TI 2019-036/V
Tinbergen Institute Discussion Paper

Targeting Disability Insurance Applications with Screening

Mathilde Godard[1]
Pierre Koning[2]
Maarten Lindeboom[3]

[1] CNRS, GATE-LSE, University of Lyon
[2] Vrije Universiteit Amsterdam, Leiden University, TI, IZA
[3] Vrije Universiteit Amsterdam, TI, IZA

Targeting Disability Insurance Applications with Screening*

Mathilde Godard† Pierre Koning‡ Maarten Lindeboom§

May 9, 2019

Abstract

We examine the targeting effects of increased scrutiny in the screening of Disability Insurance (DI) applications using exogenous variation in screening induced by a policy reform. The reform raised DI application costs and revealed more information about the true disability status of applicants at the point of the award decision. We use administrative data on DI claims and awards and merge these with other administrative data on hospitalization, mortality and labor market outcomes. Regression Discontinuity in Time (RDiT) regressions show substantial declines in DI application rates and changes in the composition of the pool of applicants. We find that the health of those who are not discouraged from applying is worse than those who are. This suggests that the pool of applicants becomes more deserving. At the same time, compared with those who did not apply under the old system of more lax screening, those who are discouraged from applying are in worse health, have substantially lower earnings and are more often unemployed. This indicates that there are spillovers of the DI reform to other social insurance programs. As we do not find additional screening effects on health at the point of the award decision, we conclude that changes in the health condition of the pool of awarded applicants are fully driven by self-screening of (potential) applicants.

JEL code : H2, I3

Keywords : Disability Insurance, Screening, Composition effects, Targeting efficiency

*We are grateful to Patrick Hullegie for his support and initial work on the data, to Christopher Cronin, Izabela Jelovac, 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 to 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 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.

†CNRS, GATE-LSE, Univ. of Lyon. 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 conditions. The intention is to reduce moral hazard in the DI program, but the risk is that those who have lost earnings capacity because of disability are left underinsured. Improved targeting weakens the trade-off faced between provision of insurance and the constraint of moral hazard. Potentially this can be achieved through more intensive screening of applications by medical examiners/gatekeepers and imposing eligibility requirements such as reintegration obligations or the use of waiting times – see Deshpande and Li (2019), Kleven and Kopczuk (2011), Markussen et al. (2017), Autor et al. (2015), Haller et al. (2019) and Liebert (2019) for recent contributions.

More scrutiny of applications may deter reporting of work absence and the filing of DI claims. If such deterrence effects are confined to workers with the least health impairments, this self-screening improves targeting efficiency. However, if increased application costs disproportionately deter those with the most severe disabilities, for instance due to their higher mortality expectations, targeting efficiency may worsen as well. This is what Parsons (1991) refers to as 'perverse self-screening'. Thus, the overall impact of these requirements on targeting efficiency is ambiguous.

We examine the targeting effects of increased scrutiny of (DI) applications to the Dutch DI program by using exogenous variation in screening induced by the introduction of a system with strict screening. The Gatekeeper Protocol (GKP) was implemented in April 2002 and changed the role of the employer and the National Social Insurance Institute (NSII) in the DI application process. Previously, the NSII, the worker and the employer had a joint responsibility for work resumption of the sick listed worker. The new law made the worker and the employer responsible for undertaking specific measures to increase the likelihood that the worker would return to work. The worker and employer had to draft a report of the causes and consequences of the impairment and a rehabilitation plan specifying the aim of the plan (for instance, work resumption at the current employer) and the steps to reach it. Next, if work resumption did not occur during the waiting period, the role of the NSII was to screen the rehabilitation plan. That is, to check on the presence and severity of the impairment, the consequences for the earnings potential of the applicant, whether sufficient measures had been undertaken to resume work and whether the disability benefit should be awarded. In comparison to the old system, the protocol implied more medical checks during

the waiting period, more work resumption efforts and, consequently, more application costs for the worker and the employer. These medical checks and work resumption measures may also have increased the information about the true disability status of the applicant, which in turn may have contributed to targeting efficiency.

The goals of the GKP were to reduce DI inflow, to increase employment rates of workers with disabilities and to ensure that benefits were provided to those who really needed them – improving the targeting efficiency of the DI system. Figure 1 suggests that the reform was very successful in reducing DI inflow rates, which dropped at the time of the introduction of the GKP. Among others, the reform contributed to the overall drop of DI enrolment from 11% in 2001 to 7.2% in 2012 (Koning and Lindeboom, 2015). However, it is unclear whether the protocol improved targeting efficiency of the DI system.

We merge large-scale administrative data from the NSII on DI claims and award decisions with administrative data on hospital admissions, mortality and labor market outcomes of applicants and non-applicants. We first estimate the effects of the introduction of the GKP (the extensive margin effects) in a Regression Discontinuity in Time (RDiT) framework. We examine DI claims and award rates for worker cohorts before and after the moment the GP came into effect. To control for potential anticipatory and delay effects, we also employ 'donut' RD regressions that exclude monthly observations that are closest to the time cutoff of interest. In line with expectations, we find a 40% decline in the DI application rates at the introduction of the GKP. The largest declines occurred in difficult-to-verify health conditions such as musculo-skeletal and mental impairments, that account for about two-third of all applications. The drastic decline in application rates raises important questions for the targeting of DI benefits. This concerns the applicant types that are screened out during the waiting period, particularly the health status of this group. As the GKP provides the medical examiners with more information that potentially reduces classification errors in the award decision, another question is whether there are further targeting gains at the point of the award decisions.

