

WP 2012 – April 2020

Application and Award Responses to Stricter Screening in Disability Insurance

Mathilde Godard, Pierre Koning, Maarten Lindeboom

Abstract:

We examine the targeting effects of stricter screening in the Dutch Disability Insurance (DI) program induced by a major nationwide reform. The drastic 2003 “Gatekeeper Protocol” raised DI application costs and revealed more information about individual true ability to work. Discontinuity-in-Time regressions on administrative data show substantial declines in DI application rates (a 40% decrease in one year), with the largest decline occurring in difficult-to-diagnose impairments and less severe health disorders. This resulted in a more deserving pool of applicants. At the same time, those who stopped applying had worse health, worked less, and were more likely to be on UI and social assistance than workers who did not apply in the old system. There are no additional targeting gains at the point of the award decision, implying that changes in average health conditions of awardees were fully driven by self-screening and work resumption in the DI waiting period.

Keywords:

Disability Insurance, Screening, Targeting efficiency

JEL codes:

H2, I3

Application and Award Responses to Stricter Screening in Disability Insurance*

Mathilde Godard[†]

Pierre Koning[‡]

Maarten Lindeboom[§]

April 6, 2020

Abstract

We examine the targeting effects of stricter screening in the Dutch Disability Insurance (DI) program induced by a major nationwide reform. The drastic 2003 “Gatekeeper Protocol” raised DI application costs and revealed more information about individual true ability to work. Discontinuity-in-Time regressions on administrative data show substantial declines in DI application rates (a 40% decrease in one year), with the largest decline occurring in difficult-to-diagnose impairments and less severe health disorders. This resulted in a more deserving pool of applicants. At the same time, those who stopped applying had worse health, worked less, and were more likely to be on UI and social assistance than workers who did not apply in the old system. There are no additional targeting gains at the point of the award decision, implying that changes in average health conditions of awardees were fully driven by self-screening and work resumption in the DI waiting period.

JEL code : H2, I3

Keywords : Disability Insurance, Screening, Targeting efficiency

*This paper previously circulated under the title “Targeting Disability Insurance Applications with Screening”. We are grateful to Patrick Hulle for his support and initial work on the data, to Christopher Cronin, Eric French, Izabela Jelovac, Helge Liebert, Nicole Maestas, Owen O’Donell, Philip Oreopoulos, Nigel Rice, Jonathan Skinner, Joachim Winter, Joe Doyle, Michele Belot, Arthur Schram, Enrica Croda, Agar Brugiavini, Yue Li and participants of the NBER/CEPRA workshop on Aging and Health, the CRES-UPF Workshop on Disability Topics in Barcelona, the 2017 Essen Health Conference and participants at seminars at PSE, York, Florence, Venice, Rome, Melbourne, Nantes and Brussels for helpful comments and suggestions. The authors acknowledge financial support from Health Chair – a joint initiative by PSL, Université Paris-Dauphine, ENSAE and MGEN under the aegis of the Fondation du Risque (FDR). Mathilde Godard acknowledges the support of the EU under a Marie Curie Intra-European Fellowship for Career Development.

[†]University of Lyon, CNRS, GATE UMR 5824, F-69130 Ecully, France. Email : godard@gate.cnrs.fr

[‡]Leiden University, Vrije Universiteit, Amsterdam, IZA and Tinbergen Institute

[§]Vrije Universiteit, Amsterdam, Tinbergen Institute, Centre for Health Economics, Monash University, IZA

1 Introduction

In many OECD countries social expenditures on Disability Insurance (DI) schemes have been growing to levels that are considerably higher than any other social insurance scheme (OECD, 2010). To curb this trend, policymakers consider decreasing DI benefits or tightening eligibility criteria. While these policies may reduce moral hazard in the DI program by lowering disincentives to work, the risk is that disabled individuals are left under-insured. A way to overcome this trade-off between incentive costs and insurance aspects of the DI program is to improve the program’s ability to target deserving beneficiaries. Identifying meritorious from non-meritorious beneficiaries is imperfect as disability is difficult to observe and costly to verify. Improved targeting may be achieved by (i) stricter screening during the process that precedes DI applications, by e.g. using waiting times or imposing requirements such as reintegration obligations and (ii) stricter screening at the point of the award decision, e.g. through medical examinations.

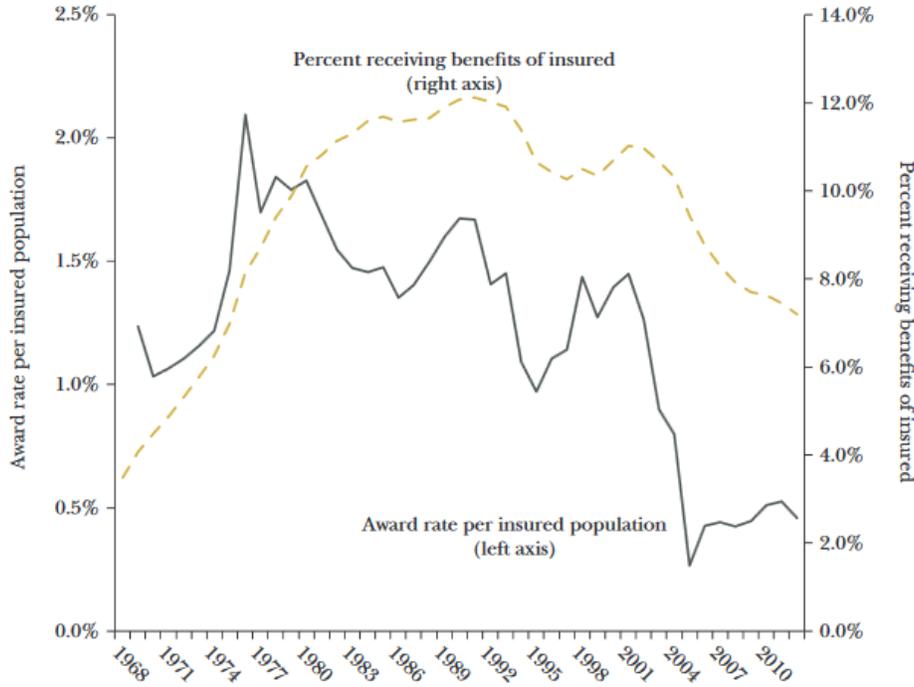
Such considerations played a role in the Netherlands, a country which until recently was characterized as the “most out of control program of the OECD countries” (Burkhauser et al., 2008). Despite a series of reforms in the late 1990s enrolment rates and inflow rates remained high until the early 2000s. This is illustrated by Figure 1, which suggests that the Gatekeeper protocol (GKP) of 2002 was the first reform that curbed this trend. This drastic reform introduced stricter screening both in the application and the award processes. It is generally considered as the most effective reform in the Netherlands that curbed DI inflow¹, and often cited as a possible blueprint for DI reform in Europe and the US (Autor, 2011; Fultz, 2015; OECD, 2009). However, beyond the drastic drop in DI inflow that was deemed “successful” we know little about the targeting effects of this nationwide reform. Did stricter screening induced by the GKP reform improve the provision of benefits to those who really needed them? This is what we investigate in this paper.

The GKP consisted of early interventions in the waiting period that precedes award determination, once a worker becomes sick. It essentially introduced a structured way to increase work resumption of workers during the waiting period, with more medical checks and more work resumption efforts of both workers and employers. These medical checks and work resumption efforts were documented in a reintegration report that was available to NSII (National Social Insurance Institute) examiners at

¹Among others, the reform contributed to the overall drop of DI enrolment from 11% in 2001 to 7.2% in 2012 (Koning and Lindeboom, 2015).

the point of award determination. Importantly, DI eligibility criteria upon award determination did *not* change with the reform, as medical examiners were explicitly instructed to maintain eligibility criteria of the old system.

Figure 1: Disability Insurance award and enrollment rate per insured worker in the Netherlands, 1968-2012



Note : The Disability Insurance award rate is the share of the insured population that started to receive disability payments in a given year. Source : UWV (2012). Borrowed from Koning and Lindeboom (2015).

As such, the GKP implied stricter screening both in the application and the award processes. We hypothesize that the stricter screening (i) revealed more information about the true ability to work, both during the application process and at the point of award determination and (ii) raised DI application costs. The combined effect of increased information and higher application costs on targeting is theoretically unclear. On the one hand, a reduction in the noise in the disability signal associated with stricter screening unambiguously improves the targeting of DI recipients. As remaining work capacity is better observed due to increased medical checks and reintegration efforts in the waiting period, able workers resume work and do not apply to DI. Further, at the point of award determination the information contained in the new reintegration report may also reduce classification errors made by NSII medical examiners. If workers take into account NSII allowance

rates in their application decisions, higher denial rates for the able will also induce self-screening among able workers (Parsons, 1991). On the other hand, stricter screening is associated with increased application costs. These increased application costs may deter sickness absence and DI claims, and its effect on the targeting of applicants is ambiguous. Hassles could deter either those most in need or least in need, depending on how potential applicants respond to these costs. If such deterrence effects are confined to workers with mild health impairments, the healthiest workers will screen themselves out, and targeting efficiency will improve. In contrast, if increased application costs disproportionately deter those with the most severe disabilities, for instance due to their higher mortality expectations and risk preferences, targeting efficiency will decrease. This is what Parsons (1991) refers to as perverse self-screening. Overall, the effect of stricter screening on the DI program ability to target deserving beneficiaries is ambiguous ex-ante.

In this paper, we use large-scale administrative registers on DI claims and award decisions as well as detailed administrative data on hospital admissions, mortality and labor market outcomes for the universe of Dutch individuals in 1999-2011. We empirically assess the effects of stricter screening due to the GKP reform in a Regression Discontinuity in Time (RDiT) framework. Our empirical strategy allows us to answer a number of key related questions. Who is screened out of the application process? Do the disabled *disproportionately* stop applying? Does the reform ‘overshoot’ – in the sense that part of the decline in applications comes from truly disabled workers? To address these related questions, we examine whether and how the GKP affected the composition of the pool of applicants and non-applicants, with respect to the type of impairment (for applicants only), health, future mortality and labour-market outcomes. We develop a simple method to uncover and characterize individuals who continued to apply after the reform (i.e., the “stayers”) and those that did not apply but would have done so without the reform (i.e., the “leavers”). We then examine whether there are further targeting gains at the award stage of the DI process. More specifically, we assess if the observed changes in the pool of awardees (if any) are primarily driven by work resumption and/or self-screening in the application phase versus classification changes in the final NSII award decision.

Our main findings can be summarized as follows. First, we find a 40% decline in DI application rates at the introduction of the GKP. The largest decline occurs in difficult-to-verify health conditions such as musculo-skeletal and mental impairments, which accounted for about two-third of all

pre-GKP applications. Second, we find that the pool of applicants under the new GKP regime has worse health and labor market outcomes. More specifically, we infer that the health of the individuals who continue to apply under the GKP is worse than the health of those who were deterred from applying. Altogether, these results suggest that the DI program is more efficient in targeting applications under the GKP protocol. At the same time, however, those who were deterred from applying are in worse health, have lower incomes and higher unemployment risks than employed workers who did not apply in the old system. In particular, we find non-negligible spillovers from the reform to the Unemployment Insurance (UI) scheme, with 14% of workers who were deterred from applying ending up in UI one year later. Third, we show that changes in the average health conditions of awarded applicants are almost fully driven by self-screening/reintegration of potential applicants during the waiting period, and not by changes in the award decision. A simple, back-of-the-envelope cost-benefit analysis reveals that savings of averted DI benefits by far exceed additional costs for NSII and employers.

While our analysis relates to a broad and mature literature about the effects of benefit generosity and DI eligibility on the labour supply of workers, two more specific strands are of most relevance.² The first strand of literature, which is mostly theoretical, addresses the implications of classification errors arising from imperfect information about the individual's true disability status.³ In this literature, Parsons (1991) was the first to distinguish Type I (incorrect denials) and Type II (incorrect awards) in the context of DI. Kleven and Kopczuk (2011) integrate the study of take-up with the study of rejections and awards. They characterize optimal programs when policy-makers chose the benefit level, eligibility rules, but also the rigour – or complexity – of screening. In particular, they analyze the implications of more rigorous screening on classification errors in DI, including incomplete take up. Finally, Low and Pistaferri (2015) develop a life-cycle model that incorporates classification errors in the DI award decisions. Using structural parameters that are derived from the Panel Study of Income Dynamics (PSID), they calibrate this model and conclude that the DI program in the US is characterized by substantial false rejections, but by fewer false acceptances.

²Empirical studies in this field of research usually point at labor supply responses to benefit generosity, eligibility criteria, elimination periods, and local benefit denial rates – see e.g. Borghans et al. (2014), Moore (2015), Karlström et al. (2008), Autor et al. (2015), Gruber and Kubik (1997), Autor and Duggan (2006), Campolieti (2002) and Maestas et al. (2013).

³A small empirical literature studies the size of classification errors in DI. Benitez-Silva et al. (2004) compare registered and self-reported disability statuses to infer the importance of award and rejection errors. In a similar vein, Croda et al. (2013) study country-differences in self-reported disability of DI benefit recipients to proxy the degree of targeting.

The second strand of the literature, which consists of recent empirical contributions, analyzes the impact of changes in the rigor of the screening process on DI application and inflow, with only limited attention to the question of targeting. A first subset of papers looks at the effects of stricter screening during the application period. Autor and Duggan (2014) and Autor et al. (2015) study deterrence effects of increases in elimination periods that precede public or private disability insurance claims. Markussen et al. (2017) evaluate the effects of compulsory dialogues for long term sick-listed workers with their employer and the family physician in Norway. Their results suggest a substantial reduction in sickness rates that are caused by the dialogues, both due to a notification effect and an attendance effect. Deshpande and Li (2019) explicitly look at the impact of increased application costs on targeting efficiency in the US. They analyze the effect of field office closures on the likelihood of SSDI applications and awards. They find a perverse effect of increased application costs: office closures lead to large and persistent reductions in the number of disability recipients, with the largest deterrence effects among applicants with moderately severe conditions, low education levels, and low pre-application earnings. A second subset of papers looks at the effects of stricter screening at the point of award decision. In particular, Liebert (2019) investigates the effect of introducing mandatory external medical reviews in the disability award determination process in a system that relied on treating physician testimony. He finds that external medical review reduces DI inflow by up to 23%, while it increases labor-market participation of claimants. Assuming no effect on self-screening, he argues that his reduced-form estimates can be interpreted as a net reduction in DI (Type II) award errors.

