ("Affect", "Pain", or "Cardiovascular disease"). In columns (1)-(2) we use the whole sample, while in columns (3)-(4) we confine attention to the sample aged 70 or less. In each sample, we run two specifications. The first controls only for vignette health condition fixed effects and the gender of the respondent. The latter identifies the "pain threshold" parameter (α_D). Regardless of sample, we find that women respondents are less likely to report that a given vignette is disabled, implying that they have higher pain thresholds than men respondents. The second specification add a dummy for whether the vignette that is being evaluated is given a female name. We find that respondents tend to be "tougher" on female vignettes – they are less likely to classify a vignette as disabled if that vignette is named "Tamara" as opposed to "Mark". We find no evidence that women respondents tend to be tougher on women vignettes (results not reported). The results in the younger sample are similar.

6.2.3 Structural parameter estimates

Table 11 reports estimates of the structural shift parameters: α_D (measuring gender differences in the underlying true, latent disability), $\rho_{\bar{A}}$ (gender differences in the cost of applying), $\gamma_{\bar{D}}$ (gender differences in disability perceptions or pain thresholds), and θ_F (which measures differences in disability standards set by SSA for women vs. men). We use the identification scheme described above, and calculate standard errors using the Block Bootstrap.

Recall that any of the explanation for Type 1 error detailed above requires the structural parameter to be negative. The estimates for α_D, $\rho_{\bar{A}}$ and $\gamma_{\bar{D}}$ are all positive (although only $\gamma_{\bar{D}}$ is statistically significant), implying that we can dismiss that lower pain thresholds, lower application thresholds or less severe impairments for women are the explanation for higher Type I error. We have indirect evidence, however, that the SSA set higher standards for women than men (or uses for women the same, perhaps inappropriate standards set for men, who were the traditional users of the program). Note that the estimate do not change much if we focus on a sample of individual who can potentially work as DDS evaluators or Administrative Law Judges (aged 70 or less).

7 Conclusions

This paper documents substantial differences in rejection rates and type I error across genders. In particular, we find that women with a severe, work-related, permanent impairment are more likely to have their disability insurance application turned down (i.e., suffer a type I error) than men with observationally equivalent characteristics. We show that supply considerations (i.e., how SSA screens applicants) are more plausible explanations than demand-side channels (i.e., differences in disability perceptions or application costs).

Our results suggest that an important policy change would consist of making disability insurance applications gender-blind. Evidence from other settings show that gender-blind evaluations of candidates matter for explaining a variety of labor market outcomes (Goldin and Rouse, 2000; Bertrand and Mullainathan, 2004; Card et al., 2018).

References

[1] Autor, D., N, Maestas, K, J. Mullen and A. Strand (2015), ”Does Delay Cause Decay? The Effect of Administrative Decision Time on the Labor Force Participation and Earnings of Disability Applicants”, NBER Working Paper No. 20840.

[2] Benitez-Silva, H., M. Buchinsky, H. M. Chan, S. Cheidvasser, and J. Rust (2004), “How Large Is the Bias in Self-Reported Disability?”, Journal of Applied Econometrics, 19 (6), 649-670.

[3] Benitez-Silva, H., M. Buchinsky, and J. Rust (2006), “How Large are the Classification Errors in the Social Security Disability Award Process”, NBER WP 10219.

[4] Bertrand, M. and S. Mullainathan (2004), ”Are Emily And Greg More Employable Than Lakisha And Jamal? A Field Experiment On Labor Market Discrimination,” American Economic Review, 94(4), 991-1013.

[5] Bound, J. and R. V. Burkhauser (1999), “Economic Analysis of Transfer Programs Targeted on People with Disabilities”, in O. Ashenfelter and D. Card (eds.), Handbook of Labor Economics. Vol. 3C. Amsterdam: Elsevier Science, pp. 3417-3528.

[6] Card, D., P. Funk, N. Irriberri and S. Della Vigna (2018), ”Are Referees and Editors in Economics Gender Neutral?”, mimeo.

[7] Chen, S. and W. van der Klaauw (2008) “The Effect of Disability Insurance on Labor Supply of Older Individuals in the 1990s”, Journal of Econometrics, 142(2), 757-784.