To address these questions empirically, we examine changes in the compositions of DI applicants and awards. Our main findings can be summarized as follows. First, we find that the pool of applicants became more deserving. The health of the individuals who were not deterred by the GKP from applying is worse than the health of those who were deterred. This suggests that targeting efficiency improved. At the same time, however, we also infer that those who were deterred from applying are in worse health, have lower incomes and have a higher unemployment risk than the

(initial) group of non-applicants. Second, we show that changes in the average health conditions of awarded applicants are fully driven by self-screening of potential applicants during the waiting period and not by changes in the award decision. Finally, additional experimental evidence of regions that further increased the stringency of the GKP – labelled as intensive margin effects – primarily affects the behavior of applicants that have difficult-to-verify impairments.

Our analysis relates to a larger literature about moral hazard in DI.[1] Of most relevance is the literature that investigates classification errors arising from imperfect information about the individual's true disability status (Kleven and Kopczuk, 2011; Parsons, 1991). Parsons (1991) was the first to distinguish Type I (incorrect denials) and Type II (incorrect awards) in the context of DI.[2] Kleven and Kopczuk (2011) argue that transaction costs increases due to imperfect information may deter individuals from applying to social insurance programs. In an international comparison Croda et al. (2018) examine how well countries target those in the poorest health across countries, or within a country over time.

As to empirical studies that are most relevant to our paper, recent contributions are made by Deshpande and Li (2019), Autor and Duggan (2014), Autor et al. (2015), Liebert (2019) and Markussen et al. (2017). Deshpande and Li (2019) analyze the effect of office closures on the likelihood of SSDI applications and awards. They find that the deterring effect of office closures decreased the targeting efficiency of SSDI benefits. Likewise, 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. In addition, Liebert (2019) investigates the effects of introducing mandatory external medical reviews in the disability award determination process in a system that relies on treating physician testimony. He finds that external medical review reduces DI incidence by up to 23% and increases labor market participation. Even if self-screening did not contribute to this effect, this reduced-form reform effect can be interpreted as a net reduction in DI award errors. 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.

[1] Empirical 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).

[2] A small literature looking at classification error rates in the US SSDI award processes suggests that both award and rejection errors are very common (Benitez-Silva et al., 2004). Also, the evidence suggests large differences in classification errors of DI systems across countries (Croda et al., 2013).

We are one of the few studies that examine the effect of screening stringency and the related increases in application costs and screening information on the composition of the applicants. We advance the literature in three ways. First, we define and implement new measures of targeting efficiency in order to evaluate the effects of the introduction of scrutiny in the screening of applications (the extensive margin). We argue that the effect of the reform may have had differential effects by type of impairment and that scrutiny in the screening process may induce perverse screening effects – that is, it may dis-proportionally deter those in bad health with difficult-to-verify impairments from applying. To test for such effects, we both characterize compliers who stop applying after the introduction of the reform – and enter in the pool of non-applicants – and non-compliers who keep applying after the introduction of the reform. As a second contribution, we are the first to consider changes in targeting efficiency in both the DI application process and the award determination process. That is, we exploit rich information on future health and labor market outcomes of individuals to detect changes in the composition of the pool of applicants and admissions and infer where most of improvements in targeting efficiency (if any) are obtained. Third, we complement these extensive margin analyses with evidence from a field experiment that was implemented at the start of the GKP. In two out of the twenty six regions stricter screening procedures were imposed than in other regions. These analyses at the intensive margin inform whether and for which type of conditions further gains in targeting efficiency are possible.

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 behavior. Data are presented in section 3. Section 4 includes the empirical empirical model and presents simple methods to separate self-screening effects during the waiting period from changes in the medical assessment at the 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. Section 5 also includes a brief description of the field experiment and the intensive margin results. We end with conclusions in Section 6.

2 The Gatekeeper reform

2.1 Institutional background

Before we proceed by explaining the screening process that precedes DI claims, it is important to note two 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, providing insurance of 70% of the corresponding loss of income. As Koning and Lindeboom (2015) argue, this broad setup of the scheme has increased the possibility of sizable screening errors in disability determinations. Second, wage payments for sick workers are continued in the waiting period that precedes disability claims and the employer bears the wage and sick-pay costs for this period. As a result, the system did not provide strong financial incentives for sick-listed workers to resume work quickly. Taken together, these two characteristics of the Dutch DI scheme laid the ground for a broad coverage of impairments and limited self-screening, thus causing 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 and to improve the targeting efficiency of the program.

The GKP stipulates the responsibilities of the 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 these data the employer and employee together draft 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. From this moment on the worker is added to the administrative database of the NSII.

The worker can file a DI application if s/he has not resumed work before the expiration of the waiting period. At the start of the GKP in April 2002 the waiting period was one year. This was later extended to two years.[3] Benefit claims are only considered admissible by the NSII if they are accompanied by the 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 transfer to disability benefits. 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.

[3]The first applications under the two year waiting period arrived in 2006 at the offices of the NSII

The employer and the sick worker are jointly responsible for reintegrating the sick worker back into his or her old job or into a new job commensurate with the worker's limitations. However, if the employee consistently rejects reasonable offers and accommodations, the employer may stop paying sickness benefit and, eventually, fire the employee. If it is clear that sufficient reintegration efforts have been undertaken by both the employer and the worker, the DI claim is processed. In a final examination, NSII determines the presence of impairments, the consequences for the earnings potential of an applicant, the degree of disability as a percentage of the worker's former wage, 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 the full costs of sick pay during the waiting period; the costs of the health service agency[4]; and the costs of reintegration measures and work accommodations to facilitate work resumption. Moreover, the employer is financially responsible in case of non-compliance or negligence and at risk of continued wage payments after the end of the waiting period. Particularly smaller employers responded to this 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. De Jong et al. (2014) also point out that additional wage premium of extended coverage varied between 0.23 and 0.43 percentage point, which in turn amounts to 7.5 % and 15 % of the average premiums that are paid, respectively.[5]

[4]In the Netherlands, private health service agency play a key role when adhering to the minimum requirement that are set in the reintegration process of workers

[5]Based on comparisons or 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.