In the Dutch literature, Koning and Lindeboom (2015) have carefully described the rise and fall of the DI enrollment in the Netherlands in a long time frame (1960-2010). We also closely relate to De Jong et al. (2011), who evaluate the effect of intensified screening of sick-listed workers on sickness absence and DI application rates. The authors exploit a field experiment that was implemented at the start of the GKP reform in 2003. In two out of 26 Dutch regions a stricter version of the GKP protocol was implemented. In the standard procedure, DI applications were screened on paper only. In the treatment regions, however, caseworkers of the NSII were instructed to screen applications more stringently, either by contacting sick-listed workers or their employers. The results indicate that intensified screening in these regions decreased long-term sickness absence and DI application rates. The question of whether intensified screening influenced the targeting of the DI program is

only marginally addressed in the paper.⁴ The authors examine potential spillovers of intensified gate-keeping on UI inflow, but find none. In this respect, it is important to stress that their study examines changes in screening at the intensive margin, whereas we examine the effects of the broad implementation of the reform (the extensive margin).

Our paper provides some distinct contributions to the literature. We are the first paper that we are aware of to estimate the effect of stricter screening on the targeting of DI benefits, where stricter screening consists of both increased *stringency* (increased application costs) and *quality* (revealed information about the true latent disability status).⁵ We argue that screening increases *both* in the DI application and the award determination process, with potential information spillovers from the application to the award stage. Its final impact on the targeting of the DI program is ambiguous, but probably much less perverse than a sole increase in application costs (as in Deshpande and Li (2019)). Second, we use detailed administrative data on application and awards, as well as information on health, future mortality, and labour outcomes at the individual level. This allows us to characterize in detail the applicant and the awarded pool. We develop and implement novel measures to characterize the health of individuals who were deterred from applying. We infer from these changes the targeting effects of stricter screening induced by the GKP.

Besides these two points, the GKP constituted a major and drastic DI reform that is often cited as a possible blueprint for reform in Europe and the US (Autor, 2011; Fultz, 2015; OECD, 2009). In particular, early interventions at the onset of sickness that leverage the existing ties between the worker and the employer are often advocated by scholars and practitioners.⁶ To that extent, evaluating the targeting effects of the GKP is paramount.

The remainder of this paper is organized as follows. Section 2 provides more information about the DI reform and the expected effects of the reform on application and award responses. Data are presented in Section 3. Section 4 includes the empirical model and presents simple methods to separate screening effects during the waiting period from changes in the NSII assessment at the

⁴De Jong et al. (2011) did not use data on award decisions, the type of impairment, hospitalization history, or death.

⁵Using the terminology of Liebert (2019).

⁶For instance, Autor and Duggan (2010) advocates early intervention for it “leverages the existing economic and social ties between the worker, the employer, and his or her coworkers with the goal of accommodating and redressing the disability before it leads to labor force withdrawal”.

point of the award decision. Section 5 provides the estimation results, compares our findings with related evidence and presents simple, back-of-the-envelope, cost-benefit calculations. We end with conclusions in Section 6.

2 The Gatekeeper reform

2.1 Institutional background

Before we proceed by explaining the process toward entering DI, it is important to stress some important features of the Dutch DI system that are different from most other OECD countries. First, the DI program covers all income losses that result from both occupational and non-occupational injuries. This broad setup of the scheme largely explains the excessively high DI rates in the past (Koning and Lindeboom, 2015). Second, unlike e.g. the US, where individuals have no income during the DI waiting period, Dutch workers receive sick pay up to 100% of their net earnings during the full waiting period of one year. Employers are financially responsible for these costs. The degree of disability is determined by considering the loss in earnings capacity due to the impairment and is measured in seven classes ranging from 15% to a 100% loss. DI benefits are 70% of pre-disability earnings when fully disabled and correspondingly lower when partially (15%-80%) disabled. DI benefits entitlements are unlimited and do not have continuing eligibility conditions, which makes it a more attractive alternative over other social insurance schemes such as UI or Social Assistance benefits.⁷ As a result, the system was relatively attractive for workers, as it provided little barriers to report sick and did not provide strong financial incentives to resume work quickly. From the employer point of view, DI was also an attractive substitute pathway into UI, as it avoided substantial firing costs.⁸ Taken together, these characteristics of the Dutch DI scheme laid the ground for high disability levels by the turn of the century. Not surprisingly, the explicit aim of the GKP in April 2002 was to reduce the number of DI applications. Another goal of the reform was to improve the targeting of the DI program.

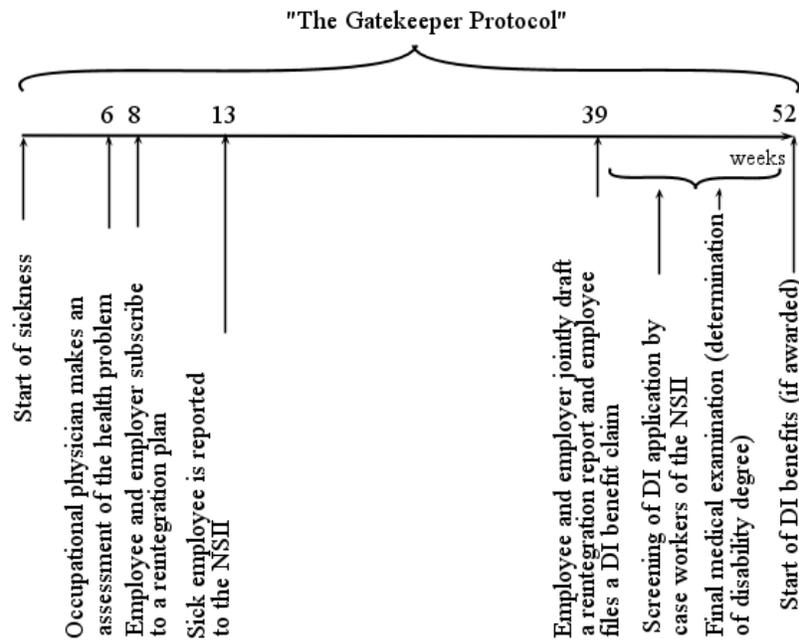
The GKP came into effect on April 2002, and the first DI applications under the new GKP protocol arrived at the NSII offices as of January 2003. The GKP stipulated the responsibilities of the

⁷For more detail about the exact calculation of the DI benefits and the DI entitlement period we refer to De Jong et al. (2011) or Koning and Lindeboom (2015)

⁸Burkhauser et al. (2008) argues that the Dutch DI system was often used – with consent of both employer organisations and unions – as an alternative to UI and social assistance. In some sectors where early retirement schemes were absent, DI benefits were explicitly used as a route for early retirement (Kerkhofs and Theeuwes, 1999).

worker and the employer for sickness spells lasting six weeks or longer. These responsibilities are represented in Figure 2, which shows different steps that need to be taken during the sickness spell. After six weeks of absence, a first assessment of medical cause, functional limitations and a work resumption prognosis has to be drafted by a medical doctor from an occupational health service agency. On the basis of this assessment the employer and employee draft together an accommodation and rehabilitation plan in which they specify an aim (resumption of work in the current job or somewhere else, whether accommodated working conditions are required, etc.) and the steps needed to reach that aim. This reintegration plan should be ready by the eighth week of sickness. It is binding for both parties, and one party may summon the other when considered negligent. After 13 weeks of absence the employer should report the sick employee to the NSII.

Figure 2: Schematic representation of the process toward entering DI



Note : Borrowed (and extended) from De Jong et al. (2011).

The worker can file a DI application if s/he has not resumed work before 39 weeks of sickness and, when the claim is approved, collects benefits after the end of the waiting period, which was initially one year.⁹ Benefit claims are only considered admissible by the NSII if they are accompanied by the

⁹The waiting period was later extended to two years. The first applications under the two-year waiting period arrived in 2006 at the NSII offices.

rehabilitation plan and an assessment as to why the plan has not (yet) resulted in work resumption. If this procedure is not followed, the employer is obliged to continue providing sick pay for some additional months rather than having the worker transferred to disability benefits.¹⁰ This implies a strong financial incentive for employers to focus their attention at the onset of sickness, when the opportunities for recovery and work resumption are probably most substantial. The employer and the worker are jointly responsible for reintegrating the worker back into his or her old job or into a new job commensurate with his/her limitations. If the employee consistently rejects reasonable offers and accommodations, however, the employer may stop paying sickness benefit and, eventually, fire the employee. When sufficient reintegration efforts have been undertaken by both the employer and the worker, the DI claim is processed. In a final examination, medical doctors and job experts from the NSII jointly determine the presence of impairments, the consequences for the earnings potential, and the corresponding disability benefit level. If awarded, the applicant starts receiving benefits in week 52.

While employers were already financially responsible for the continued payment of wages during absence, the GKP implied a substantial increase in costs related to the reintegration of their sick-listed workers. In effect, it meant that employers started bearing sick-pay costs; the costs of the health service agency¹¹; and the costs of reintegration measures and work accommodations to facilitate work resumption. The employer was also financially responsible for continued wage payments after the end of the waiting period if DI claims were deemed inadmissible. Smaller employers have particularly responded to the GKP reform by buying extended insurance products of private insurance companies that are labeled as “Gatekeeper-proof”; they do not only insure employers against the risk of continued wage payments, but also provide services to adhere to all the formal responsibilities that are described in the GKP. In particular, they provide reintegration services and send out caseworkers that take care of all requirements that need to be met in order to file a DI application. Based on a survey among about 2,500 employers with less than 100 employees, De Jong et al. (2014) find that about 80% of them had bought private insurance, and 88% of these employers opted for coverage that was Gatekeeper-proof. The additional wage premium of extended coverage varied between 0.23 and 0.43 percent of the total wage costs. These amounts correspond to 7.5 % and 15

¹⁰Since the start of the GKP, employers have been sanctioned only in rare occasions (Koning and Lindeboom, 2015).

¹¹In the Netherlands, private health service agency play a key role when adhering to the minimum requirements that are set in the reintegration process of workers.

% of the insurance premiums, respectively.¹²

2.2 Expected effects of the reform

The introduction of the GKP implied strong changes in the responsibilities for both the employer and the sick-listed worker in the waiting period that preceded the DI claims assessment. Employers who did not yet comply to the new requirements now needed to increase their reintegration activities. Through increased tests and reintegration activities, employers and workers may have learned more about (i) the worker’s true ability to resume work and (ii) the required level of reintegration efforts for a DI claim to be considered admissible by the NSII. As a result, we expect an increase in work resumption rates during the waiting period and a decrease in the number of DI claims that are filed. Importantly, the impact of increased requirements on work resumption is likely to differ across impairment types. For “easy-to-verify” impairment types, the cause, the treatment and prospects for work resumption are clear. These may include acute and curable health problems that last for a limited time period, for which – if work resumption is possible – effective reintegration plans are relatively easy to implement. Alternatively, these easy-to-verify impairments may include severe conditions (e.g. last stage cancers) for which prospects of work resumption are null. In such situations, no additional medical tests are required and work resumption is not an option. For these easy-to-verify impairments (either mild or severe), the value added of the protocol is then probably limited. In contrast, for “difficult-to-verify” impairments (e.g. mental conditions such as burnout and depression, or musculo-skeletal diseases), the cause is more difficult to identify, the severity more difficult to assess, and whether or not reintegration measures are effective is uncertain. For such difficult-to-verify diseases, the protocol is potentially more effective in increasing work resumption rates.

Changing the perspective to that of the worker, the key question is whether the introduction of the GKP changes his/her decision to enter the DI application process. The GKP reform affects this decision in two ways. First, it increases the worker’s costs of participating in the DI waiting period. The protocol requires more verification tests, more visits to the doctor, more reintegration efforts and more documents to support the claim. These costs are likely to deter workers from entering the DI process.¹³ Application costs may vary according to disability status, and could even be larger

¹²Based on comparisons of reintegration activities for employers with and without insurance coverage, De Jong et al. (2014) argue that buying insurance does not reduce the potency of the GKP reform.

¹³These costs may also be non-monetary, as there may be an (emotional) hassle to go through the application

for disabled workers because of higher mortality expectations, a higher degree of risk aversion or because they are put under high strain by reintegration efforts (Parsons, 1991). Second, the GKP reform decreases the noise of the disability signal of applicants upon award determination. With the reintegration report at hand, the NSII obtains more information about the true disability status of the applicant. While formal eligibility standards remain constant and medical examiners are explicitly instructed to maintain the DI criteria of the old system, the protocol may thus reduce classification errors. This is especially relevant for workers with conditions that are difficult to verify, where the disability signal is noisier. If workers take into account NSII allowance rates in their application decisions, higher denial rates for the able with difficult-to-verify impairments will induce higher self-screening among these workers (Parsons, 1991). This effect will work in the other direction for truly disabled individuals with impairments that are difficult to verify who now face higher acceptance rates.

Based on these insights, we make a number of predictions that guide the empirical analysis. First, we expect the reform to decrease DI applications and awards among able workers, especially among those with difficult-to-verify disorders. This is due to a combination of (i) higher application costs and (ii) a reduction in the noise about their true ability to work, both in the application period and at the point of award determination. For these reasons, we expect both increased self-screening and increased work resumption among able workers. Second, the effects of the reform on DI applications and awards for disabled workers are ambiguous. Stricter screening may increase the award rate (conditional on applying) of the disabled, but increased application costs may also induce incomplete take-up. As a result, the final effect on application and award rates is ambiguous. Increased application costs could even induce disabled applicants to disproportionately screen themselves out of the application process (perverse self-screening in the terminology of Parsons (1991)). In that case, the targeting of DI applicants would worsen. We also expect most of the action to come from disabled individuals applying for difficult-to-verify diseases, though, because this is where a reduction in noise in the disability signal matters the most. Overall, whether the GKP increases the targeting of DI benefits is essentially an empirical question.

process.