[8] Collins, K. P. and Erfle, A. (1985), ”Social Security Disability Benefits Reform Act of 1984: Legislative History and Summary of Provisions”. Social Security Bulletin, 48(4), 5-32.

[9] Daly, M. C. and R. V. Burkhauser (2003), ”The Supplemental Security Income Program.” Robert Moffitt (ed.), Means Tested Transfer Programs in the United States. Chicago, IL: University of Chicago Press for the NBER, 79- 140.

[10] Duggan, M. and S.A. Imberman (2009), “Why Are the Disability Rolls Skyrocketing? The Contribution of Population Characteristics, Economic Conditions, and Program Generosity”, in *Health at Older Ages: The Causes and Consequences of Declining Disability among the Elderly*, NBER.

[11] Fillingim, Roger B., Christopher D. King, Margarete C. Ribeiro-Dasilva, Bridgett Rahim-Williams, and Joseph L. Riley III,(2009), "Sex, Gender, and Pain: A Review of Recent Clinical and Experimental Findings", Journal of Pain, 10(5): 447–485.

[12] Goldin, C. and C. Rouse (2000) "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." American Economic Review, 90 (4): 715-741.

[13] Golosov, M. and A. Tsyvinski (2006), "Designing Optimal Disability Insurance: A Case for Asset Testing", Journal of Political Economy, 114, 257-279.

[14] Haveman, R. and B. Wolfe (2000). "The Economics of Disability and Disability Policy", in A.J. Culyer and J.P. Newhouse (eds.), Handbook of Health Economics, Vol. 1, Amsterdam: Elsevier Science, pp. 995-1051.

[15] Kapteyn, Arie, James P. Smith and Arthur van Soest (2007), "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands", The American Economic Review, 97(1), 461-473.

[16] Low, Hamish and Luigi Pistaferri (2015), "Disability insurance and the dynamics of the incentive-insurance tradeoff", American Economic Review 105(10): 2986-3029.

[17] Nagi, S. Z. (1969), Disability and Rehabilitation. Columbus, OH: Ohio State University Press.

[18] Racine, Mélanie, Yannick Tousignant-Laflamme, Lorie A. Kloda, Dominique Dion, Gilles Dupuis, and Manon Choinière (2012), "A systematic literature review of 10 years of research on sex/gender and experimental pain perception – Part 1: Are there really differences between women and men?", Pain 153, 602-18.

[19] United States General Accounting Office (1994), "Social Security Disability. Most of Gender Difference Explained. Report to the Ranking Minority Member, Special Committee on Aging, U.S. Senate". GAO/HEHS-94-94.

Table 1: Health conditions by self-reported work limitation status

	Not disabled	*Disabled*
Difficulty walking across room	0.0243	0.1983
Difficulty dressing	0.0412	0.2306
Difficulty stooping, kneeling or crouching	0.3471	0.7885
Difficulty getting out of bed	0.0380	0.2410
Difficulty grocery shopping	0.0375	0.2784
Difficulty preparing meals	0.0199	0.1485
Hospital stay	0.1576	0.4229
Nights in hospital	1.1074	5.7242
Obese	0.3124	0.4029
Underweight	0.0091	0.0211
Died in sample	0.2538	0.4456
Doctor diagnosed high blood pressure	0.4112	0.6285
... psychological condition	0.1492	0.3681
... heart condition	0.1220	0.3315
... arthritis	0.4292	0.7065
... diabetes	0.1298	0.2750
... lung condition	0.0654	0.2079
... stroke	0.0288	0.1153
... cancer	0.0702	0.1115

Note: The unit of observation is a person-HRS wave for all variables except death, where it is just person. Respondents are defined as ”Disabled” if they report to have an impairment or health problem that limits the kind or amount of paid work they can do; if the condition is not temporary (lasting less than three months?); and if the limitation keeps them from working altogether. The sample is people aged 20-65 only.