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 precedes the DI claims assessment. Employers that did not yet comply to the new norms needed to increase their reintegration activities.[6] As a result, on average work resumption rates may have increased and consequently the number of DI applications at the end of the waiting period may have decreased. From the increases in reintegration activities, employers (and workers) may also have learned more about the ability to resume work and prospects of future acceptance into the DI program. This may also have affected the decision to proceed with the application process and hence reduced overall DI application rates.

It is likely that the impact of increased requirements for employers on testing and reintegration activities differed across disease types. For some diseases the cause of the disease, the treatment and prospects for work resumption are clear so that – if work resumption is possible – effective reintegration plans are relatively easy to implement. This may include acute and curable health problems that last for a limited time period, for which the value added of the protocol was probably limited. Alternatively, there may be severe known conditions (e.g. last stage cancers) for which prospects of work resumption are zero. In such situations no additional medical tests are required and work resumption is no option. In contrast, there are also less clear-cut conditions, e.g. mental conditions such as burnout and depression, for which the causes are more difficult to identify, the severity is more difficult to assess and where it is uncertain which reintegration measures are most effective. For such, "difficult-to-verify" diseases, the protocol is potentially more effective in increasing work resumption rates and reducing DI applications.

Changing the perspective to that of the worker, the key question is whether the introduction of the GKP changed the decision to enter the DI process (i.e., the waiting period). Essentially, the introduction of the protocol affected this decision in two ways. On the one hand, it increased the worker's costs of participating in the DI waiting period. Increases in the scrutiny during the application process implied more interventions, more verification tests, more visits to the doctor, more documents to support the claim and hence more costs.[7] These costs are likely to have deterred workers from entering the DI process. On the other hand, the reform may also have induced stronger self-screening among workers due to decreases in the noise of the disability signal of applicants. At the point of the award decision, the NSII has obtained more information about the true disability

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

[7] Note that these costs may also be non-monetary, as there may be (emotional) hassle of going through the application process.

status of the applicant. Holding eligibility standards constant, the NSII thus may have reduced classification errors.[8] This element of the screening effect is most relevant for workers with conditions that are more difficult to verify. Particularly for the most able workers in this group, the reduction in Type 2 errors reduces the conditional award rate, which (combined with increased costs) may have induced them to not enter the DI application process. At the same time, individuals with difficult-to-verify diseases that initially were among the Type I errors may have been induced to apply for DI benefits. This type of self-screening would then increase the number of DI applications for this group.

Summing up, the overall consequences of the reform were most likely to decrease the DI applications for the most able workers with conditions that are difficult to verify. For this group, there were strong increases in the expected costs of the application and the reduction in noise in the disability signal reduced the expected gains from the application. This may have reduced opportunistic behavior among these type of workers. For the less able workers with difficult-to-verify diseases, however, the effects of the reform on DI applications and awards are ambiguous. More scrutiny in the screening process may have increased their award rates, but at the same time application costs increased. These increases in application costs were potentially large for workers with high mortality expectations, risk averse workers or worker that were put under high strain by the reintegration efforts.

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*, 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 application (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%). For each application we observe the date of the award decision, but not the exact date of the application. As Figure 2 shows, applications usually occur after 39 weeks of

[8] At this point, it should be stressed that the GKP did not imply any changes in the DI eligibility criteria and medical examiners were explicitly instructed to maintain the DI criteria of the old system.

absence and award decisions 13 weeks later. In principle, changes in DI applications due to the GKP reform should start at January 2003. We can, however, not rule out that there are some delays in the system, causing some applications under the old regime to arrive later than December 2002. For this reason and for reasons argued later on, we will employ a research design that uses a "donut hole" around the cutoff date.

We merge the *DI application register* with several administrative records of Statistics Netherlands covering the entire population. The tax records measure earnings (in 2010 Euros) from 1999 onwards. Based on this information we consider a worker employed if he or she has any positive earnings in a given calendar year.[9] 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.[10] 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, hospitalization spells due to mental disorders are poorly measured. For each individual in the LMR data we observe (i) the admission and discharge date, and (ii) 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 his/her (potential) waiting period. For each individual, we also construct the Charlson Comorbidity Index that aggregates information over multiple hospitalization spells since 1995 (see Data Appendix D for more information on the construction of this Charlson Comorbidity Index). Finally, we merge our data with the death register. With this information we can track future mortality outcomes for applicants and non-applicants.

Our sample of interest consists of prime-age individuals – aged 25-65 – who were employed in the previous year. We exclude cases with conflicting information in the different databases. Such cases include, for instance, applicants who are (according to the NSII register) rejected, but who nevertheless appear as DI recipients in the tax register (11.3% of rejected applicants). Conversely,

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

[10]With the ZIP codes we can identify for each individual the relevant regional office of the NSII. This is of relevance for the field experiment that was performed in two out of the 26 regions in the Netherlands. See Section 5.4 for more information.

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). We also exclude those who apply more than once during the period under study, as we cannot be sure which application is linked to the final award decision. 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).[11] For each month in 2001-2004, we take all individuals who apply that month as well as a 1% random sample of all non-applicants we observe in that month. Our final sample consists of a maximum of 95,338 individuals in each month, measured over the years 2001-2004.