3 Data

3.1 Data setup

The data used in this paper come from several administrative sources. Key information on DI applications from 1999 onwards is obtained from the *DI application register*, which is administered by the NSII. The register contains information on DI applications by impairment type (musculo-skeletal, mental, cardiovascular, nervous, respiratory, endocrine or “other” non-categorized disorders) and the date and outcome of the award decision (award or reject). For awarded claims we have information on the degree of disability measured in categories (15-25%, 25-35%, 35-45%, 45-55%, 55-65%, 65-80% and 80-100%). Award decisions are typically made within a maximum of two months after application.¹⁴ While we observe the date of the award decision, we do not observe the application date, nor the starting date of the sickness spell. Because of this (unobserved) delay between the application and the award decision, we cannot precisely determine around the January 2003 cutoff – when the first DI applications under the new GKP protocol arrived at the NSII – whether an applicant was treated according to the GKP protocol. In particular, we cannot rule out that some award decisions observed after January 2003 stemmed from workers who applied in December 2002 (under the old regime). For this reason and for reasons argued later on, we will employ a research design that uses a donut hole around the cutoff date. Note that for the purpose of exposition, we align the award and application date and refer to the “DI award decision month” as the “DI application month” in what follows.

We merge the *DI application register* with several administrative records of Statistics Netherlands covering the entire population. The tax record measures earnings (in 2010 Euros) from 1999 onwards. Based on this information we consider a worker employed if s/he has any positive earnings in a given calendar year.¹⁵ The tax register also provides yearly information on disability receipt, welfare or unemployment benefits. Demographics (month-year of birth, sex, nationality and place of residence as ZIP codes) are obtained from the municipality registers. The Dutch national hospital register (LMR) contains data on inpatient and day-care patients of all general and academic hospitals as well as specialized hospitals in the Netherlands from 1995 to 2005. Unfortunately, the LMR data do not include specialized hospitals for mental care or psychiatric clinics. As a consequence,

¹⁴See <https://www.uwv.nl/particulieren/ziek/ziek-wia-uitkering/beoordeling-beslissing-wia-aanvraag/index.aspx>

¹⁵Note that according to this definition, an employed individual can combine both earnings and DI benefits.

hospitalization spells due to mental disorders are poorly measured. For each individual in the LMR data we observe the admission and discharge date, as well as the main diagnosis. Following the International Classification of Diseases (ICD-9) codes, we categorize the hospital admissions into eight diagnosis types: mental, musculo-skeletal, neoplasms, endocrine, cardio-vascular, nervous, respiratory and other diseases. To proxy the individual’s health status we construct indicators for an individual being hospitalized (all-cause or cause-specific) in the three years preceding the current year (for a DI applicant, this corresponds to the three-year period preceding his/her year of waiting period). For each individual, we also construct the Charlson Comorbidity Index that aggregates information over multiple hospitalization spells since 1995 (see Data Appendix B for more information on the construction of this index). Finally, we merge our data with the death register. With this information we can track future mortality outcomes of applicants and non-applicants.

Our sample of interest consists of prime-age individuals (aged 25-65) who were employed in the previous year.¹⁶ We exclude cases with conflicting information in the different databases¹⁷, and individuals who apply more than once during the period under study, as we cannot be sure which application is linked to the final award decision. We also exclude individuals from two out of the 26 regional NSII offices that participated in the field experiment at the start of GKP De Jong et al. (2011).¹⁸ Our analysis is restricted to the pre-GKP period (2001 and 2002) and the first two years the GKP started affecting new applicant cohorts (2003 and 2004).¹⁹ For each month in 2001-2004, we take all individuals for which we observe an award decision that month (‘applicants’) as well as a 1% random sample of individuals for which we do not observe any award decision that month (“non-applicants”). Our final sample consists of a maximum of 95,338 individuals in each month, measured over the years 2001-2004.²⁰

¹⁶We make this restriction since the GKP only applies to employed individuals. Potential selection into employment in response to the GKP reform is not a concern here, as our RDiT regressions effectively use information for 2002 and 2003 (see next Section). Imposing the restriction that workers treated according to the GKP (observed in 2003) were employed in 2002 (pre-reform) is thus not likely to be endogenous.

¹⁷Such cases include applicants who are initially rejected, but who nevertheless appear as DI recipients in the tax register (11.3% of rejected applicants). Conversely, we exclude individuals who are awarded DI benefits according to the NSII records, but who do not appear as a DI benefit recipient according to the Tax records (6.3% of awarded applicants).

¹⁸We stated earlier that the NSII implemented a field experiment at the moment of the introduction of the GKP. In two out of the twenty six regions screening procedures were stricter than in the other twenty four regions. De Jong et al. (2011) used this field experiment to examine the effect of marginal changes in screening intensity.

¹⁹Our RDiT regressions (see section 5) effectively use information from 2002 and 2003.

²⁰Note that due to our sampling design, individuals can appear more than once in our final sample.

3.2 Summary statistics

Summary statistics for our sample, which are computed by taking a 1% random sample of applicants and non-applicants in each month, are shown in Table 1. The table is split into the pre-GKP period (2001-2002, column (1)) and the post-GKP period (2003-2004, column (2)). The table shows that characteristics such as age, gender, ethnicity remain stable over time. In contrast, the monthly DI application rates sharply decline: from 0.07% prior to the GKP reform (2001-2002) to 0.04% in post-reform years (2003-2004). Similarly, monthly DI award rate as a percentage of the total worker population decrease from 0.04% to 0.02% over the same period.

Figure 3 shows how the number of DI application and awards have evolved over 1999-2004 at the monthly level. The figure suggests an immediate impact at the time when the first DI applications under the new system arrived at the offices of the NSII. We also observe substantial end-of-year effects in application numbers in all years in our sample.²¹

Turning to labor-market characteristics, Table 1 shows that gross earnings slightly increase over time (from €26,262 to €27,389), and so is the fraction of those on unemployment benefits (from 3% to 5%). Similarly, the health index and the fraction of individuals who were hospitalized in the previous three years slightly increase over time. The shares by type of hospitalization stay rather constant over time. As expected, the fraction of individuals who have been hospitalized for mental reasons is low. This is due to the absence of specialized hospitals for mental care in the LMR database. Five-year mortality rates also stay stable over time (and are less than one percent in both periods).

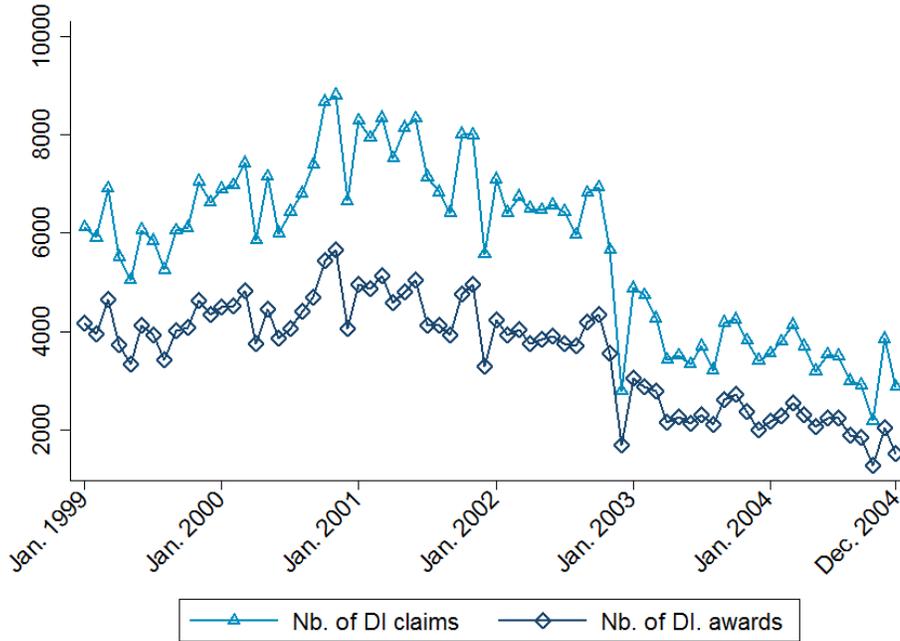
²¹This is due to the closing of the regional offices around Christmas in the last one or two weeks of December. Note, however, that the drop in the number of DI applications and awards in December 2002 seems larger than in other years. We discuss this point in detail in Section 4.

Table 1: Summary statistics of the data (1% random sample)^(a).

	Pre-Gatekeeper 2001-2002 (1)	Post-Gatekeeper 2003-2004 (2)
Demographics		
Age	38.37 (9.46)	39.02 (9.53)
Male	0.56 (0.50)	0.56 (0.50)
Native	0.82 (0.38)	0.82 (0.39)
Labour-market^(b)		
Employed	0.95 (0.21)	0.95 (0.22)
Gross earnings (2010 euros)	26,262 (22,261)	27,389 (23,770)
UI recipient	0.03 (0.18)	0.05 (0.22)
Welfare recipient	0.02 (0.13)	0.02 (0.13)
DI		
DI (monthly) application rate	0.0007 (0.026)	0.0003 (0.018)
DI (monthly) award rate (uncond.)	0.0004 (0.021)	0.0002 (0.014)
Health status		
Health index	0.026 (0.308)	0.032 (0.342)
Hospitalized in the three previous years ^(c)	0.156 (0.363)	0.161 (0.367)
Hospitalization type (among hospitalized) ^(d)		
Musculo-skeletal disorders	0.176 (0.381)	0.170 (0.376)
Neoplasms	0.051 (0.219)	0.052 (0.223)
Cardiovascular diseases	0.077 (0.267)	0.080 (0.271)
Mental disorders	0.009 (0.093)	0.008 (0.087)
Endocrine problems	0.013 (0.113)	0.013 (0.114)
Nervous disorders	0.054 (0.227)	0.058 (0.234)
Dead within five years	0.008 (0.088)	0.008 (0.087)
Number of observations	2,095,914	2,186,913

Notes : ^(a) To compute these summary statistics we take a 1% random sample of applicants and non-applicants in each month (2001-2004); ^(b) Our sample excludes individuals not employed in the previous year (see Section 3). Labour-market characteristics are measured at the yearly level; ^(c) Hospitalized in t-1; t-2 or t-3; ^(d) Note that only the main hospitalization types are listed here. Individuals can be hospitalized for several reasons in the previous three years, so lines could add up to more than 100%.

Figure 3: Disability Insurance application and award inflow in the Netherlands, 1999-2004



Note : The DI application inflow is the number of prime-age individuals for which we observe an award decision in a given month; the DI award inflow is the number of prime-age individuals who were awarded DI benefits in a given month.

Table 2 breaks the sample into the group of applicants and the group of non-applicants, for both the pre- and the post-GKP period. A comparison between column (1) and (3) of this table suggests that the reform induced substantial changes in the composition of the pool of applicants. Under the GKP regime, applicants are slightly older, more likely to be men and non-native, have much lower earnings, are less employed, and more often unemployed or on welfare. These applicants are also in worse health upon application (as measured by the Charlson index and past hospitalization rates). This decline in health is a first indication that the GKP protocol works as an efficient screening device to target deserving DI applicants. From this alone we cannot rule out, however, that some of the truly disabled screen themselves out of the application process. To investigate this, we look at changes in the health in the pool of non-applicants – see columns (2) and (4). A simple look at these columns suggests that the health conditions of non-applicants do not change with the introduction of the GKP. Taken together, these descriptive results thus provide first suggestive evidence that the introduction of the GKP increased the DI system ability to screen out able workers during the

application process – without deterring the disabled to apply.²²

Table 2: Summary statistics of DI applicants and non-applicants on the month of (potential) application^(a), before and after the introduction of the Gatekeeper protocol

	2001-2002		2003-2004		Diff-in-diff (5)
	Applicants (1)	Non-appl. (2)	Applicants (3)	Non-appl. (4)	
Demographics					
Age	43.30 (9.82)	38.37 (9.46)	43.81 (9.88)	39.02 (9.53)	-0.151*** (0.044)
Male	0.41 (0.49)	0.56 (0.50)	0.48 (0.50)	0.56 (0.50)	0.075*** (0.002)
Native	0.79 (0.40)	0.82 (0.38)	0.77 (0.42)	0.82 (0.39)	-0.019*** (0.002)
Labour-market^(b)					
Employed	0.90 (0.30)	0.95 (0.21)	0.83 (0.37)	0.95 (0.22)	-0.063*** (0.001)
Gross earnings (2010 euros)	26,267 (22,319)	30,894 (22,263)	20,778 (18,263)	27,391 (23,772)	-1,703.6*** (104.7)
UI recipient	0.11 (0.32)	0.03 (0.18)	0.18 (0.38)	0.05 (0.22)	0.044*** (0.001)
Welfare recipient	0.03 (0.18)	0.02 (0.13)	0.04 (0.19)	0.02 (0.13)	0.006*** (0.001)
Health status					
Health index	0.18 (0.88)	0.03 (0.34)	0.27 (1.14)	0.03 (0.34)	0.089*** (0.002)
Hosp. past three years ^(c)	0.34 (0.47)	0.16 (0.36)	0.40 (0.49)	0.16 (0.37)	0.048*** (0.002)
Dead within five years	0.03 (0.17)	0.01 (0.10)	0.04 (0.20)	0.01 (0.10)	0.011*** (0.000)
Number of observations	146,134	2,094,450	75,691	2,186,129	4,502,404

Notes: ^(a) To compute these summary statistics we take – in each month – the full sample of applicants (that month) as well as a 1% random sample of non-applicants (that month). ^(b) Our sample excludes individuals not employed in the previous year (see Section 3). Labour-market characteristics are measured at the yearly level. ^(c) Hospitalized in t-1; t-2 or t-3. Column (5) presents the difference-in-differences mean, i.e. for each variable Y ($Y_{applicant,after} - Y_{applicant,before} - (Y_{non-applicant,after} - Y_{non-applicant,before})$).

Finally, Table 3 sheds further light on applicant characteristics, as well as on award decisions. The

²²These results should be interpreted with caution, however. First, simple differences in means between the pre- and post-GKP periods could reflect secular time trends. Second, the group of non-applicants is much larger than the group of applicants. In particular, the share of deterred applicants in the non-applicant group is low, which makes it difficult to detect deterrence effects in the means of the group of non-applicants.

top panel shows shares by impairment type among applicants at either side of the reform date. Before the reform (column (1)), about 65% of DI applicants apply for musculo-skeletal or mental impairments. These are impairments where the applicant’s true ability to work is typically difficult to verify, and where the GKP protocol is potentially more effective. After the reform, the share of these difficult-to-verify diseases declines by about eight percentage points, to about 57%. The bottom panel of Table 3 reports the award rates of applicants, first for all impairment types, and next for each impairment separately. Before the reform, almost 60% of applicants are awarded DI benefits. Conditional award rates, however, vary substantially across impairments: they are typically lower for difficult-to-verify diseases than easy-to-verify ones.²³ After the start of the GKP, we see some changes in the award rates of applicants. In particular, the (conditional) award rate for all impairments increases from 59.9% to 61.6%, which – assuming no changes at the point of award determination – is indicative of a more deserving applicant pool. Interestingly, the increase in the award rate for all impairments is driven by increases in award rates for workers applying for mental or “other” non-categorized impairments. Again, assuming no changes at the point of award determination, these results suggest that (i) the GKP was efficient in screening out able workers during the application process and (ii) targeting improved mainly due to individuals with difficult-to-verify impairments who disproportionately stop applying. However, as argued earlier, assuming no changes in the award determination process may be problematic. After all, the reintegration report may provide NSII more information about the applicant’s true ability to work at the point of award determination. To that extent, we cannot interpret changes in the (conditional) award rate as sole changes in application behaviour. We return to this in the next section.