Table 2: Descriptive statistics

	Men		*Women*		*All*	
	Mean	SD	Mean	SD	Mean	SD
Type I error, appl. round	0.39	0.49	0.62	0.49	0.54	0.50
Type II error, appl. round	0.32	0.47	0.25	0.44	0.28	0.45
Awarded, round	0.45	0.50	0.32	0.47	0.37	0.48
Type I error, appl. cycle	0.33	0.47	0.55	0.50	0.47	0.50
Type II error, appl. cycle	0.35	0.48	0.29	0.45	0.32	0.47
Awarded,cycle	0.50	0.50	0.38	0.49	0.42	0.49
Disabled	0.47	0.50	0.54	0.50	0.51	0.50
Applied SSI only	0.21	0.40	0.25	0.44	0.23	0.42
Applied DI + SSI	0.19	0.39	0.18	0.39	0.18	0.39
College degree	0.34	0.48	0.28	0.45	0.30	0.46
Black	0.26	0.44	0.29	0.46	0.28	0.45
Married	0.62	0.49	0.47	0.50	0.53	0.50
Widowed	0.04	0.20	0.12	0.33	0.09	0.29
Lab. mark. experience	20.27	8.81	17.11	8.15	18.36	8.55
Age	57.10	4.74	55.62	6.01	56.21	5.59
Type of condition in F831						
Musculoskeletal	0.40	0.49	0.42	0.49	0.41	0.49
Respiratory	0.04	0.19	0.07	0.25	0.05	0.23
Cardiov.	0.20	0.40	0.10	0.30	0.14	0.35
Endocrine	0.06	0.23	0.07	0.25	0.06	0.24
Neurol.	0.08	0.27	0.07	0.26	0.07	0.26
Mental dis.	0.08	0.28	0.10	0.30	0.09	0.29
Cancer	0.03	0.17	0.04	0.20	0.04	0.19
Immune def.	0.03	0.18	0.02	0.15	0.03	0.16
Dig. & Urin.	0.01	0.10	0.03	0.17	0.02	0.15
Other	0.07	0.26	0.08	0.27	0.08	0.26
Number of obs.	374		575		944	

Note: The sample includes all first-round applications (of all application cycles) observed in F831 data for HRS respondents.

Table 3: Probit regression for award rates at initial consideration

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.127***	-0.151***	-0.121***	-0.118***	-0.123***	-0.128***	-0.127***
	(0.033)	(0.032)	(0.032)	(0.033)	(0.032)	(0.032)	(0.036)
Disabled		0.191***	0.202***	0.186***	0.163***	0.154***	0.166***
		(0.031)	(0.030)	(0.031)	(0.031)	(0.031)	(0.031)
College degree			0.024	0.015	0.015	0.015	0.017
			(0.035)	(0.035)	(0.035)	(0.034)	(0.036)
Black			-0.017	-0.010	-0.012	-0.017	-0.019
			(0.036)	(0.037)	(0.036)	(0.036)	(0.035)
Lab. mark. experience			0.001	0.002	0.002	0.002	0.002
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Applied SSI only			0.093**	0.092**	0.089**	0.100**	0.092**
			(0.040)	(0.040)	(0.040)	(0.040)	(0.040)
Applied DI + SSI			0.008	0.021	0.009	0.009	0.012
			(0.043)	(0.043)	(0.043)	(0.042)	(0.041)
Married			0.007	0.012	0.004	-0.000	0.010
			(0.036)	(0.036)	(0.035)	(0.034)	(0.034)
Widowed			0.024	0.032	0.017	0.025	0.056
			(0.058)	(0.061)	(0.060)	(0.059)	(0.055)
Age			0.016***	0.018***	0.018***	0.018***	0.018***
			(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Health cond. FE	No	Yes	Yes	Yes	Yes	Yes	Yes
HRS Objective FE	No	No	No	Yes	Yes	Yes	Yes
Year FE	No	No	Yes	Yes	Yes	Yes	Yes
ADL FE	No	No	No	No	Yes	Yes	Yes
BMI+Hosp.	No	No	No	No	No	Yes	Yes
Occupation FE	No	No	No	No	No	No	Yes
(P-value)							0.001
Observations	953	915	915	881	881	875	875
Pseudo R^2	0.013	0.090	0.153	0.171	0.191	0.202	0.233

Note: Standard errors in parentheses, clustered at the individual level. Reported coefficients are marginal effects. *Experience* is years with non-zero wage income.