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 pre-GKP period (2001-2002, column 1) and 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 Gatekeeper system (2001-2002) to 0.04% in the post-gatekeeper years (2003-2004). Similarly, the DI award rates as a percentage of the total population decrease from 0.04% to 0.02% over the same period. In addition, Figure 3 shows how the number of DI application and awards have evolved over 1999-2004 at the monthly level. [12] 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.[13] When 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 increases 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 hospital database.

Table 2 breaks the pre- and post-GKP period for the groups of applicants and the non-applicants in our sample. A comparison between column 1 and 3 of this table suggests that the reform induced

[11] In the empirical analyses we use RDiT regressions that effectively use information from 2002 and 2003.

[12] For the purpose of exposition we align the award and application date in the figure.

[13] The drop in the number of DI applications and awards in December 2002 seems larger than in other years. We will discuss this point in detail in Section 4.

substantial changes in the composition of the pool of applicants. Applicants are more often older, female and non-native, have lower average income, lower employment rates and are more often unemployed or on welfare. Measured in terms of the Charlson index and the hospitalization rates, applicants in the GKP system are also in worse health. This decline in health is a first indication that benefits are now more tailored towards the deserving. From this alone, however, we cannot rule out that the truly disabled may have screened themselves out of the application process. For this one needs to look at the change in the health in the pool of non-applicants (columns 2 and 4). These columns show that the health condition of the non-applicants has not changed. Taken together, this is first tentative evidence that the introduction of the Gatekeeper protocol may have increased targeting efficiency in the application process. It has to be noted, however, that group of non-applicants is much larger than the group of applicants. This makes it difficult to detect application deterrence effects in the means of the group of applicants. For this reason we employ a different route in the empirical analyses.

To shed some light on changes in the DI determination process, we show two year average shares by impairment type at either side of the reform date in Table 3. The top panel of the table shows that before the reform (column 1), about 65% of DI applicants apply for musculo-skeletal or mental impairments. These are impairments that are generally more difficult to verify. After the reform the share of these difficult-to verify diseases declines with about eight percentage points to about 57%. The bottom panel of the table shows that the award rate of applicants is generally high, but also that it varies substantially across impairments: it is typically lower for difficult-to-verify diseases, such as musculo-skeletal, mental or 'other' non-categorized impairments. After the start of the gatekeeper protocol we see some changes in the award rates. However, as argued earlier, it is difficult to interpret these changes directly as they reflect changes in application behavior (the type of applicant that applies under the new system) as well as changes in information about the health of the applicant available to the medical examiner at the point of the award decision. We turn to this in the next section.

Overall, the tables above have suggested that the introduction of the GKP has led to a sharp decline in application rates, and that the pool of applicants becomes on average less healthy. At the same time, there is no worsening of health in the pool of non-applicants. This is indicative of improvements in targeting efficiency. These first results also provide suggestive evidence that most declines in applications are in difficult-to-verify diseases.

4 Estimation strategy

The idea of targeting efficiency in DI is that only truly disabled workers will be awarded benefits, while those with mild or without any health conditions are screened out in the DI application process. Improvements in targeting efficiency may result from increases in the rigor of the screening process that reduce the noise of the disability signal and the scope for opportunistic behavior. As argued earlier, however, the GKP also increased the application costs for workers. Particularly for disabled workers with difficult-to-verify conditions, deterrence effects stemming from costs increases associated with additional tests and reintegration measures may have been substantial. While increased scrutiny in the screening process will most likely lead to a fall in the number of applications, it does not necessarily imply improvements in targeting efficiency. Targeting efficiency will improve only if the fall in applications comes primarily from a reduction in applications of more able workers.

With this in mind, the challenge of this paper is to investigate how the introduction of the GKP changed the composition of health conditions of DI applicants and non-applicants. To address these questions, we first investigate the overall effect of the increased scrutiny by comparing the numbers of DI applications just before and after the start of the GKP. More specifically, we use a Regression Discontinuity in Time (RDiT) specification that exploits the sharp discontinuity in January 2003. Second, we employ a similar RDiT estimation approach to uncover changes in the composition of DI applications, including the shares of diagnosis types, the average health conditions (hospitalizations and mortality rates) and average labor market outcomes. Interestingly, the combined estimation results for changes in application rates and changes in shares of applicant types allow us to infer the share averages for both the remaining applicants (i.e. non-compliers) and compliers to the reform that have become non-applicants. Finally, we broaden our perspective to the effects of the reform on award decisions. Bearing in mind that medical examiners were provided with more information on the reintegration process in the waiting period that precedes DI claims, they could potentially use this to reduce classification errors at the point of the award decision. Therefore, we aim to test whether changes in the medical screening induced additional gains in targeting efficiency. In what follows, we explain these three steps in detail.

4.1 Specification of the RDiT model

In all steps of our empirical analysis, we compare the number and composition of DI applications and conditional DI awards just before and after the start of the GKP. We use Regression Discontinuity

in Time (RDiT) specifications that exploit the sharp discontinuity in January 2003 when the first applicants arrived that followed the GKP requirements (Hausman and Rapson, 2017). In light of the reorganizations that went together with the reform, we extend this specification by dropping observations that are closest 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 cutoff effects 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 cut-off. The vertical line indicates the exact month the GKP started to affect DI claims.[14] A quick look at the graph indeed shows a sharp and substantial drop in DI application rates at the cutoff, which is much larger than discontinuities at other 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.

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:

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

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 and award rates of 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 $k = 2, .., 12$ is a set of month indicators and therefore δ_k capture month-of-year fixed effects.[15] X_i is a vector of time-invariant characteristics (age, gender and ethnic background). Standard errors are clustered

[14]The NSII implemented a field experiment that introduced a differential screening policy in two out of the 26 regions in the Netherlands. The two treated regions (Apeldoorn and Hengelo) are excluded in the current figure. We return to this issue in Section 5.4.

