Disability Insurance: Error Rates and Gender Differences (Low & Pistaferri 2020)

Gabriella Fleischman & Kelsey Pukelis

Health Equity Reading Group

September 23, 2021

Outline

- summary by discussion
- context:
  - DI/SSI context
  - framework for evaluating screening policies (Kleven & Kopczuk 2011)
- highlights about this paper's approach
  - defining "true" disability status
  - conceptual framework, model
  - identification
- relation to discrimination literature
- speculation about gender differences in error rates
- policy implications

summary by discussion

- a brave soul or two to summarize the main results of the paper?

summary by discussion

- Main results
  - Women are more likely to be falsely rejected for DI than men
  - Women who are rejected do *not* return to work
  - Structural estimates suggest the mechanism is different thresholds for accepting applications from men and women
- Definitions
  - Type 1 Error: Someone who truly has a work-limitation is rejected from DI
  - Type 2 Error: Someone who does *not* have a work-limitation is accepted for DI

Disability insurance context

- rising disability rolls in the U.S. since mid-1980s (Autor & Duggan 2003)
- ...especially for mental health and musculoskeletal conditions — low mortality impairments
- labor supply disincentives, decreasing with severity of impairments (Maestas et al 2013)
- benefits of DI receipt
  - reduced mortality for low-income beneficiaries (Gelber et al 2019 WP)
  - estimated WTP is greater for single than married applicants (Autor et al 2019, Norway Judge IV)
  - fewer adverse financial events (e.g. bankruptcy) (Deshpande et al 2019)
- judge IV designs
  - do error rate differences violate identification assumptions?

Targeting framework (*a la* Kleven & Kopczuk (2011)

- governments can choose:
  - degree of screening complexity
  - benefit level
  - eligibility threshold
- and tradeoff between:
  - Type Ib errors (false rejections, conditional on applying; rejecting a truly disabled applicant)
  - Type II errors (false acceptances, conditional on applying; awarding benefits to applicants who are not truly disabled)
  - Type Ia errors (the truly deserving don't apply)
- across public programs, disability insurance probably has the *most* complex screening process, with the *highest* benefit level
- should have few Type Ia errors here, so main tradeoff is between Type Ib/II errors

What counts as the "truth"? - this paper

- Health and Retirement Study (HRS), self-reports
  - **work impairment**: "an impairment or health problem that limits the kind or amount of paid work you could do"
  - temporary or not (less than three months)
  - prevents work altogether or not
- their definition of disability is stricter than SSA's
  - ignores that people can still earn up to a small amount
- effort to ensure timing of survey response is close to date of DI/SSI application
  - survey interview that is no more than 12 months after the application date
  - robustness to other close timing

What counts as the "truth"? - other papers

- other studies of disability error rates have also used surveys
  - Benítez-Silva, Buchinsky, and Rust (2004) — Health and Retirement Study (HRS)
  - Duclos 1995 — Family Expenditure Survey data (UK)
  - Low and Pistaferri (2015) — Panel Study of Income Dynamics (PSID)
- assume that the survey response has classical (mean-zero) measurement error
- alternative sources of "truth"?
  - medical assessment by independent team (Nagi 1969)
  - others?

Pros and cons of using survey responses as the "truth"

- (+) no concern that applicants will misreport to the government
- (-) but maybe will still misreport to remain consistent with DI application (but results are robust to timing of survey before or after DI application)
- (-) recall errors
- (-/+) individual is basing their response on their own (non-statutory) definition of work impairment/disability
  - (+) this may be more comprehensive than SSA's definition, which has to be based on observable information
  - (-) inter-personal comparability
  - (-) self-report of disability could be an ex-post rationalization of decision to leave the work force (endogeneity problem)
- the authors show self-reported disability is correlated with more objective or diagnostic measures — but would we expect anything different from these results...?

Why might women have higher false rejection rates than men? (conceptual framework)

1. (Demand, from applicants) women may have a **lower "pain" threshold** for labeling a work-limitation as severe enough to apply
2. (D) women's work limitations may be **objectively less severe**
3. (D) women may have a **lower cost of applying**
4. (Supply, from SSA) women may face **tougher standards** set by SSA
5. (S) women may **exhibit noisier signals** about the extent of their work limitation than men.