²³For instance, 82.6% of claims related to nervous disorders are approved. Similarly, 79.2% and 77.5% of claims related to (resp.) respiratory diseases and endocrine problems are approved.

Table 3: Summary statistics of DI applicants on the month of application, before and after the introduction of the Gatekeeper protocol (GKP).

	DI applicants	
	Pre-GP	Post-GP
	2001-2002	2003-2004
	(1)	(2)
Disability impairment type (% of applicants)		
Musculo-skeletal disorders	0.288 (0.453)	0.285 (0.451)
Mental disorders	0.366 (0.482)	0.288 (0.453)
Cardiovascular diseases	0.048 (0.213)	0.059 (0.236)
Nervous disorders	0.037 (0.190)	0.046 (0.210)
Respiratory disorders	0.015 (0.123)	0.018 (0.132)
Endocrine problems	0.011 (0.106)	0.014 (0.116)
Other	0.234 (0.424)	0.290 (0.454)
Award rate among DI applicants, by type of impairment		
All impairment types	0.599 (0.490)	0.616 (0.486)
Musculo-skeletal disorders	0.676 (0.468)	0.580 (0.494)
Mental disorders	0.592 (0.491)	0.676 (0.468)
Cardiovascular diseases	0.818 (0.386)	0.776 (0.417)
Nervous disorders	0.826 (0.380)	0.805 (0.396)
Respiratory disorders	0.792 (0.406)	0.770 (0.421)
Endocrine problems	0.775 (0.418)	0.727 (0.446)
Other	0.412 (0.492)	0.514 (0.500)
Degree of disability (among awarded)		
Fully disabled (disability degree > 80%)	0.516 (0.500)	0.531 (0.499)
Number of applicants	146,134	75,691

Notes : (1) To compute these summary statistics we take – in each month – the full sample of applicants (that month).

4 Empirical strategy

This paper assesses the effects of stricter screening due to the GKP reform on the targeting of the DI program. Our empirical strategy follows three successive steps. First, we investigate the overall effect of the increased screening by comparing the numbers of DI applications just before and after the start of the GKP. To do so, we use a Regression Discontinuity in Time (RDiT) specification that exploits the sharp discontinuity in January 2003, when the protocol started to affect new DI claims. Second, we employ a similar RDiT estimation approach to uncover changes in the composition of DI applications, including the shares of impairment types, the average health conditions (hospitalizations and mortality rates) and average labor market outcomes. We develop a simple method to characterize the health of workers who continued to apply for DI benefits and workers who were deterred from applying. This is instrumental when looking at shifts in the composition of the applicant pool, but also when looking at incomplete take-up of benefits by disabled workers. Finally, we change our perspective to examine the effects of the reform on award decisions. More specifically, we test whether the information provided by the reintegration report at the point of NSII award determination induced additional gains in the targeting of DI benefits. In what follows, we explain these three steps in detail.

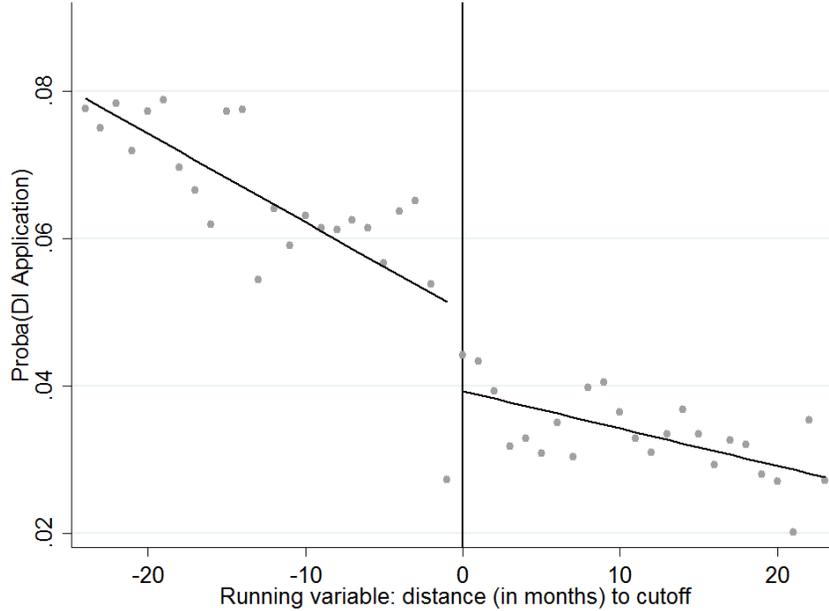
4.1 Specification of the RDiT model

Throughout our analysis, we compare the number and composition of DI applications and awards just before and after the start of the GKP. We use Regression Discontinuity in Time (RDiT) specifications (Hausman and Rapson, 2017) that exploit the sharp discontinuity in January 2003. Because of data restrictions (see previous section) and reorganizations that went together with the reform, we extend this specification by dropping observations that are close to the cutoff point. As such, we estimate “donut” RD specifications with various holes, dropping individuals within one to four months of the cutoff.

To gauge the potential importance of the reform, Figure 4 plots the average monthly DI application rate for the period 2001-2004, which is a four-year (symmetric) window around the cutoff. Observations close to the cutoff point are not dropped at this point. The vertical line indicates the exact month at which the GKP started to affect DI claims. A quick look at the graph shows a sharp and substantial drop in DI application rates at the cutoff, which is much larger than discontinuities

at other January placebo cutoffs – see Appendix Figure A1. Similar eyeball tests in the Appendix Figure A2 with a donut hole of one month on each side of the cutoff and with flexible polynomial forms also suggest that there is a sizable and robust treatment effect.

Figure 4: Regression Discontinuity plot for DI application rate, linear fit



Note: For each month, we take all applicants that month and a 1% random sample of non-applicants that month to compute the monthly DI application rate – see Section 3. Thus, DI monthly application rates should be divided by 100 to reflect the average rate in the population.

We specify the linear RDiT model with a total bandwidth of 13 months on each side of the threshold and a donut bandwidth of one month on each side of the threshold. As we will discuss later on, alternative specifications with varying bandwidths and various donut holes produce similar results. More specifically, we estimate the following linear model:

$$Y_{it} = \alpha + \beta T_t + \gamma_1 t + \gamma_2 t * T_t + \sum_{k=2}^{12} \delta_k M_{kt} + \theta X_i + \epsilon_{it} \quad (1)$$

where Y_{it} is the outcome measure of interest of individual i in month t . While our initial interest lies in the effect of the reform on DI application rates, subsequent analyses will consider the effect on the composition of the pool of applicants and non-applicants. T_t is an indicator that takes the value of 1 after December 31, 2002. We allow the linear time trend to differ at each side of the cutoff

by including t and the interaction term $t * T_t$. The parameter γ_1 thus captures any pre-existing linear trend in DI application, while γ_2 captures the linear trend after the policy change. M_{kt} with $k = 2, \dots, 12$ is a set of month indicators and δ_k the associated month-of-year fixed effects.²⁴ X_i is a vector of time-invariant characteristics (age, gender and ethnic background). Standard errors are clustered along both the month-year and the individual dimensions.²⁵

The parameter β measures the impact of the reform on the outcome measure of interest. DI applications are filed after 39 weeks (nine months) and therefore the first applications under the new GKP reform arrived at the regional offices of the NSII in January 2003. The RDiT framework of Equation (1) uses a window of 13 months on either side of the cut-off date (December 31, 2002). Therefore, applications arriving at the regional offices of the NSII in 2003 are the first cohorts who reported sick from April 2002 up to the start of 2003. β thus reflects (i) changes in the decision of the worker and employer not to enter the waiting period after 13 weeks of sickness; (ii) increased work resumption, conditional on the worker entering the waiting period; and (iii) changes in the decision of the worker to stop the application process prematurely. The decision to drop out of the application process/waiting period may be due to unexpected costs and worker and employer learning more about the worker’s ability to resume work and his prospect of being awarded DI benefits.

It is important to stress two interpretative issues that are inherent to our RDiT design. To start with, the RDiT model estimates the initial effects of the GKP, rather than learning effects that took more time to materialize. For the first application cohorts after January 2003, the use of reintegration report in the waiting period and for DI determination was new. The NSII, employers, workers and the occupational health services had no experience with reintegration reports prior to that date. Related to this point, our estimation method does not capture any compositional effects the reform may have had on the total pool of employed workers. While it is likely that the GKP reform may have deterred employers from hiring workers with high sickness absence and DI risks, any of such labor demand effects are not relevant for our analysis that uses worker observations in the first year of the new protocol. We therefore expect parameter β to largely reflect mechanisms

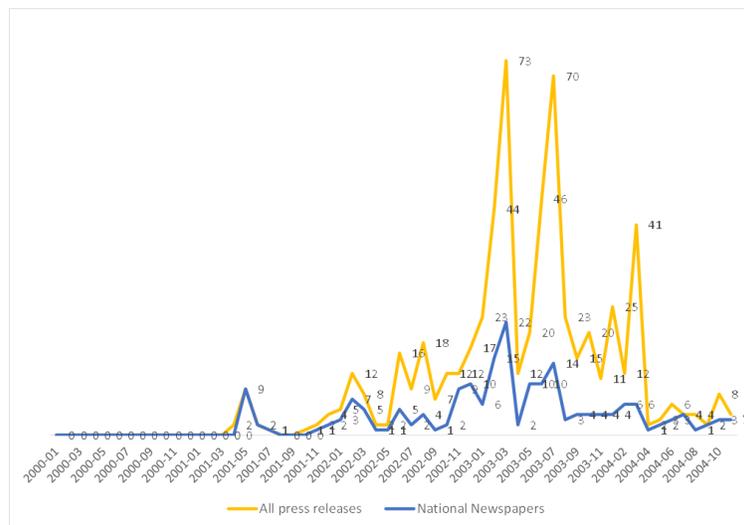
²⁴Controlling for month-of-year fixed effects seems crucial given the substantial end-of-year effects (see Figure 4).

²⁵Lee and Lemieux (2010) argue that in a regression discontinuity framework where the treatment-determining variable is discrete, the observations should be clustered at the level of the right-hand side variable. Here, additional interest lies in a “primary” dimension of clustering (i.e. the individual). For that reason, we use two-way clustering following Cameron et al. (2011).

(ii) and (iii).

Possible anticipation effects may be an issue in the current context. The relevant time window for anticipation would be between December 2001 – i.e. the moment the reform was announced – and April 2002. This in turn implies an increase in DI application rates 39 weeks ahead, i.e. in September-December 2002. A simple look at Figure 4 shows no such increases, suggesting that such anticipation effects were limited. The absence of (substantial) anticipation effects is also confirmed by Figure 5, which shows that the press coverage of the new GKP protocol in national newspapers was low when the reform was proposed but not decided upon yet.²⁶ Between December 2001 and March 2002, the months between the announcement and the implementation, we see a small increase in press coverage. Altogether, we see this as suggestive evidence that employers and workers did not anticipate the new law. It then took until the first months of 2003 to have substantial press coverage; this rise was triggered by the substantial drop in DI applications that was reported by the NSSI one year after the enactment of the law.

Figure 5: Press Coverage of Gatekeeper Protocol (2000-2004)



Note : Article numbers referring to the Gatekeeper Protocol are derived from the LexisNexis Academic database. This database includes all newspaper and magazine articles since 2000.

²⁶The GKP was voted in the Dutch House of Representatives on July 5, 2001, but it took until Nov. 29, 2001 to have it passed on by the senate. All details of the law were made public in December 2001. Details of the law can be found in the following document (in Dutch) : https://www.eerstekamer.nl/behandeling/20011218/publicatie_wet_8/document3/f=w27678st.pdf.

4.2 Characterizing workers who are (not) deterred from applying

While our initial focus is on the reform effects on (total) DI applications, we now turn to the composition of the pool of applicants. As a first step, we simply re-estimate Equation (1) on the sample of applicants, taking as an outcome variable any given characteristic (i.e. age, gender, health, future mortality and labor-market outcomes). To better assess the targeting effects of the GKP reform, we develop a simple non-parametric method to characterize those who continue to apply for DI benefits, and those who would have applied under the old scheme but were deterred from applying in the new one. We denote the former group as the “stayers” and the latter group (the counterfactual group) as the “leavers”.²⁷ As we cannot directly observe those who are deterred from applying by the reform, we have to infer their characteristics indirectly. Below we explain the method in detail.

For inference on the characteristics of leavers and stayers, our key assumption is that of monotonicity in responses to the reform (no “defiers”). This means that while the reform may have induced some workers to stop applying, it did not induce workers who would normally not have applied to start applying after the reform. We believe that this assumption is highly plausible in the Dutch context of pervasive moral hazard and notoriously high DI inflow rates (prior to the GKP reform).

To formalize our approach, we first define Y as an indicator for “leavers” ($Y=1$) and “stayers” ($Y=0$). In line with our earlier notation, we also define β_Y as the causal reform effect on the DI application rate and β_X as the causal reform effect on the exogenous applicant characteristics X . The following defining equations then hold:

$$E_Y(E(X|Y)) = (1 - \beta_Y) E(X|Y = 0) + \beta_Y E(X|Y = 1) \quad (2)$$

and

$$\beta_X = E(X|Y = 0) - E(X) = \beta_Y (E(X|Y = 0) - E(X|Y = 1)) \quad (3)$$

Rewriting equation (3) yields:

$$E(X|Y = 1) = E(X|Y = 0) - \frac{\beta_X}{\beta_Y} \quad (4)$$

²⁷In the instrumental variables literature, the leavers can be referred to as the compliers, the stayers as always-takers, and non-applicants in both periods as never-takers (Angrist and Pischke, 2009).