[15]Recall from Figure 4 that particularly end-of-year effects may cause a strong drop in DI application and award decisions. This is the effect of the closing of the regional offices due to Christmas in the last one or two weeks of December.

along both the month-year and the individual dimensions.[16]

The parameter β measures the impact of the reform on the outcome measure of interest. DI applications are filed after nine months, at the end of the waiting period. Therefore, in theory, β should reflect (i) 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) the decision of worker to drop out during the application process. 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 disability status of the worker and the prospect of being awarded DI benefits. The RDiT framework of Equation (1) focuses on the first applicant cohorts after the cutoff. Applications in 2003 are from workers who first reported sick in 2002 or at the start of 2003 when employers and workers had no knowledge about the details of the screening policy. We therefore expect the parameter β to largely reflect mechanisms (ii) and (iii).

Possible anticipation effects are also an issue in our context. The relevant time window for anticipation effects would be between December 2001 – i.e. the moment the reform was announced – and April 2002. This in turn implies an increase in DI award decision 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 act in national newspapers was low when the reform was proposed but not decided upon yet.[17] Between December 2001 and March 2002, the months between the announcement and the implementation, we see a small increase in press coverage, also rendering it unlikely that employers and workers anticipated 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.

[16] Lee 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).

[17] The 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 Inferences on targeting efficiency: DI applications

While our focus is on the reform effects on (total) DI applications, we also consider the composition of the pool of applicants as an outcome measure. The intention is to infer the average characteristics of workers that continued to apply for DI benefits and those who would apply under the scheme but do not so under the new scheme. The latter group gives direct insight into the deterrence effects of the GKP reform. Following the standard literature on Instrumental Variable estimation, we can characterize those who continued to apply as‘“always-takers” (or: non-compliers) and those that do not apply but would so under the counterfactual as “compliers” to the GKP reform (Angrist and Pischke, 2009). Our aim is to uncover the characteristics of both non-compliers and compliers.

To formalize this approach, we first define β_Y as the causal reform effect on the DI application rate and β_X as the causal reform effect on the averages of exogenous applicant characteristics X. $\overline{X}_C$ includes the value averages of X of compliers to the reform, $\overline{X}_{NC}$ the averages of X for non-compliers and $\overline{X}$ denotes initial averages of all applicants before the reform. This implies the following defining equations:

$$\overline{X} = (1 - \beta_Y)\ \overline{X}_{NC} + \beta_Y\ \overline{X}_C \tag{2}$$

and

$$\beta_X = \overline{X}_{NC} - \overline{X} = \beta_Y\ (\overline{X}_{NC} - \overline{X}_C) \tag{3}$$

Rewriting equation (3) yields:

$$\overline{X}_C = \overline{X}_{NC} - \frac{\beta_X}{\beta_Y} \tag{4}$$

The above expression makes apparent that the difference in the value average of X for compliers and non-compliers can be inferred from the effect estimates of the reform on DI applications (i.e., β_Y) and on the value average of X (i.e., β_X). By definition, increases in the value averages of X among non-compliers imply that the sample of compliers have lower average values of X. With the effect estimates of β_Y, β_X and the observed value average of X among non-compliers (i.e., $\overline{X}_{NC}$), we can derive average characteristics of the compliers that are not (directly) observed in our sample. Depending on the type of variables at hand, we can thus assess whether the remaining pool

of applicants has become more deserving due to the reform.[18] These results provide insight of the reform effects on targeting efficiency among the pool of applicants and non-applicants.

4.3 Inferences on targeting efficiency: DI award rates

To examine the effects of the reform on targeting efficiency for the awarded DI benefits, we finally broaden our perspective to the effects of the reform on conditional award decisions. Again, we use the RDiT specification to obtain such effects. Following Deshpande and Li (2019), one may argue that increases in conditional award rates can be interpreted as improvements in targeting efficiency stemming from self-screening of applications. In their analysis, increases in application costs induced by DI office closures may reduce applications. Targeting efficiency improves if the pool of applicants becomes more deserving, i.e. if conditional award rates increase.[19] In our context, however, changes in conditional award rates may reflect both changes in application behavior – that is, the composition of the pool of applicants changes – and changes in the award decision process that stem from changes in the disability signal. While self-screening typically increases conditional award rates, the reduction in the noise of the disability signal has ambiguous effects. A reduction in false negatives increases award rates, whereas a reduction in false positives decreases award rates.[20]

In light of these arguments, our key challenge is to separate the effects the reform had on the composition of applicants from the effect that stems from changes in the medical assessment. We address this issue by combining information on changes in application rates and changes in conditional award rates. More specifically, we focus on dichotomous health indicator H that is included in matrix X. We define p as the share of applicants with bad health. Applicants with bad health have conditional award probabilities equal to a_H, versus a_L for applicants with good health. For notational convenience, we next define $\delta = \frac{a_H}{a_L}$ as the relative conditional award rates for those with bad health. We observe the average health of all applicants, $\overline{H}_{app}$, and of all awardees, $\overline{H}_{award}$. From this, we can express observed average health rates in terms of p and δ:

$$\overline{H}_{app} = (1-p)\ 0 + p = p \tag{5}$$

[18] The rationale behind this argument is that in the old system, without scrutiny in the screening process and with excessively high DI rates, moral hazard was substantial – see e.g. Koning and Lindeboom (2015)

[19] More precisely, the approach of Deshpande and Li (2019) relies on changes in costs (office closures) and its effect on applications. In the award decision there is no new information available and the adjudicator's standards remain constant.