Theoretical framework

- Four equations with five unknowns:
  1. $L_i^* = \alpha_0 + \alpha_L F_i + \epsilon_i$ (true work limitation)
  2. $\overline{L}_i = \gamma_0 + \gamma_{\overline{L}} F_i$ (threshold for reporting work limitation)
  3. $\overline{A}_i = \overline{L}_i + \delta_0 + \delta_{\overline{A}} F_i$ (threshold for applying for DI)
  4. $S_i^* = L_i^* + \theta_{SSA} F_i + \zeta_i$ (noisy signal of work limitation)
- Three decisions:
  1. $L_i = \mathbb{1}\{L_i^* > \overline{L}_i\}$ - individual reports to be work-limited
  2. $A_i = \mathbb{1}\{L_i^* > \overline{A}_i\}$ - individual applies for DI
  3. $DI_i = \mathbb{1}\{S_i^* > \overline{L}_{SSA}\}$ - agent accepts DI application
- Parameters of interest and implications:
  1. $\gamma_{\overline{L}} < 0 \implies$ women have a lower work-limitation-reporting threshold
  2. $\alpha_L < 0 \implies$ women have less severe work limitations
  3. $\delta_{\overline{A}} < 0 \implies$ women have a lower opportunity cost of applying
  4. $\theta_{SSA} < 0 \implies$ SSA judges women more strictly
  5. $\sigma_\zeta^2(F) > \sigma_\zeta^2(M) \implies$ SSA receives less precise signal for women

Identification: vignette approach

- Example: "[Name] 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 [Name] limited in the kind or amount of work [he/she] could do?"
- Measuring parameters with vignettes
  - $\gamma_{\overline{L}}$: Are female respondents more or less likely to describe a character in a vignette as having a disability (pain threshold parameter)?
    - Less likely, indicating women have a higher pain threshold
  - θ_{SSA}: Combination of actual rejection of applications and, are all HRS respondents more or less likely to classify a female character in a vignette as having a disability?
    - Men are less likely, women are equally likely, indicating men are "tougher" on women

Problems with vignette approach

- Assumes respondents' view of characters mimics (1) SSA agents' view of applicants' work limitations, and (2) respondents' view of their own work limitation
- Vignettes capture threshold at which respondents would classify someone as work-limited, which may not be the same as the threshold at which they decide to report their own work limitation on the HRS
- Ignores particular context (geographical, labor market) individual is in in terms of ability to find a job

Why might women have higher false rejection rates than men? — demand evidence

1. (Demand, from applicants) women may have a **lower "pain" threshold** for labeling a work-limitation as severe —likely the opposite
   - after rejection, women are less likely to work, suggesting their limitations were truly severe
   - conditional on many observable characteristics, women are also less likely to apply for DI/SSI
   - men tend to be more lenient in marking a disability when evaluating vignettes
2. (D) women's work limitations may be **objectively less severe**
   - if anything, they are (insignificantly) *more* severe (structural estimate)
3. (D) women may have a **lower cost of applying**
   - if anything, application costs are (insignificantly) *higher* for women (structural estimate)

Why might women have higher false rejection rates than men? — supply evidence

4. (Supply, from SSA) women may face **tougher standards** set by SSA
   - when the vignette subject is a woman, she is less likely to be classified as disabled $\Rightarrow$ evidence in favor of this explanation if SSA reviewers have similar tendencies
   - support of this by structural estimates
5. (S) women may **exhibit noisier signals** about the extent of their work limitation than men.
   - actually, the noise of the signal is estimated to be *lower* for women (structural estimate)

Test for discrimination: Residual regression approach[1]

- control for a bunch of observables that might explain difference in outcomes
- assumes no omitted variable bias
  - can't control for unobservables
- assumes no *included* variable bias
  - otherwise, potential post-treatment bias with RHS variables that are functions of discrimination elsewhere or earlier
  - therefore assumes away any lateral or historic discrimination
- assumes no differential effect of observables by race — e.g. assumes symptoms of the same diagnosis don't present differently by gender
- captures only "in-market discrimination"
- distinguishing taste-based from statistical discrimination?

[1] Thanks Emma Rackstraw!