In the empirical analyses we replace the expectation operators by their respective sample averages. With the estimates of β_Y , β_X and the sample average of X among the “stayers” ($\bar{X}_{Y=0}$), Equation (4) identifies the average characteristics of those who are deterred from applying ($\bar{X}_{Y=1}$). Depending on the type of variables at hand, we can assess whether the remaining pool of applicants has become more deserving due to the reform.

4.3 Further selection in the award determination

With constant eligibility standards and no changes in classification errors upon award determination, a straightforward way to infer the targeting effects of stricter screening would be to look at changes in conditional award rates. Targeting efficiency improves if the probability of acceptance (conditional on application) increases. This is the empirical strategy followed by Deshpande and Li (2019).²⁸ However, as argued earlier, NSII examiners have more information at the point of award determination under the GKP, which may lead to a reduction in award classification errors. To that extent, we cannot simply infer the targeting effects of the reform from changes in conditional award rates. While self-screening and/or increased work resumption of able applicants typically increase conditional award rates, the reduction in the noise of the disability signal has ambiguous effects. A reduction in false negatives increases conditional award rates, whereas a reduction in false positives decreases conditional award rates.²⁹ In our context, changes in conditional award rates thus reflect both changes in application behavior – that is, changes in the composition of the pool of applicants – and changes in the award decision process that stem from changes in the disability signal. In light of these arguments, our aim is to separate the effects of the GKP reform on the composition of applicants from effects on awards due to a reduction in the disability signal. For the latter effect, it is important to remember that the GKP reform did not change DI eligibility standards.³⁰

We address this issue by combining information on changes in application rates and changes in conditional award rates. More specifically, we define a dichotomous variable H to indicate the health status of applicants, with $H=0$ for applicants in good health and $H=1$ for applicants in bad health. We define p as the share of applicants in bad health. For the award outcome we use the binary

²⁸In their context, office closures increase application costs but there is no new information available for the NSII at the point of award determination.

²⁹Liebert (2019) provides explicit conditions under which the inflow reduction can be interpreted as a higher relative reduction in award errors than rejection errors.

³⁰As a matter of fact, the medical examiners who had to assess the DI claims were explicitly instructed to use the eligibility requirements of the old regime (see also De Jong et al. (2011)).

indicator A , with $A=1$ for an approved application and $A=0$ for a rejected application. Applicants in bad health have conditional award probabilities equal to $a_{H=1}$, versus $a_{H=0}$ for applicants in good health, respectively. Then we can write:

$$E(H) = (1 - p) 0 + p = p \quad (5)$$

and

$$E(H|A = 1) = \frac{p a_{H=1} 1 + (1 - p) a_{H=0} 0}{p a_{H=1} + (1 - p) a_{H=0}} \quad (6)$$

For notational convenience, we define $\delta = \frac{a_{H=1}}{a_{H=0}}$ as the relative conditional award rates for those in bad health. Then we can write Equation (6) as:

$$E(H|A = 1) = \frac{p \delta}{(1 - p) + p \delta} \quad (7)$$

This yields the following expression for the relative conditional award rate for those in bad health:

$$\delta = \frac{E(H|A = 1)}{1 - E(H|A = 1)} \frac{1 - E(H)}{E(H)}. \quad (8)$$

The above expression shows that the relative conditional award rate δ can be inferred from averages of H among the samples of applicants (\bar{H}) and awardees ($\bar{H}_{A=1}$). We can run RDiT regressions to estimate the effect of the reform on H for both samples, and eventually infer changes in δ . Changes in δ will be indicative of additional targeting at the award stage. Increases in δ indicate that the conditional award rate of those in bad health increases relative to that of those in good health. This is indicative of improvements in the targeting of DI benefits upon award decision. If δ remains constant, however, the conditional award rate of those in bad and good health move in tandem. Any changes in the composition of the pool of awarded applicants are then fully driven by self-screening and/or work resumption during the application process.³¹ So while Deshpande and Li (2019) make inferences about targeting efficiency on the basis of changes in conditional award rates, in our context – with changes in application costs and the disability signal – we exploit changes in *relative* conditional award rates.

³¹Estimates for (changes in) δ can also be stratified across different applicant types. For instance, it is conceivable that increases in the disability signal are particularly relevant for workers with conditions that are difficult to verify.

5 Results

5.1 Effects on the number of DI applications

Table 4 presents the main RDiT and Donut RDiT estimates for DI applications using Equation (1). Standard errors are in parentheses and the implied percentage change (relative to the pre-reform sample average) is in brackets. All models in Table 4 use a bandwidth of 13 months on each side of the threshold.³² The first panel shows the RD estimate, whereas the four other panels present Donut RDiT estimates with varying widths of the “hole” around the cutoff. Removing observations within one month from the cutoff increases the treatment effect substantially and reduces the standard errors. In contrast, dropping additional months around the cutoff – up to four months on both sides – does not substantially change the results. In line with the arguments made earlier, these findings suggest that both data limitations (the discrepancy between the application and the award decision dates) and organizational changes introduce noise in the data close to the threshold. As it seems, delays in the award procedure increased in December 2002, where end-of-year effects are common anyway. This in turn would lead to underestimation of the treatment effect in the conventional RDiT specification. We therefore proceed with Donut RDiT specifications without observations within one month of the threshold.

The size effect of our preferred specification implies a 39.3% decrease in DI applications as a result of the introduction of the GKP. This corresponds to the observed drop displayed in Figure 4 and is consistent with the claim of Koning and Lindeboom (2015) that the reform was one of the major contributors to the decrease in DI applications since the turn of the century. This drastic decline in DI applications is likely to stem from two sources (see discussion in Section 4). First, effective reintegration measures may have increased work resumption rates. Second, workers may have decided to drop out during the waiting period. This may be due to unexpected high costs associated with reintegration efforts and worker and employer learning more about the disability status of the worker and the prospect of being awarded DI benefits. Ex-ante self-screening (the decision not to enter the waiting period at all) is probably limited, as the the reform was new to all parties involved.

³²Varying bandwidths do not influence the results (see Figure A3 in Appendix).

Table 4: (Donut) Regression Discontinuity estimates of the effect of the Gatekeeper Protocol on DI application rates.

	DEPENDENT VARIABLE
	DI application
	Coeff. [% change] (se)
RD estimate	-0.009* [-15.5%] (0.005)
Mean of dependent variable	0.058
Number of observations	2,537,250
Donut RD dropping those within one month of the cutoff	-0.024*** [-39.3%] (0.003)
Mean of dependent variable	0.061
Number of observations	2,257,510
Donut RD dropping those within two months of the cutoff	-0.027*** [-44.3%] (0.002)
Mean of dependent variable	0.061
Number of observations	2,070,088
Donut RD dropping those within three months of the cutoff	-0.027*** [-44.3%] (0.002)
Mean of dependent variable	0.061
Number of observations	1,881,742
Donut RD dropping those within four month of the cutoff	-0.029*** [-47.5%] (0.003)
Mean of dependent variable	0.060
Number of observations	1,693,411

Notes : (1) Each model uses a bandwidth of 13 months on each side of the threshold and includes a linear trend that is flexible on either side of the cutoff, as well as month-of-year dummies. All estimates include controls for sex, age and ethnic background (see Equation 1). (2) Standard errors are clustered both at the individual level and at the month-of-year level. (3) Estimates in brackets are presented in percentage change of the baseline DI application for the sample of interest.

Table 5 reports estimation results of the reform effect by type of impairment. The top row reproduces the estimate obtained with our preferred specification for all impairments. Subsequent rows of Table 5 show that the most sizable declines are observed for difficult-to-verify impairment types (musculo-skeletal, mental and non-categorized (“other”) disorders). Our results suggest a 41.3% decrease in applications for difficult-to-verify impairment types versus a 32.6% decrease for easy-to-verify ones. Bearing in mind that mental and musculo-skeletal disorders account for up to 65 percent of all applications (see Table 3), the total decline in applications primarily comes from mental and musculo-skeletal disorders.³³

³³We observe a sizable decline (45.9%) for respiratory disorders. These disorders, however, account for only 1.5%

Table 5: The effect of the Gatekeeper Protocol on DI application rates. Donut-RD estimates.

	DEPENDENT VARIABLE
	DI application
	Coeff. [% change] (se)
<i>Panel A: All impairments</i>	-0.024*** [-39.3%] (0.003)
<i>Panel B: Difficult-to-verify impairments</i>	-0.022*** [-41.3%] (0.002)
Musculo-skeletal	-0.008*** [-45.8%] (0.001)
Mental disorders	-0.008*** [-38.1%] (0.001)
“Other” disorders	-0.005*** [-38.5%] (0.001)
<i>Panel C: Easy-to-verify impairments</i>	-0.002*** [-32.6%] (0.000)
Cardiovascular diseases	-0.001*** [-35.8%] (0.000)
Nervous disorders	-0.001*** [-24.3%] (0.000)
Respiratory disorders	-0.0004*** [-43.7%] (0.000)
Endocrine problems	-0.0002** [-30.2%] (0.000)
Nb. of obs.	2,071,474

Notes : (1) Each line presents the estimated coefficient associated with the treatment (GKP reform) for a different outcome (i.e. DI application for a specific impairment type). (2) Donut RD dropping those within one month of the cutoff. We use a bandwidth of 13 months on each side of the cutoff and include a linear trend that is flexible on either side of the cutoff, as well as month-of-year dummies. All estimates include controls for sex, age and ethnic background (see Equation 1). (2) Standard errors are clustered both at the individual level and at the month-of-year level. (3) Estimates in brackets are presented in percentage change of the baseline DI application rate for the sample of interest.

5.2 Characterizing workers who are (not) deterred from applying

A key question is whether the observed decline in applications stems from disabled or able individuals. We argued earlier that a characterization of workers who continue to apply for DI benefits (the “stayers”) as well as those who are deterred from applying (the “leavers”) is informative about the targeting effects of the GKP reform. As the group of “leavers” cannot be observed directly, we developed a simple non-parametric method in Section 4.2 to indirectly infer the characteristics of this group. Using the RDiT estimates of the reform effect on DI applications and on various worker characteristics, we can characterize both “leavers” and the “stayers”.

of all DI applications.

Table 6: The effect of the Gatekeeper Protocol on the composition of the pool of DI **applicants** – Donut-RD estimates

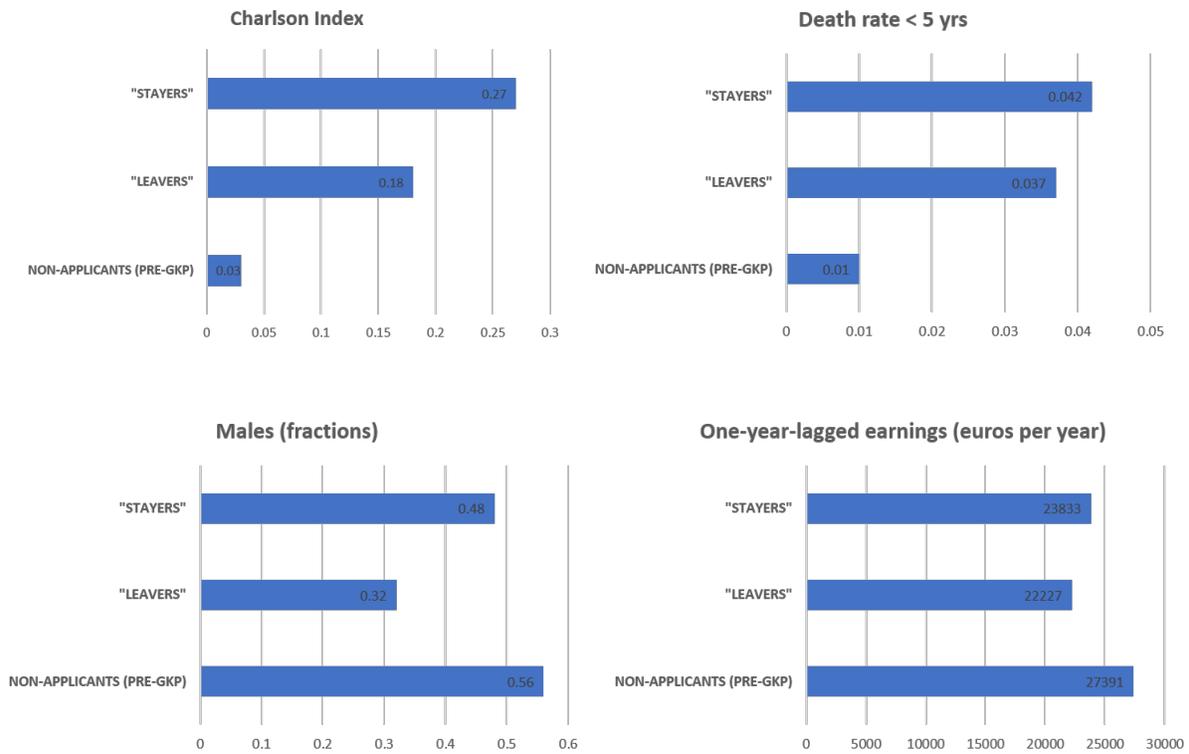
	Applicant pool Change in mean/share [% change]	Implied difference “stayers” vs. “leavers”
<i>Panel A: Health and Mortality</i>		
Health (Charlson) index	0.036*** [+18.6%] (0.002)	0.087*** (0.024)
Death rate within five years	0.002 [+6.0%] (0.002)	0.004 (0.005)
<i>Panel B: Socio-Demographics</i>		
Male	0.068*** [+16.5%] (0.007)	0.166*** (0.023)
Native	-0.103*** [-13.0%] (0.007)	-0.025 (0.016)
Age		
Young (25-34)	0.027*** [+11.7%] (0.004)	0.066*** (0.013)
Prime age (35-49)	-0.026*** [-5.7%] (0.007)	-0.063*** (0.017)
Senior (50+)	-0.001 [-0.32%] (0.005)	-0.003 (0.013)
One-year lag earnings (gross, in 2010 euros)	658** [+2.9%] (294)	1,606** (755)
<i>Panel C: Share of impairment types</i>		
Difficult-to-verify impairments	-0.024*** [-2.7%] (0.008)	-0.058*** (0.008)
Easy-to-verify impairments	0.024*** [+20.0%] (0.008)	0.058*** (0.008)
Nb. of individuals	107,297	107,297

Notes : (1) Each line in column (1) presents the estimated impact of the GKP reform on the average value of a characteristic (e.g. average health index) or the proportion of applicants with a given characteristic (e.g. female). Each line in column (2) presents the implied difference between the “leavers” and “stayers” for a given characteristic. (2) Donut RD dropping those within one month of the cutoff. We use a bandwidth of 13 months on each side of the cutoff and include a linear trend that is flexible on either side of the cutoff, as well as month-of-year dummies (see Equation 1). (3) Standard errors in column (i) are clustered both at the individual level and at the month-of-year level. Standard errors in column (ii) are obtained from seemingly unrelated regression, and clustered at the month-of-year level.