Table 4: Probit regressions for Type I errors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.224***	0.221***	0.200***	0.180***	0.191***	0.194***	0.206***
	(0.046)	(0.045)	(0.046)	(0.047)	(0.048)	(0.047)	(0.054)
College degree			-0.036	-0.011	-0.011	-0.018	-0.037
			(0.053)	(0.054)	(0.054)	(0.053)	(0.054)
Black			0.014	0.009	0.007	0.019	-0.010
			(0.059)	(0.059)	(0.058)	(0.057)	(0.056)
Lab. mark. experience			-0.005	-0.005	-0.004	-0.004	-0.001
			(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Applied SSI only			-0.081	-0.079	-0.084	-0.091	-0.081
			(0.058)	(0.059)	(0.058)	(0.058)	(0.056)
Applied DI + SSI			0.016	-0.007	0.003	0.016	-0.002
			(0.064)	(0.065)	(0.066)	(0.064)	(0.064)
Married			0.041	0.018	0.020	0.035	0.020
			(0.054)	(0.054)	(0.052)	(0.052)	(0.050)
Widowed			-0.020	-0.039	-0.017	-0.024	-0.066
			(0.082)	(0.087)	(0.086)	(0.084)	(0.081)
Age			-0.016***	-0.018***	-0.018***	-0.017***	-0.015***
			(0.005)	(0.005)	(0.004)	(0.005)	(0.005)
Health cond. FE	No	Yes	Yes	Yes	Yes	Yes	Yes
HRS Objective FE	No	No	No	Yes	Yes	Yes	Yes
Year FE	No	No	Yes	Yes	Yes	Yes	Yes
ADL FE	No	No	No	No	Yes	Yes	Yes
BMI+Hosp	No	No	No	No	No	Yes	Yes
Occupation FE	No	No	No	No	No	No	Yes
(P-value)							0.001
Observations	471	471	471	450	450	446	445

Note: Standard errors in parentheses, clustered at the individual level. The reported coefficients are marginal effects. *Experience* is years with non-zero wage income.

Table 5: Probit regressions for Type II errors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.067	-0.083*	-0.060	-0.076*	-0.081*	-0.089**	-0.065
	(0.045)	(0.042)	(0.041)	(0.042)	(0.042)	(0.042)	(0.045)
College degree			-0.005	-0.008	-0.003	-0.011	-0.013
			(0.046)	(0.046)	(0.045)	(0.045)	(0.048)
Black			-0.023	-0.046	-0.036	-0.038	-0.057
			(0.045)	(0.047)	(0.046)	(0.046)	(0.046)
Lab. mark. experience			-0.000	-0.000	-0.001	-0.001	0.001
			(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Applied SSI only			0.125**	0.109**	0.095*	0.108*	0.131**
			(0.055)	(0.055)	(0.053)	(0.055)	(0.053)
Applied DI + SSI			0.052	0.068	0.051	0.061	0.079
			(0.057)	(0.058)	(0.057)	(0.056)	(0.055)
Married			0.077*	0.062	0.061	0.063	0.074*
			(0.046)	(0.046)	(0.046)	(0.046)	(0.045)
Widowed			0.032	0.029	0.026	0.043	0.054
			(0.087)	(0.088)	(0.085)	(0.087)	(0.085)
Age			0.016***	0.016***	0.017***	0.018***	0.020***
			(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Health cond. FE	No	Yes	Yes	Yes	Yes	Yes	Yes
HRS Objective FE	No	No	No	Yes	Yes	Yes	Yes
Year FE	No	No	Yes	Yes	Yes	Yes	Yes
ADL FE	No	No	No	No	Yes	Yes	Yes
BMI+Hosp	No	No	No	No	No	Yes	Yes
Occupation FE	No	No	No	No	No	No	Yes
(P-value)							0.004
Observations	444	444	420	414	414	412	410

Note: Standard errors in parentheses, clustered at the individual level. Reported coefficients are marginal effects. *Experience* is years with non-zero wage income.

Table 6: Probit regressions for Type I errors: Robustness

			Timing assumptions			*Different disab. definitions*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Baseline	DI sample	$-3 \leq t \leq 9$	$0 \leq t \leq 9$	$0 \leq t \leq 12$, weighted	Less strict. disab. def.	At least two ADL's
Female	0.206***	0.250***	0.109*	0.165**	0.189***	0.117***	0.167***
	(0.054)	(0.060)	(0.064)	(0.065)	(0.055)	(0.040)	(0.058)
Observations	445	334	404	338	445	762	325

Note: Standard errors in parentheses, clustered at the individual level. Reported coefficients are marginal effects. All regressions include the controls of Table 4, column (7).