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

and

$$\overline{H}_{award} = \frac{p\delta}{(1-p)+p\delta} \tag{6}$$

This implies

$$\delta = \frac{\overline{H}_{award}}{1-\overline{H}_{award}} \quad \frac{1-\overline{H}_{app}}{\overline{H}_{app}}. \tag{7}$$

The above expression shows that an estimate of the relative conditional award rate δ can be identified from the averages of H among the samples of awardees and all applicants. Analogous to our estimates on compositional changes among DI applicants, we can run RDiT regressions of the reform effect on H for both these samples. This then allow us to infer the effect of the reform on δ, the conditional relative award rate for applicants with bad health. Increases in δ indicate that the conditional award rate of those in bad health increases relative to those in good health and thus that medical examiners manage to screen out additional workers in good health. In this case targeting efficiency improves in the final stage of the application process. If δ remains constant, this means that the conditional award rate of those in bad health and those in good health move in tandem. Or, conversely, that any changes in the composition of the pool of awarded applicants are fully driven by self-screening in the application phase of the waiting period. So, while Deshpande and Li (2019) make inferences about targeting efficiency on the basis of changes in conditional award rates, our context requires us to look at *relative* conditional award rates. In this way we can separate self-screening effects in the waiting period from screening effects of the medical examiner at the end of the waiting period.

5 Estimation results

5.1 The extensive margin: DI applications

Table 4 presents the main RDiT and Donut RDiT estimates for DI applications using equation (1), with next to the coefficient standard errors in parentheses and the implied percentage change in brackets. All models in Table 4 use a bandwidth of 13 months on each side of the threshold.[21] The first panel shows the RD estimate, while the four other panels present Donut RDiT estimates with varying width of the "hole" around the cut-off. Removing observations within one month from the cut-off increases the treatment effect substantially, while reducing the standard errors.

[21] Varying bandwidths do not influence the results (see Figure A3 in Appendix A).

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 organizational changes introduce additional noise in the data close to the threshold. As it seems, any delays in the application system increased in December 2002, where end-of-year effects are common anyway. This in turn may lead to underestimation of the treatment effect in the conventional RDiT specification. We therefore proceed with Donut RDiT specifications where observations within one month of the threshold are removed.

The effect sizes of our preferred specification imply a 39.3% decrease in the number of DI applications as a result of the introduction of the GKP. Note that this corresponds with the observed drop displayed in Figure 4. Consistent with Koning and Lindeboom (2015), this implies that the reform was one of the major contributors to the decrease in DI applications since the turn of the century. Recall from the discussion in the previous sections that the decrease in DI applications is likely to stem from two sources.[22] First, effective reintegration measures may increase work resumption rates. Second, self screening effects: workers decide to drop out during the waiting period. This decision may be due to unexpected 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.

Table 5 reports estimation results of the reform effect by type of disorders. We categorize disorders into difficult-to-verify and easy-to-verify impairment types. The severity of musculo-skeletal, mental and non-categorized ("other") disorders is typically more difficult to verify with tests than other disorders. The top row of Table 5 repeats the result of our preferred specification for all impairments in Table 4. Subsequent rows of Table 5 show that the most sizable declines are observed for difficult-to-verify impairment types: 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), most of the total decline in applications comes from mental and musculo-skeletal disorders.[23] While this suggests that after the reform there was more scope to reduce DI applications for these conditions, we cannot be certain that the drop comes from the least unhealthy workers. For this, we need additional analyses that further characterize the decline in applications.

[22] See the discussion in Section 4 of the treatment effect β where we argued that the reform was new to all parties involved, that anticipation effects are likely to be limited and that it is expected that that ex-ante self-screening effects are limited.

[23] We observe a sizable decline (45.9%) for respiratory disorders. These disorders, however, account for only 1.5% of all DI applications.

5.2 Characterizing compliers and non-compliers

We argued earlier that changes in the composition and average health conditions of the pool of applicants may be informative on targeting efficiency of the DI screening system. A worsening of the health condition of the applicants may be indicative of improvements in targeting efficiency, but does not rule out that some of the deserving, truly disabled, decide not to apply. To uncover such effects, we estimate the RDiT estimates of various measures for the composition of workers. Based on these estimates and the effect on (total) DI applications, we next calculate the implied averages of composition variables of individuals who also applied after the reform and compliers to the reform that did not. 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). Column 1 presents RDiT estimates of the impact of the reform on a given characteristic, columns 2 and 3 the average X of non-compliers and compliers, and column 4 the differences of averages between these two groups.

Of most importance for targeting efficiency among the applicants are the results shown in panel A. The effect of the reform on the Charlson index is positive and substantial, implying a worsening of the health conditions in the pool of applicants. The results for the health index show that the non-compliers are in much worse health (average index of 0.27) than the compliers (average index of 0.18). Additional results (not in the table) show that the Charlson index increases by 16.8% for applicants with difficult-to-verify disorders, but only by 9.7% for applicants with easy-to-verify ones (both effects are significant at the 5% level, and significantly different). Even though we find no significant composition effect on mortality rates, these results suggest that the reform disproportionately decreased the DI applications of most able workers.[24] At the same time, the average index for compliers is considerably higher than the averages for non-applicants – see Table 2. So while compliers are healthier than non-compliers, they are still far from being as healthy as the non-applicants.