The results of this exercise are reported in Table 6 for a set of health variables (panel A), socio-demographic variables (panel B) and for impairment types (panel C). The first column of the table presents RDiT estimates of the impact of the reform on the average value of a characteristic in the

pool of applicants (e.g. average health index) or the proportion of applicants with a given characteristic. The second column presents the implied difference between the “leavers” and “stayers” for each characteristic. Figure 6 complements these estimates with plots of the implied averages of “leavers” and “stayers” for a subset of variables, together with value averages of the initial group of non-applicants that are directly observed in the data.

Figure 6: Implied averages for the Charlson index, death rate, fraction of males, and earnings for workers that are deterred from applying (“leavers”), those who continue to apply (“stayers”) and (pre-GKP) non-applicants.



The combined results in Panel A of Table 6 and Figure 6 suggest that stricter screening induced by the GKP reform increased the efficiency in targeting deserving applicants. We find a significant and positive effect of the reform on the Charlson index of applicants (a 18.6% increase, significant at the 1% level), indicating a worsening of health in the pool of the applicants. The top panel of Figure 6 on health proxies is particularly informative on the targeting effects of the GKP reform. A positive difference in means between “leavers” and “stayers” suggests that stricter screening is more efficient in targeting deserving applicants. Accordingly, the first panel of Figure 6 shows that the implied average of the index for “leavers” (0.18) is about 30% lower than for those of “stayers”

(0.27). We also find a small and insignificant effect of the reform on the 5-year mortality rate of applicants (see column (1) of Table 6). The implied difference in mortality rates between “leavers” and “stayers” is accordingly small. This insignificant effect on future mortality rate is most likely due to the lack of power (the average age is 43 years, see Table 1), with low mortality rates (<5%).

Turning to Panels B and C of Table 6, we observe important shifts in the applicant pool in terms of age, gender, earnings and impairment types. The emerging picture is that stricter screening induced by the GKP reform was efficient in targeting deserving applicants. To start with, results in column (1) show that the applicant pool under the GKP consists of more men, and less prime-aged workers. Figure 6 shows implied averages for the gender of “leavers” and “stayers”. The share of women among the “leavers” is about 16 percentage-points higher than the share of women among the “stayers”, indicating that women disproportionately stopped applying. Similar changes occur for prime-aged workers, who are over-represented among the “leavers”. Interestingly, we find that low-earning workers disproportionately stopped applying due to the GKP (see also Figure 6). This observed shift in earnings is not straightforward to interpret. Several interpretations are consistent with this fact, with opposite implications in terms of targeting. It could be that reintegration activities were particularly successful for these low-earning workers, or that the increase in application costs disproportionately discouraged them to apply for DI benefits. Supposing low-earning workers are on average in better health than other applicants, the fact that they disproportionately screen themselves out of the application process could improve the targeting of DI applicants. We do find support for this in the data.³⁴ Recall that prior to the GKP reform, DI was also functioning as a long-term program for workers who were less employable (typically with low earnings), rather than being restricted to those having substantial health problems. We return to this issue in Section 5.4, where we put our findings in a broader perspective. Finally, and consistent with earlier results in Table 6, the share of individuals with hard-to-verify impairments decreases in the applicant pool (a 2.7% decrease). To examine these findings in more detail, we reproduced column (1) of Table 6 for females and males separately (not shown). The results of these analyses indicate that the effects for age and impairment types are primarily driven by females³⁵, but that health (proxied by the

³⁴In particular, we find that the (pre-GKP) allowance rate for applicants in the bottom of the earnings distribution is lower than the allowance rate for other applicants, suggesting that these applicants were on average in better health than other applicants.

³⁵The decline in the share of difficult-to-verify impairments is -0.029 and statistically significant for females and -0.006 and insignificant for males. For prime-aged workers, the effects are -0.037 and -0.007 for females and males, respectively.

Charlson index) worsens for both female and male applicants.

Although stricter screening induced by the GKP reform was efficient in targeting deserving applicants, it most likely also induced some individuals with worse health and weaker labor market positions (possibly disabled) to no longer apply. This is suggested by the top panel of Figure 6, which shows that the average Charlson index and 5-year mortality rate of “leavers” are substantially higher than for the initial group of non-applicants. To investigate such effects in more detail, we change our perspective to changes in the pool of non-applicants. The idea is that the disabled who stopped applying due to the GKP reform (if any) ended up in the pool of non-applicants. With no remaining capacity to work, they would most likely have turned to UI insurance, and/or eventually to welfare benefits. To indirectly assess the presence of disabled individuals in the pool of non-applicants, we look at one-year leads in hospitalization, death rates, and labour-market outcomes.³⁶ Worsening along these dimensions should reflect a increased presence of disabled individuals in the non-applicant pool.

We first compare pools of non-applicants, before and after the reform, using our RDiT strategy. Table 7 shows the results of this exercise, while Figure 7 complements these with the implied averages of non-applicants under the counterfactual without the GKP reform (“stayers”) and those who are deterred from applying due to the reform (“leavers”) for a subset of variables.³⁷ Results in column (1) of Table 7 suggest that the pool of non-applicants worsens along all dimensions (health, death, and labour-market outcomes). From Figure 7, we infer that “leavers” who are added to the pool of non-applicants are in worse health. Hospitalization rates among that group are twice the rate of the initial group of non-applicants and mortality rates are as much as eight times higher. Perhaps more strikingly, “leavers” have substantially lower incomes, lower employment rates and much higher UI benefit incidence than the group of non-applicants. More specifically, UI rates of those who stop applying is 14%. The implied quasi-elasticity of UI receipt due to less DI application is about 0.36.³⁸ This suggests that there are non-negligible spillovers to UI from the reform. At the

³⁶We look at future outcomes to capture the transition to UI and welfare benefits. We focus on one-year-lead variables, as longer-run outcomes could be interpreted as both a composition effect and the ‘pure’ (behavioural) effect of foregone DI application (and possibly foregone DI benefits).

³⁷For this we conduct similar regressions and calculations for the sample of non-applicants as for the sample of applicants.

³⁸Recall that the overall effect of the reform on DI applications amounted to a decrease of 39.5%, whereas the implied probability of UI receipt is 14% for those who no longer apply. Consequently, about 36% of former DI applicants enters into UI.

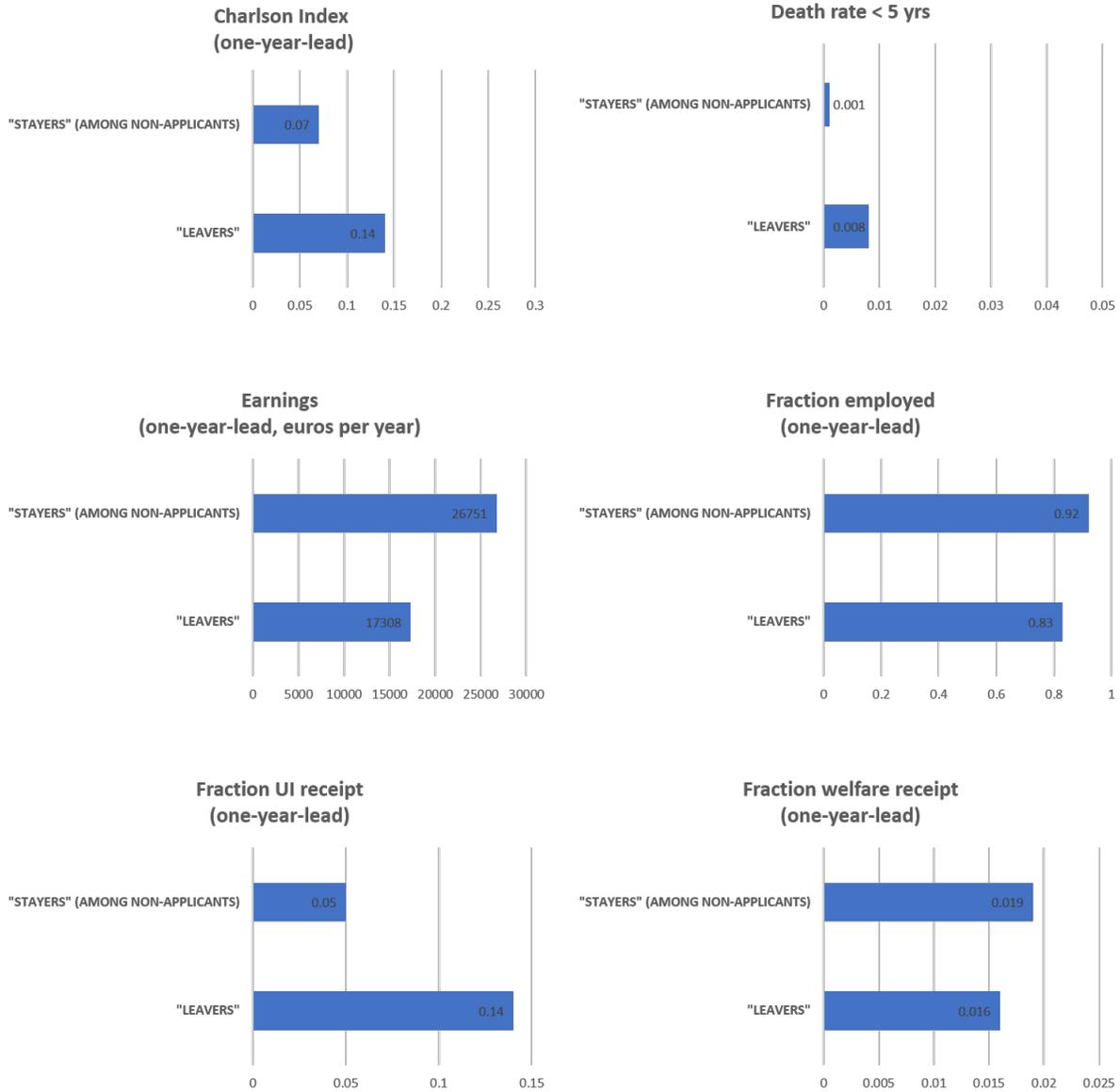
same time – and on a more positive note – a substantial part of the this group (83%) are one year later still at work.

Table 7: The effect of the Gatekeeper Protocol on future outcomes (one-year-lead) of **non-applicants** – Donut RD estimates.

Future (one-year lead):	Change in mean [% change]	Implied difference “leavers” vs. non-participants
<i>Panel A: Health and Mortality</i>		
Hospitalization (any type)	0.004*** [+5.9%] (0.001)	0.071*** (0.010)
Death rate	0.0002*** [+16.2%] (0.0001)	0.007*** (0.003)
<i>Panel B: Labor-market outcomes</i>		
Earnings (gross, 2010 euros)	-438*** [-1.6%] (54)	-9,442*** (896)
Employment (fraction)	-0.004*** [-0.4%] (0.001)	-0.086*** (0.023)
UI receipt (fraction)	0.005 ^μ [+9.5%] (0.003)	0.086*** (0.021)
Welfare receipt (fraction)	0.001* [+5.0%] (0.000)	-0.003 (0.006)
Number of individuals	2,150,213	2,150,213

Notes : (1) Each line in column (i) presents the estimated impact of the GKP reform on the average value of a characteristic (e.g. average health index) or the proportion of applicants with a given characteristic. (2) Donut RD dropping those within one month of the cutoff. We use a bandwidth of 13 months on each side of the cutoff and include a linear trend that is flexible on either side of the cutoff, as well as month-of-year dummies (see Equation 1). (3) Standard errors in column (i) are clustered both at the individual level and at the month-of-year level. Standard errors in column (ii) are obtained from seemingly unrelated regression, and clustered at the month-of-year level.

Figure 7: Implied averages the Charlson index, death rate, earnings, fraction employed and UI and welfare receipt for workers that are deterred from applying (“leavers”) and those who continue not to apply (“stayers”). Inference is based on applicant-data.



5.3 Further targeting gains at the point of the award decision?

The picture that emerges from our findings so far is that the GKP reform induced substantial compositional changes in the pool of DI applicants. The evidence on health outcomes suggests stricter screening was generally efficient in targeting deserving applicants. From this alone, however, we cannot assess the overall implications of the reform on targeting efficiency in terms of DI

awards. As argued in Section 2, the GKP provided medical examiners with information on the reintegration activities that were planned in the waiting period. While these examiners were explicitly instructed to maintain the DI eligibility criteria, more information at the point of the award determination may have reduced classification errors and further improved targeting efficiency. To examine these effects we follow the approach that was unfolded in Section 4.3. More specifically, we interpret changes in the relative conditional award rate of those in bad health, δ . We proxy health H by a set of dichotomous variables: (i) a dichotomized Charlson index which captures previous hospitalization history (which is equal to one if the index is larger than or equal to one, zero otherwise) and (ii) a dummy indicating whether the individual died within the next five years.³⁹

Table 8 presents the findings that follow from this approach. Columns (1) and (2) show the averages of health proxies for applicants and awardees before the GKP reform, respectively. Based on these averages, column (3) infers the relative conditional award rate (δ) of individuals with bad health. As expected, Table 8 makes apparent that health conditions of awarded applicants are worse than of the total pool of applicants. This reflects the fact that the allowance rate for applicants in bad health is higher than the allowance rate for individuals in good health who apply. This is mirrored by a value of δ (the conditional award rate of individuals in bad health relative to those in good health) higher than one. Specifically, prior to the GKP protocol, those in bad health – as defined by the Charlson index – have 34.9% higher chance of obtaining a DI benefit than those in good health. This rate is substantially higher for difficult-to-verify impairments (panel B) than for easy-to-verify impairments (panel A), suggesting that medical examiners have more added-value when screening difficult-to-verify conditions.