Table 7: Probit regressions for Type I errors: Further evaluation stages

	(1) *Baseline*	(2) *Rejected,* *DDS level*	(3) *Never on DI*
Female	0.206^{***}	0.200^{***}	0.058
	(0.054)	(0.052)	(0.058)
Observations	445	445	291

Note: Standard errors in parentheses, clustered at the individual level. Reported coefficients are estimates of marginal effects. All regressions include the controls of Table 4, column (7).

Table 8: Probit regressions for disability self-reports and DI/SSI application

	(1)	(2)
	Self-report	*Applies for*
	a disability	*disab. insur.*
Female	-0.00492	-0.0035**
	(.0035)	(0.0017)
College degree	-0.01339***	-0.0069***
	(.00336)	(0.0018)
Black	0.01488***	0.0080***
	(.00392)	(0.0019)
Lab. mark. exp.	-0.00249***	0.0005***
	(.000188)	(0.0001)
Married	-0.00875***	-0.0085***
	(.00334)	(0.0016)
Widowed	-0.00054	-0.0076**
	(.00567)	(0.0032)
Age	0.00107***	-0.0009***
	(.000279)	(0.0001)
Health cond. FE	Yes	Yes
Year FE	Yes	Yes
ADL FE	Yes	Yes
BMI+Hosp	Yes	Yes
Occupation FE	Yes	Yes
Observations	38937	40845

Note: Standard errors in parentheses, clustered at the individual level. Reported coefficients are marginal effects. *Experience* is years with non-zero wage income.

Table 9: Vignettes: Descriptive Statistics

	All resp.		*Male resp.*		*Female resp.*	
	Mean	SD	Mean	SD	Mean	SD
Mean vignette rating						
overall	3.11	0.59	3.16	0.59	3.08	0.59
female hypoth. person	3.09	0.73	3.15	0.73	3.05	0.73
male hypoth. person	3.13	0.74	3.16	0.73	3.11	0.74
Rating of own health cond.	1.94	1.11	2.02	1.14	1.90	1.09
Age	65.41	10.55	66.31	10.25	64.83	10.69
Observations	4457		1758		2699	

Note: 1=Not limited, 2=A little limited, 3=Moderately limited, 4=Severely limited, 5=Extremely limited

Table 10: Probit regressions for the probability that vignette is disabled

	All respondents		*Respondents aged* ≤ 70	
	(1)	(2)	(3)	(4)
Female Respondent	-0.031***	-0.031***	-0.038***	-0.038***
	(0.007)	(0.007)	(0.009)	(0.009)
Female Vignette		-0.017***		-0.013**
		(0.006)		(0.007)
Vign. health cond. FE	Yes	Yes	Yes	Yes
Observations	40,113	40,113	28,224	28,224

Note: Standard errors in parentheses, clustered at the individual level. Reported coefficients are marginal effects.

Table 11: Structural estimates

		Whole sample	*Aged* $\leq$ 70
Health distr. shifter	α_D	0.038	0.057
		(0.039)	(0.040)
		[0.29]	[0.13]
Application threshold	$\rho_{\bar{A}}$	0.027	0.027
		(0.052)	(0.052)
		[0.62]	[0.62]
Disability report threshold	$\gamma_{\bar{D}}$	0.088^{***}	0.107^{***}
		(0.019)	(0.023)
		[<0.01]	[<0.01]
SSA threshold	θ_F	-0.048^{***}	-0.037^{*}
		(0.016)	(0.020)
		[0.01]	[0.06]

Note: Block bootstrap s.e. in round brackets; p-values in square brackets.