Turning to panels B and C of Table 6, we observe important shifts in disease types, age and gender. The share of women among the compliers is about 16 percentage point higher than the share of women among the non-compliers. Similar changes occur for prime aged workers: prime age workers are over represented among the compliers. To examine this in greater detail, we re-

[24] Although differences are not statistically significant, Figure A4 indeed suggests that applicants under the GKP regime are at a higher risk of death within the next five years due to composition effects. The difference in the two distributions reflects the mortality differential that is "explained" by group differences in observable characteristics of applicants.

produced column 1 of Table 6 for females and males separately. The results of these analyses show that the effects for age and impairment types are primarily driven by females. For instance, the 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 between the age of 35 and 49, the effects are -0.037 and -0.007 for females and males, respectively. Interestingly, the coefficients of the Charlson index for males and females are about equal in size, but only statistically significant for men. So the changes in the shares of easy- and difficult-to-verify diseases for women do not translate into significant changes in the health index for women. In this respect, it should be noted that difficult-to-verify impairments primarily consists of mental and musculo-skeletal impairments; these are conditions that are poorly represented in the Charlson Health Index (see the definition that is explained in Data Appendix D). From all this, it follows that the average health condition of the pool of applicants worsens and that the most important shifts come from the application behavior of prime aged women with difficult-to-verify impairments.

For a complete assessment of the effects of the reform on targeting, we also consider changes in the pool of non-applicants. Stated differently, we compare compliers – that is, 'former applicants' that are added to the pool of non-applicants – with the pool of non-applicants under the counterfactual without the GKP reform. For this we conduct similar regressions and calculations for the sample of non-applicants as for the sample of applicants. Table 7 shows the results of this exercise. Column 1 reports the effect estimate, column 2 the average of the characteristic of compliers and column 3 the average of the characteristic of non-compliers among the non-participants. We focus on one year lead variables of the health, mortality and labor market variables, as longer run outcomes may render it more difficult to be interpreted as the result of the reform. When comparing columns 2 and 3, we infer that compliers who are added to the pool of non-applicants are in worse health. Hospitalization rates among the compliers are twice the rate of the initial group of non-applicants and mortality rates are as much as eight times higher. Both differences are significant at the one percent level. Perhaps more strikingly, we also find that the compliers have substantially lower incomes, lower employment rates and much higher UI benefit incidence than the initial group of non-applicants. In particular, the implied quasi-elasticity of UI receipt due to less DI application is about 0.36.[25] This suggests that there are non-negligible spillovers to UI from the reform.

Even though the effects on health outcomes for the applicants indicate improvements in targeting

[25]Recall 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.

efficiency, we conclude that compliers to the reform generally consists of individuals with weak labor market positions – that is, they are in worse health, have lower incomes, lower employment rates and higher unemployment rates. On the other hand (and on a more positive note), a substantial part of the this group (83%) are one year later still at work.

5.3 Are there 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 these changes generally have improved the targeting efficiency of the screening system. From this alone, however, we cannot assess the overall implications of the reform on targeting efficiency in terms of DI awards. As argued earlier in Section 2, the GKP provided medical examiners with information on the reintegration activities that were planned in the sickness period. While these examiners were explicitly instructed to maintain the DI eligibility criteria of the old system, more information at the point of the award decisions may have reduced classification errors and further improved targeting efficiency.

To examine how the additional information for medical examiners has affected outcome decisions, we follow the approach that was unfolded in the previous section. The intention is to compare relative changes in averages for the samples of applicants and awarded applicants before and after the reform. Panel A of Table 8 shows results for all impairments, panels B and C for difficult-to-verify and easy-to-verify impairments, respectively. As a dichotomous measures for (bad) health conditions, we define a dummy value that is equal to one if the Charlson index is larger or equal to one (and zero otherwise) and an indicator for mortality within five years. We also include the fraction of individuals with earnings that are in the bottom quartile of the earnings distribution of applicants and non-applicants. With this variable, we aim to infer the relative changes in application and award rates for workers with a vulnerable position in the labor market.

In line with expectations, panel A of Table 8 also makes apparent that health condition of awarded applicants is worse than the health condition of the total pool of applicants. This also holds for the mortality rate for individuals with difficult-to-verify impairments (panel B) and for those with easy-to-verify impairments (panel C). Note that in general the health and mortality indicators of panel C are much higher (i.e. worse) than those in panels A and B. For workers in the bottom quartile of the income distribution, however, we see opposite effects, with fewer low wage workers among the awarded applicants. Using equation (7), we infer for all impairments an estimate

of the relative conditional award rate δ_0 equal to 1.349. This implies that in the old system 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 lower for easy-to-verify impairments (panel C) than for difficult-to-verify impairments (panel B), suggesting that medical examiners have more value added in terms of screening for applicants with difficult-to-verify conditions.

When comparing columns 1 and 4 of Table 8, we again observe health declines between the pre- and post-Gatekeeper periods. However, the odd ratios in columns 3 and 6 are virtually the same, leading to a ratio of the δ's (column 7) of about unity. Indeed, from the comparison of columns 2 and 5 we can infer that the health condition of the awarded applicants also worsens and that the relative health declines of the awarded applicants are similar to the health declines of all applicants. Stated differently, changes in the health condition of the pool of awardees are fully driven by self-screening during the waiting period rather than by additional gains in screening out healthy applicants in the final award decision.

5.4 The intensive margin: additional evidence from a field experiment

The results above show that the implications of the GKP reform were substantial, both in terms of numbers of DI applications and the composition of applicants. In this respect, it is important to stress once more that the treatment under investigation applies to the transition of a screening system without strong worker and employer obligations in the sickness period to one with a high level of scrutiny. This means we may characterize effect estimates as extensive margin effects. To supplement these findings with intensive margin effects, one needs information on the effects of changes in the screening regimes at the level of the NSII offices. These findings could inform to what extent further intensification of the screening process improves targeting and for which type of impairments this is most relevant. As we will argue later on, it will also provide additional insight in the potential mechanisms underlying the screening effects.