To see whether the GKP reform induces further targeting gains at the point of award decision, we interpret changes in δ throughout the reform. An increase in δ indicates that the targeting of DI benefits upon award decision improves, as the conditional award rate of those in bad health increases relative to that of those in good health.

³⁹Obviously, the set of health indicators H that we consider is limited. Medical examiners probably have access to a richer set of health information from medical checks and reports on reintegration efforts. Our variables at hand are a noisier measure of the underlying health information accessible to the medical examiner. We therefore expect that any changes observed in the final award stage are an underestimate of the true effect.

Table 8: Inference on targeting efficiency from DI award rates: value averages for applicants and awardees and relative conditional acceptance rates (“ δ ”), before and after the reform.

Variable:	Pre-Gatekeeper Protocol			Post-Gatekeeper Protocol		
	Applicants (average) (1)	Awarded (average) (2)	Relative conditional award rate (δ) (3)	Applicants (average) (4)	Awarded (average) (5)	Relative conditional award rate (δ) (6)
Panel A: All impairments						
P(Health (Charlson) Index ≥ 1)	0.083	0.109	1.349	0.109	0.141	1.334
Dead within five years	0.033	0.045	1.39	0.042	0.056	1.38
Panel B: Difficult-to-verify impairments						
P(Health (Charlson) Index ≥ 1)	0.058	0.075	1.32	0.079	0.102	1.32
Dead within five years	0.027	0.037	1.39	0.035	0.049	1.39
Panel C: Easy-to-verify impairments						
P(Health (Charlson) Index ≥ 1)	0.269	0.290	1.11	0.297	0.322	1.13
Dead within five years	0.074	0.085	1.17	0.080	0.094	1.20

Notes : (1) We use data in a bandwidth of 13 months on each side of the cut-off.

When comparing the results for the pre-GKP period (columns (1)-(3)) to those for the post-GKP period (columns (4)-(6)), we observe health declines for both applicants and awardees. However, δ stays virtually unchanged before and after the reform (see columns (3) and (6)). This means that changes in the average health condition of the pool of awardees are fully driven by self-screening/work resumption during the waiting period, without additional gains in screening out healthy applicants in the final award decision. As a note of caution and in all fairness, we use relatively crude measures for health H and therefore may fail to fully pick up changes in the award stage due to a reduction in the disability signal. At the same time, it is reassuring to see that the estimates of δ also stay constant for stratified samples of applications with easy-to-verify and hard-to-verify diseases. This suggests no major changes in (unmeasured) health conditions for a given level of the applicants Charlson index.

5.4 Our findings in perspective

Our analysis suggests that the main mechanisms driving compositional changes among DI awardees are increased self-screening and increased work resumption among potential applicants. These effects are substantial: DI applications fall on average by 39.3%. With average award rates of about 60 to 70%, this leads a decline in DI inflow rates of 24-28%. We find strong changes in the composition of the applicant pool, in terms of age, gender, earnings, type of impairment and health. Altogether, these outcomes raise the question how our findings compare to those from other related studies that consider the effect of stricter screening on DI application behavior, DI receipt, and the earnings and employment of DI claimants.

To start with, Parsons (1991) shows that an increase in the rigor of initial eligibility screening discourages potential applicants. Specifically, a 10 percent increase in the initial denial rate induces a 4 percent decrease in DI applications. Liebert (2019) shows that including an external medical review in the disability award determination process reduces DI incidence by 23%. Autor et al. (2015) exploit exogenous variation in decision times induced by differences in the processing speed of disability examiners. They find that longer processing times reduce the employment and earnings of SSDI applicants for multiple years following application. Markussen et al. (2017) examine the effect of compulsory dialogues between the worker, the employer and the family physician in the sickness waiting period. Their findings indicate that dialogues substantially increase work resumption rates. Staubli (2011) investigates the effect of tightening DI eligibility criteria of older workers and find

declines in disability enrollment of 6 to 7.4 percentage points (26-32%). Interestingly, like in our study, they find important spillover effects into the UI scheme. Finally, Deshpande and Li (2019) find that field office closings lead to large and persistent declines in the number of DI recipients (a 13% decline in surrounding areas).

While our results confirm the idea that stricter screening has large effects on DI application and inflow rates, they also suggest that the GKP reform was highly cost-effective. Yearly applications fell from about 145,000 to about 75,000 applications in the first year of the reform. Inflow rates remained constant in the next years. From the 75,000 applications, a little more than 60% resulted in a DI benefit of on average about 12,000,- euros per year.⁴⁰ This means that about 42,000 DI benefits were averted in the year of the introduction of the GKP. With an average DI duration of 12.9 years and an annual discount factor of 2%, the expected present value of future DI benefit payment savings is a little more than 135,000 euros per averted DI claim and more than 7 billion euro in total. These savings were accompanied by cost increases of both the NSII and employers. The total administrative costs of monitoring the sickness period and conducting DI claims by the NSII amounted to 510 million euro (UWV, 2018); this amount can be interpreted as an upper bound of costs associated with increased screening by the NSII. As for employers, De Jong et al. (2014) estimate the additional costs of preventive activities and hiring occupational health services to be between 0.23 and 0.43 percent of the total wage costs, which would imply increases in employer costs associated with the GKP in the order of 600 million to 1.2 billion euro. All in all, this means that savings associated with averted benefits by far exceed the additional costs of the NSII and employers.⁴¹

For a broader assessment of the welfare consequences of the reform, insight is needed in changes in the amount of incorrect denials (Type 1 classification errors) and incorrect awards (Type 2 classification errors). If one would define DI eligibility to workers with severe health conditions only, the GKP is likely to have reduced Type 2 errors. Knowing that the Dutch DI program showed very high DI inflow rates before the GKP reform, the increase of health conditions among the remaining group of applicants then suggests that reductions in Type 2 errors dominate possible increases in Type 1 errors. Accordingly, targeting efficiency of the DI program has most likely increased under

⁴⁰Note that this may also include partial DI benefits.

⁴¹This finding corresponds to e.g. Markussen et al. (2017), who show that compulsory dialogues with sick-listed workers are highly cost-effective.

this definition. However, if one was to define eligibility as for those workers with a low work capacity in a broader sense – as e.g. the SSDI program allows for – the picture gets more nuanced (Maestas, 2019). Following this broader definition, it is important to stress that about 14% and 1.6% of workers that stop applying end up in UI benefits and means-tested welfare programs, respectively. Moreover, after exhaustion of UI benefits entitlements, those on UI revert to means-tested social welfare, with substantially lower benefits. So if these workers were hindered to engage in substantial and gainful working activities, about 16 % of them could be characterized as incomplete take-up (Kleven and Kopczuk, 2011).

Obviously, another potential source of welfare loss may come from changes in hiring practices by employers. The introduction of the GKP implied a strong increase in employer obligations. Employers may have responded to this in ways that were not intended by policymakers. Most notably, the reform could have induced them to hire high-risk workers on a temporary basis only. This effectively releases the employer of any obligation after the contract expires. Indeed, in years following the reform – particularly as from 2006 – high-risk workers became less likely to have permanent contracts and there was an increase of temporary and unemployed workers in the inflow into the DI system (Koning and Lindeboom, 2015). At this point, it should be stressed once more that our research design only assesses the short-term effects of the reform and not changes in hiring practices that took a longer time period to materialize.

6 Conclusion

We examine the targeting effects of stricter screening in the Dutch DI program, which was until recently the “most out-of-control program of the OECD countries” (Burkhauser et al., 2008). To this end, we exploit exogenous variation in stricter screening induced by the introduction of the drastic 2003 Gatekeeper protocol (GKP). In comparison to the old system, the new protocol implied higher application costs and more information about the true disability status, both during the application process and at the point of the award decision. We make a number of predictions that guide the empirical analysis. We then estimate a Regression Discontinuity in Time (RDiT) model, and look at the impact of the GKP reform on DI applications, awards, the composition of the pools of applicants, non-applicants, and awarded applicants.

We find that the introduction of the reform reduced application rates by about 40% in one year. We argue that this strong decline most likely reflects the effect of increased work resumption and/or the decision of the worker to pull out of the application process during the waiting period (self-screening). In line with our predictions, we observe the strongest decline in application behavior for individuals with difficult-to-verify impairments, and primarily for musculo-skeletal and mental impairments (which account for about two-thirds of all pre-GKP applications). Furthermore, we find differential application responses by health status, gender, earnings and age, which generate substantial changes in the pool of applicants. We implement a simple method to uncover and characterize individuals that continued to apply after the reform (i.e. “stayers”) and those that do not apply but would have done so without the reform (i.e. “leavers”). We find that the health of the “stayers” is worse than the health of “leavers”, implying that the pool of applicants becomes more deserving and that targeting efficiency has improved. At the same time, a worsening of the pool of applicants does not rule out that some of the deserving workers decide not to apply for DI benefits. To investigate the relevance of such effects, we compare the characteristics of “leavers” with the characteristics of those of non-applicants in the old system. We find that the “leavers” are in worse health than this group of non-applicants, have substantially lower incomes and are more often unemployed, with 14% of workers who were deterred from applying ending up in unemployment one year later. This suggests that there were non-negligible spillovers of the GKP reform to other social insurance programs (cf. Low and Pistaferri (2015) and Borghans et al. (2014)). To sum up, stricter screening was efficient in screening out undeserving applicants, but most likely induced some individuals with worse health and a weak labor market position to no longer apply. Further results suggest that there were no additional targeting effects at the point of the award decision, implying that changes in average health conditions of awardees were fully driven by self-screening and work resumption in the application process.

From a policy perspective, these findings are of particular interest for the design and implementation of DI screening systems. All DI programs in developed countries have a more or less structured screening procedure as part of their application processes. The screening procedure examined here consists of early interventions at the onset of sickness that leverage the existing ties between the worker and the employer. This set-up leads to strong screening effects and a more efficient targeting of DI recipients. What is more, early work resumption (83% of screened-out applicants are in work one year later) is a first guard against possible scarring effects of DI receipt and adverse long term

labor market outcomes. In light of the initial system with high DI rates and where moral hazard was important, the gains of stricter screening were likely to be substantial.

Bibliography

- AKERLOF, G. A. (1978): “The economics of ‘tagging’ as applied to the optimal income tax, welfare programs, and manpower planning,” *The American Economic Review*, 68, 8–19.
- ANGRIST, J. D. AND J.-S. PISCHKE (2009): *Mostly Harmless Econometrics*, Princeton University Press.
- AUTOR, D. (2011): “The unsustainable rise of the disability rolls in the United States: Causes, consequences, and policy options,” National Bureau of Economic Research NBER WP. 17697.
- AUTOR, D. AND M. DUGGAN (2014): “Moral Hazard and Claims Deterrence in Private Disability Insurance,” *American Economic Journal: Applied Economics*, 6, 110–141.
- 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,” National Bureau of Economic Research (NBER) Working Paper No. 20840.
- AUTOR, D. H. AND M. DUGGAN (2010): “Supporting work: a proposal for modernizing the US disability insurance system,” *Center for American Progress and The Hamilton Project*.
- AUTOR, D. H. AND M. G. DUGGAN (2006): “The rise in the disability rolls and the decline in unemployment,” *The Quarterly Journal of Economics*, 118, 157–206.
- BARRECA, A. I., J. M. LINDO, AND G. R. WADDELL (2016): “Heaping-Induced Bias in Regression-Discontinuity Designs,” *Economic Inquiry*, 54, 268–293.
- BENITEZ-SILVA, H., M. BUCHINSKY, AND J. RUST (2004): “How large are the classification errors in the social security disability award process?” National Bureau of Economic Research (NBER) Working Paper No. 10219.
- BLINDER, A. S. (1973): “Wage discrimination: reduced form and structural estimates,” *Journal of Human resources*, 436–455.
- BORGHANS, L., A. C. GIELEN, AND E. F. LUTTMER (2014): “Social support substitution and the earnings rebound: Evidence from a regression discontinuity in disability insurance reform,” *American Economic Journal: Economic Policy*, 6, 34–70.

- BOUND, J. (1989): “The health and earnings of rejected Disability Insurance applicants,” *The American Economic Review*, 79, 482–503.
- BURKHAUSER, R. V., M. C. DALY, AND P. R. DE JONG (2008): “Curing the dutch disease: Lessons for united states disability policy,” Michigan Retirement Research Center Working Paper 2088-188, University of Michigan.
- CAMERON, A. C., J. B. GELBACH, AND D. L. MILLER (2011): “Robust Inference With Multiway Clustering,” *Journal of Business & Economic Statistics*, 29, 238–249.
- CAMPOLIETI, M. (2002): “Moral hazard and disability insurance: On the incidence of hard-to-diagnose medical conditions in the Canada/Quebec Pension Plan Disability Program,” *Canadian Public Policy/Analyse de Politiques*, 419–441.
- CHARLSON, M. E., P. POMPEI, K. L. ALES, AND C. R. MACKENZIE (1987): “A new method of classifying prognostic comorbidity in longitudinal studies: development and validation,” *Journal of chronic diseases*, 40, 373–383.
- CRODA, E., J. SKINNER, AND L. YASAITIS (2018): “The health of Disability Insurees: An international comparison,” Working Paper 2018:28 Department of Economics, University of Venice “Ca’ Foscari”.
- CRODA, E., J. S. SKINNER, AND L. YASAITIS (2013): “An International Comparison of the Efficiency of Government Disability Programs,” NBER Disability Research Center Paper No. NB 13-08.
- DE JONG, P., M. GIELEN, AND V. HAANSTRA-VELDHUIS (2014): “Verzekeringsgraad kleine werkgevers,” APE report number 1209.
- DE JONG, P., M. LINDEBOOM, AND B. VAN DER KLAAUW (2011): “Screening disability insurance applications,” *Journal of the European Economic Association*, 9, 106–129.
- DE JONG, P. R. (2011): “Sickness, disability and work: Breaking the barriers – A synthesis of findings across OECD countries – By OECD,” *International Social Security Review*, 64, 103–104.
- DELL, M. (2010): “The persistent effects of Peru’s mining mita,” *Econometrica*, 78, 1863–1903.
- DESHPANDE, M. AND Y. LI (2019): “Who Is Screened Out? Application Costs and the Targeting of Disability Programs,” Forthcoming *American Economic Journal: Economic Policy*.