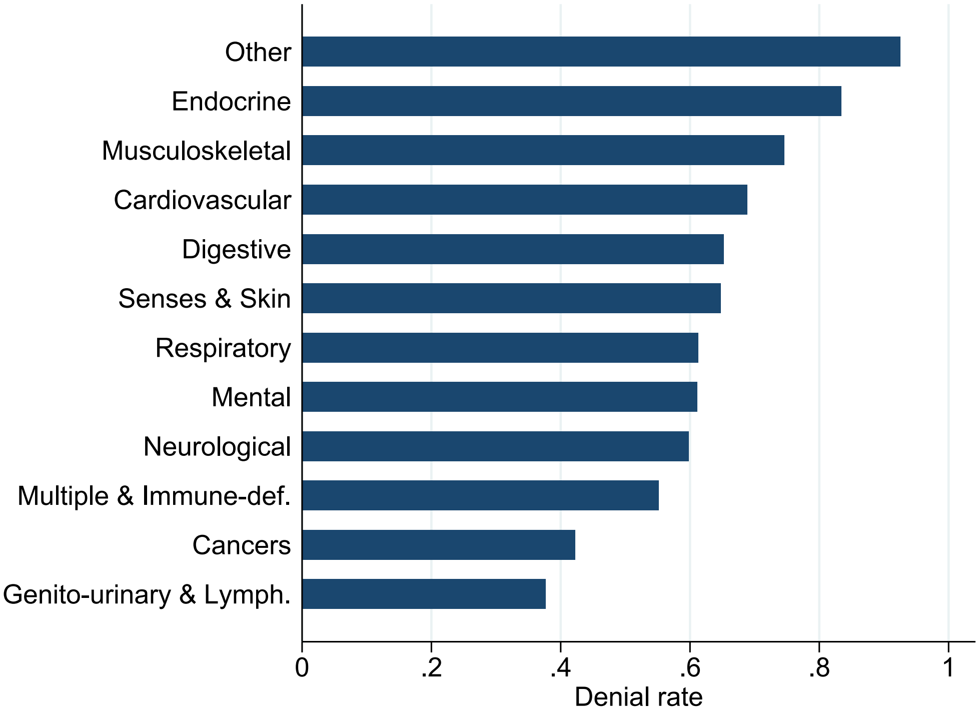

Figure 1: Denial rates by primary disability code

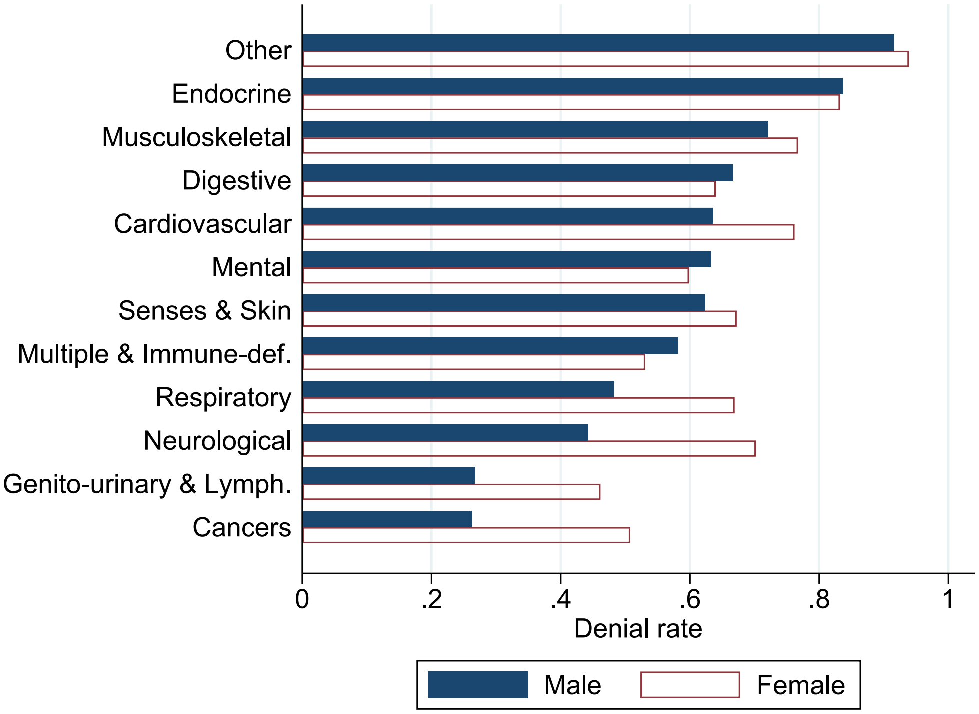

Figure 2: Denial rates by primary disability code and gender