Other approaches from the discrimination literature

- **outcome test**: ex-post, are women more or less likely to work than men? (✓)
- **correspondence/audit study**: vignettes, but done with survey respondents, not doctors or reviewers (incorporated)
- **concordance test**: different outcomes with gender-concordant doctor or reviewer? (not done)
  - other study finds female patients with female doctors are more likely to go on to collect benefits than those with male doctors (no difference for male patients) (Cabral & Dillender 2021)

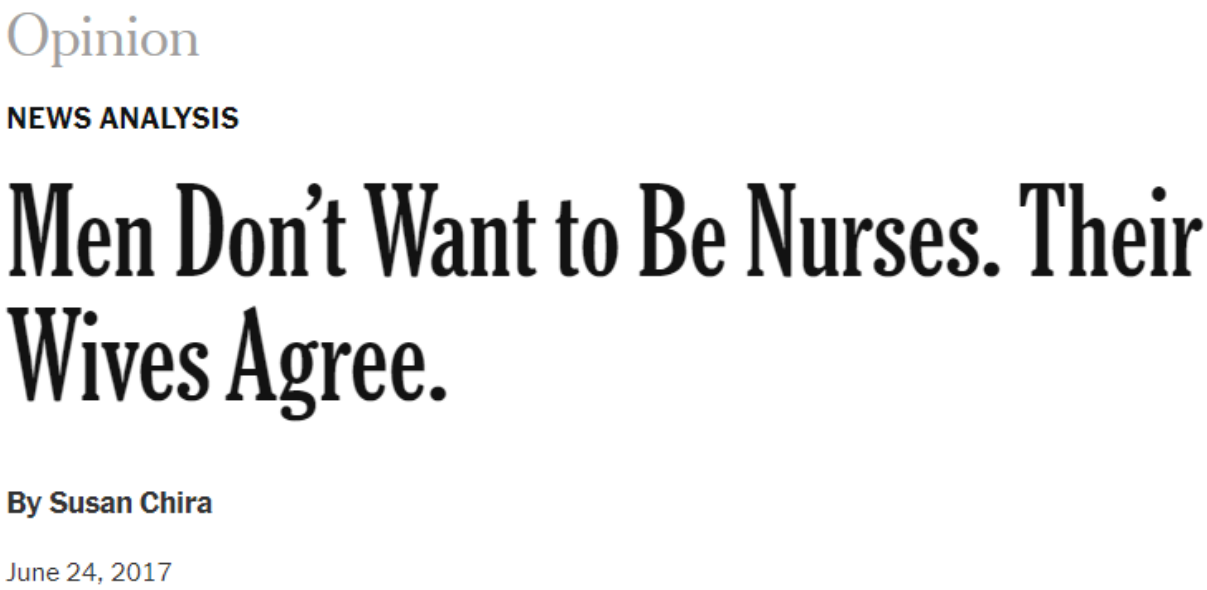
Reflection: the influence of social norms

- Relevant margin: judgment about ability to find other work — a fundamentally *social* concept
- social norms and judgements about work are *encoded in policy*
  - the social state has intentionally reduced labor supply based on social judgements (e.g. child labor prohibitions, compulsory schooling, retirement age, overtime and vacation regulations) (Saez 2021)
- social judgements also, perhaps inevitably, influence implementation of policies by *individuals with discretion*
  - DI/SSI case reviewers
  - doctors
  - lawyers?

Speculation about gender difference results

- women are traditionally *secondary wage earners*
  - BUT gender differences hold even after controlling for being the primary earner
- SSA has extra information not available to the econometrician (OVB)
  - e.g. daily activities, who takes care of the house, other activities, and so on
  - if women report doing more household activities (even if they are not capable of work), then this could perhaps explain the difference
- results are "consistent with the idea that women applicants are '**less believed**"'
  - echoes results in healthcare that women (Hoffmann & Tarzian 2021) and racial minorities (Warraich 2020 article) are "less believed" when it comes to pain

Real difference in willingness or ability to change jobs, based on work identity?



Opinion

NEWS ANALYSIS

Men Don't Want to Be Nurses. Their Wives Agree.

By Susan Chira

June 24, 2017

- "*Work is at the core of what it means to be a man, in a way that work is not at the core of femininity*" — Ofer Sharone, sociology prof at UMass Amherst
- many reasons why men don't take "pink-collar" jobs (both identity and expectations of others)

Policy implications

- insignificant difference at medical stage still highlights importance of gender-specific medicine
- gender-blind DI/SSI applications? probably infeasible
  - some gender-specific illnesses
  - for same diagnosis, symptoms may present differently between men/women
- AI/ML algorithms, with explicit objective to reduce gender differences?
- Paper's speculation: "It is also possible that the screening system evolves (with lags) to fit the gender composition of applicants, who were initially mostly men."
  - Testable: Did screening of men improve over time?
  - If this is the case, what to do?
- include more women in training examples?
- more incentives for gender equity in outcomes/error rates?
- update guidelines?