- FULTZ, E. (2015): “Disability Insurance in the Netherlands: A Blueprint for US Reform,” *Washington, DC: Center for Budget and Policy Priorities*.
- GRUBER, J. (2000): “Disability insurance benefits and labor supply,” *Journal of Political Economy*, 108, 1162–1183.
- GRUBER, J. AND J. D. KUBIK (1997): “Disability insurance rejection rates and the labor supply of older workers,” *Journal of Public Economics*, 64, 1–23.
- HALLER, A., S. STAUBLI, AND J. ZWEIMULLER (2019): “Tightening Disability Screening Or Reducing Disability Benefits? Evidence and Welfare Implications,” Mimeo.
- HAUSMAN, C. AND D. S. RAPSON (2017): “Regression discontinuity in time: Considerations for empirical applications,” National Bureau of Economic Research (NBER) WP No. 23602.
- JOHANSSON, P., L. LAUN, AND T. LAUN (2014): “Screening Stringency in the Disability Insurance Program,” *B.E. Journal of Economic Policy and Analysis*, 14, 1–19.
- KARLSTRÖM, A., M. PALME, AND I. SVENSSON (2008): “The employment effect of stricter rules for eligibility for DI: Evidence from a natural experiment in Sweden,” *Journal of Public Economics*, 92, 2071–2082.
- KERKHOF, MARCEL, L. M. AND J. THEEUWES (1999): “Retirement, Financial Incentives and Health,” *Labour Economics*, 6, 203–228.
- KLEVEN, H. J. AND W. KOPCZUK (2011): “Transfer program complexity and the take-up of social benefits,” *American Economic Journal: Economic Policy*, 3, 54–90.
- KONING, P. AND M. LINDEBOOM (2015): “The rise and fall of disability insurance enrollment in the Netherlands,” *The Journal of Economic Perspectives*, 29, 151–172.
- KUHN, A., J.-P. WUELLRICH, AND J. ZWEIMÜLLER (2010): “Fatal attraction? Access to early retirement and mortality,” Institute for the Study of Labor (IZA) Discussion Paper no. 5160.
- LARSSON, L. (2006): “Sick of being unemployed? Interactions between unemployment and sickness insurance,” *The Scandinavian Journal of Economics*, 108, 97–113.
- LEE, D. S. AND T. LEMIEUX (2010): “Regression discontinuity designs in economics,” *Journal of Economic Literature*, 48, 281–355.

- LIEBERT, H. (2019): “Does external medical review reduce disability insurance inflow?” *Journal of Health Economics*, 64, 108–128.
- LOW, H. AND L. PISTAFERRI (2015): “Disability Insurance and the Dynamics of the Incentive Insurance Trade-Off,” *The American Economic Review*, 105, 2986–3029.
- MAESTAS, N. (2019): “Identifying work capacity and promoting work: A strategy for modernizing the SSDI program,” *The Annals of the American Academy of Political and Social Science*, 686, 93–120.
- MAESTAS, N., K. J. MULLEN, AND A. STRAND (2013): “Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt,” *The American Economic Review*, 103, 1797–1829.
- MARKUSSEN, S., K. ROED, AND R. SCHREINER (2017): “Can compulsory dialogues nudge sick-listed workers back to work?” *The Economic Journal*.
- MCCRARY, J. (2008): “Manipulation of the running variable in the regression discontinuity design: A density test,” *Journal of econometrics*, 142, 698–714.
- MOORE, T. J. (2015): “The employment effects of terminating disability benefits,” *Journal of Public Economics*, 124, 30–43.
- OAXACA, R. (1973): “Male-female wage differentials in urban labor markets,” *International Economic Review*, 693–709.
- OECD (2009): “Maladie, invalidité et travail – Garder le cap dans un contexte de ralentissement de l’activité économique,” *Document de base, Forum Å Ĥ Haut Niveau, Stockholm, 14-15 mai 2009*.
- (2010): “Sickness, Disability and Work: Breaking the Barriers. A Synthesis of Findings Across OECD Countries,” OECD.
- PARSONS, D. O. (1991): “Self-screening in targeted public transfer programs,” *Journal of Political Economy*, 99, 859–876.
- STAGG, V. ET AL. (2015): “CHARLSON: Stata module to calculate Charlson index of comorbidity,” *Statistical Software Components*.

STAUBLI, S. (2011): “The impact of stricter criteria for disability insurance on labor force participation,” *Journal of Public Economics*, 95, 1223–1235.

TREITEL, R. (1979): “Disability claimants who contest denials and win reversals through hearings,” Tech. rep., Staff Report No. 34, Washington : Soc. Security Admin., Office Res. and Statis.

UWV (2018): “UWV Jaarplan 2018,” UWV Amsterdam.

VON WACHTER, T., J. SONG, AND J. MANCHESTER (2011): “Trends in Employment and Earnings of Allowed and Rejected Applicants to the Social Security Disability Insurance Program,” *The American Economic Review*, 101, 3308–3329.

A Figures

Figure A1: Placebo analysis: Regression-Discontinuity plots for DI application with cutoff at placebo dates.

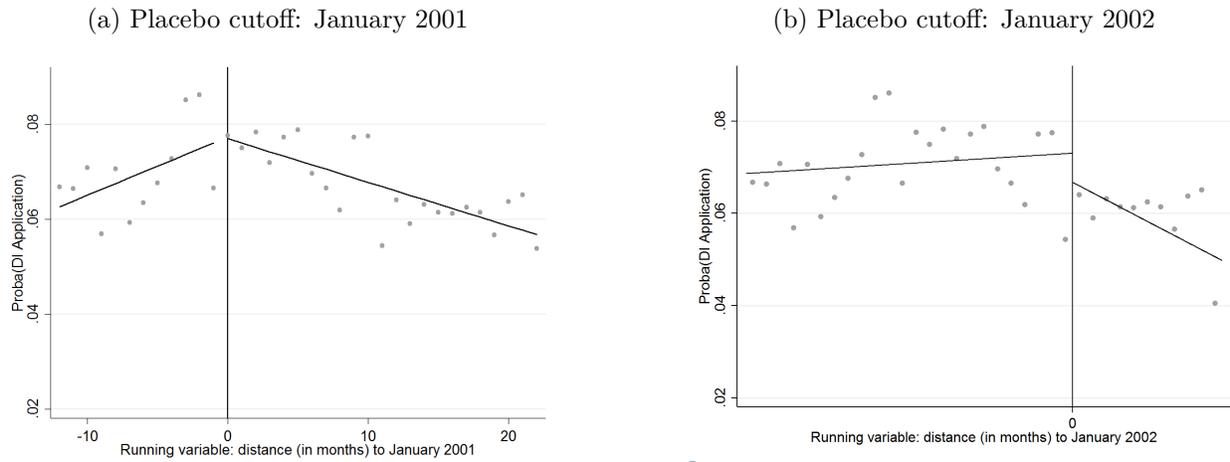


Figure A2: Donut Regression-Discontinuity plots (dropping those within one month of the cutoff) for DI application with various choices of polynomial time control.

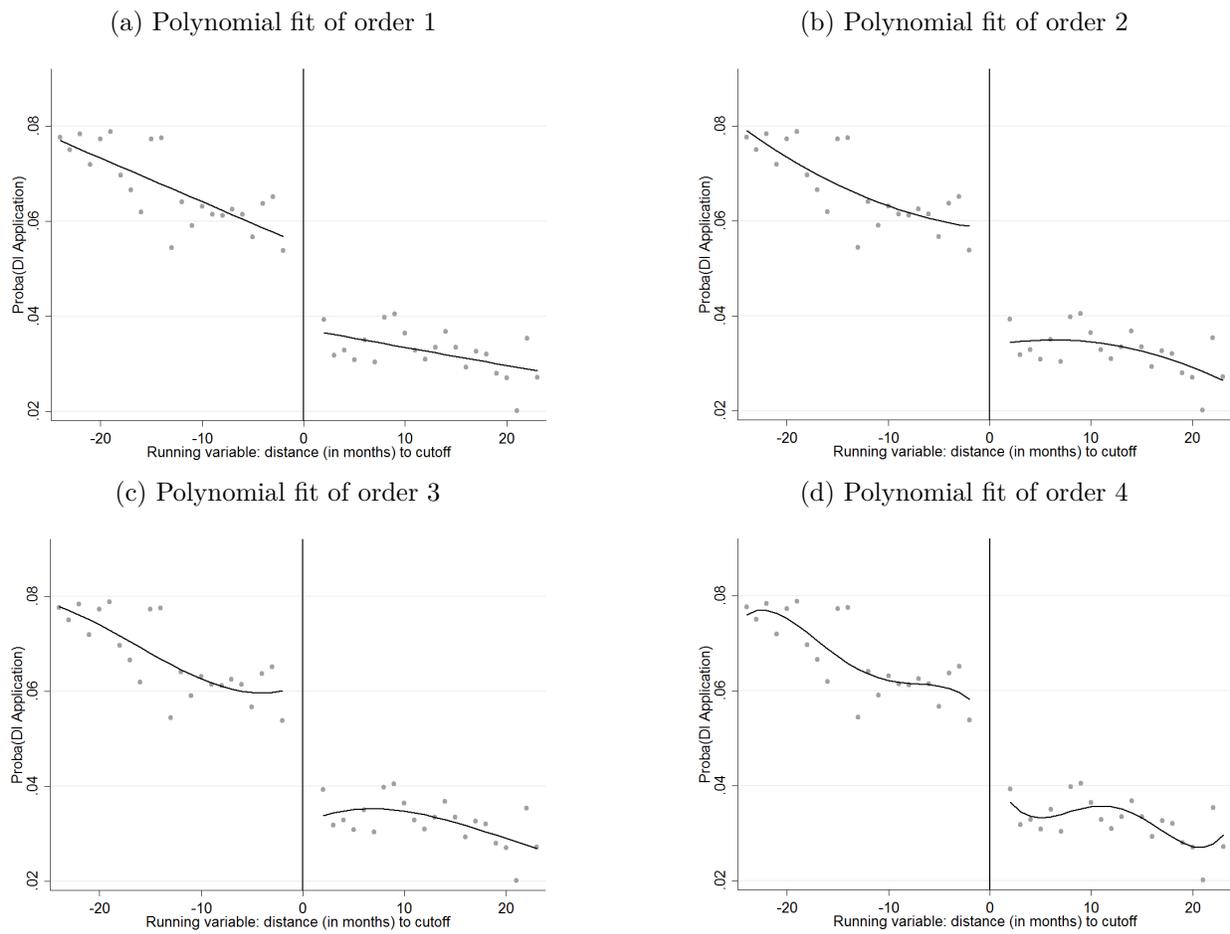
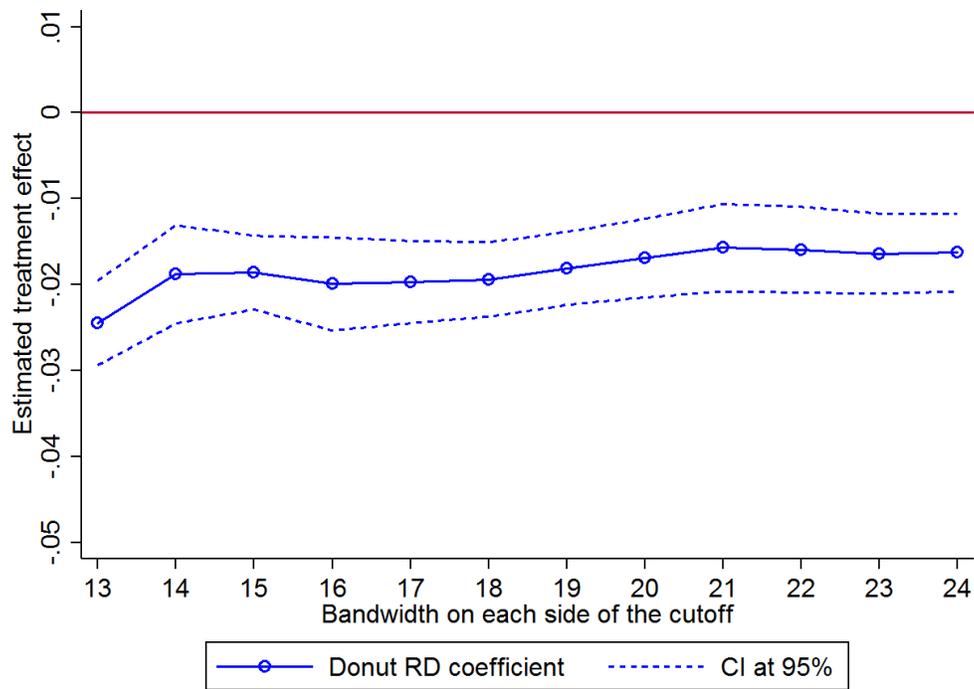


Figure A3: The effect of the Gatekeeper Protocol on DI application behaviour. Donut Regression-Discontinuity estimates (dropping those within one month of the cutoff) with varying bandwidths.



Note: Estimates are obtained by running Equation (1). Note that in the presence of month-of-year fixed effects, the model is not identified for bandwidths below 12 months on each side of the threshold.

B Data appendix : The Charlson Comorbidity Index

The Charlson Comorbidity Index (CCI) is a popular tool for predicting mortality by classifying or weighting comorbid conditions (comorbidities) – see Charlson et al. (1987). The CCI can be constructed from medical record abstract or administrative data. We use the coding algorithm developed by Stagg et al. (2015) to derive the CCI from ICD-9-CM administrative data. For each individual, in each year, we compute the CCI index as a weighted sum of 17 comorbidities. The 17 comorbidities and their associated weights – that allow for adjustment for severity of illness – are listed in Table A1. As we have longitudinal data, we then compute a time-varying comorbidity index that aggregates information over multiple hospitalization spells since 1995.

Table A1: Charlson Comorbidities and Weights

Comorbidity	Assigned weight
Acute Myocardial infection	1
Congestive Heart Failure	1
Peripheral Vascular Disease	1
Cerebrovascular Disease	1
Dementia	1
Chronic pulmonary disease	1
Rheumatic disease	1
Peptic Ulcer Disease	1
Mild Liver Disease	1
Diabetes without chronic complications	1
Diabetes with end organ damage	2
Hemiplegia / Paraplegia	2
Renal (kidney) Disease	2
Cancer (Any malignancy/lymphoma/leukemia)	2
Moderate or severe liver disease	3
Metastatic Cancer	6
AIDS/HIV	6