In the first year the GKP was implemented, the NSII conducted a field experiment in two regions: Apeldoorn and Hengelo.[26] The experiment started in January 2003 when the first DI applications under the new GKP act arrived at the regional offices of the NSII and ended in October 2003. For 24 out of the 26 regional offices in the control group, caseworkers were instructed to adhere to the standard protocol. That is, to screen the reintegration reports 'on paper' and to only deviate from this and contact the employer and/or the worker only if there was some suspicion of fraud

[26]For this reason, these two regions were left out of the analyses in the previous section.

and/or negligence. In the remaining two treatment regions the default was to contact the worker and/or the employer. To make sure that the treatment regions implemented the stricter screening policy, they were provided additional resources. Table B1 in Appendix B provides more detail on the distribution of the screening methods across the treated and control regions. The table shows that the treatment regions used more time intensive measures such as contacting the worker and/or employer, rather than checking on paper. In the region of Apeldoorn caseworkers visited sick workers more often, while in the Hengelo region they more often contacted the employer.[27]

To estimate the effect of the intensified screening on applications, we need random selection of the treatment regions conditional on regional fixed effects. Similar to De Jong et al. (2011), we therefore test if in the treatment regions the applications changed differently between the pre-experiment years 2001 and 2002 than in the control regions – as shown in Figure 6. Without important differences in pre-experiment DI application rates, we can thus follow a Difference-in-Differences (DiD) approach that compares individuals in treated and control regions before and after the start of the experiment started. While De Jong et al. (2011) analyze the treatment effect of the experiment on sickness absence and inflow into DI in the first two years (2003-2004), our aim is to complement these findings with effects on the type of impairments.[28] More details about the empirical strategy can be found in Appendix C.

Table 9 shows the treatment effects that follow from estimation of the DiD model for all applications and by type of impairment. Most notably, the effect of intensified screening on DI applications is considerably smaller than the 'extensive' margin or overall effect of the introduction of GKP. Marginal increases in screening intensity in the treatment regions reduce applications by 4.5% which is comparable to the effect sizes found in De Jong et al. (2011). As it seems, this effect is solely confined to applications for mental disorders (a 16.7% decrease). Indeed, it is conceivable that for such impairments follow-up contact via telephone or face-to-face is more effective than for other impairments for which verification is easier. This is in line with findings of studies that find effects of compulsory dialogues (Markussen et al. (2017)) and external reviews (Liebert (2019)).

To shed more light on the mechanisms underlying the effects of intensified screening, we finally stratified the intensive margin effect by treatment region. We then find that DI applications dropped with 8.6% in the region of Apeldoorn, whereas the decrease was 3.6% in the region of Hengelo. This

[27] De Jong et al. (2011), using the same experiment, show that the screening in the treatment regions was indeed stricter than in the control regions. In particular, the time spent on screening reports in treatment regions by caseworkers was 40% higher than in control regions.

[28] As in the RDiT regressions of the previous sections, we only include the 13 months on either side of the date of the reform. This is different from De Jong et al. (2011) who exploit 2001-2004.

suggests that increasing screening effort on the worker is more effective than on their employer.

5.5 Our findings in perspective

Our analysis suggests that the main mechanisms driving compositional changes among DI applicants are increased self-screening among potential applicants and increased work resumption. Based on our estimates, these effects are substantial: DI applications fall on average with 39.3%. With average award rates of about 60-70%, this leads a decline in DI inflow rates of 24-28 percentage point. We also find strong changes in the composition of the applicant pool. As we will show below, these findings are roughly in line with findings from other related studies that consider the effect of changes in application costs or award rates.

To start with, Parsons (1991) shows that an increase in the rigor of initial eligibility screening discourages potential applicants. In particular, 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. Their findings indicate 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. These dialogues also 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. Low and Pistaferri (2015) find that disability insurance interacts with welfare programs. Finally, Deshpande and Li (2019) study the effect of application costs on DI inflow and targeting using the closings of Social Security Administration field offices. They find that field office closings lead to large and persistent declines in the number of DI recipients. They also conclude that office closings primarily discourage the most vulnerable workers: those with low income, low education and moderately severe health conditions.

While our results confirm the idea that increased screening may have large effects on DI application and inflow rates, they also suggest that the GKP reform was highly cost-effective. Applications fell from about 145,000 applications to about 75,000 applications in the first year of the reform. In the next years inflow rates remained constant. From the 75,000 applications, a little more than

60% resulted in a DI benefit of on average about 12.000,- euros per year.[29] 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 a upper bound of costs associated with increased screening by the NSII. As for employers, De Jong et al. (2014) estimate the additional costs of preventative activities and hiring occupational health services to be between 0.23 and 0.43 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.[30]

These simple back-of the envelope calculations do not take individuals welfare losses into account. In this respect, our results show that the larger part of those who are induced by the reform to stop applying are one year later still in employment (83%). About 14% and 3% of these workers end up in UI benefits and means tested welfare programs, respectively. After exhaustion of UI benefits entitlements, those on UI revert to means tested social welfare with substantially lower benefits. 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 are likely to respond to this in ways that were not intended by policymakers. Most notably, the reform may induce employers 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 high-risk workers have become less likely to have permanent contracts and there has been a strong increase of temporary and unemployed workers in the inflow into the DI system (Koning and Lindeboom, 2015).

[29] Note that this may also include partial DI benefits.

[30] This finding corresponds to e.g. Markussen et al. (2017), who show that the economic gains from compulsory dialogues with sick-listed workers are highly cost-effective.

6 Conclusions

In this study we examine the targeting effects of increased scrutiny in the application period of DI benefits. To this end, we exploit exogenous variation induced by the introduction of a system with strict screening (the Gatekeeper protocol; GKP) and a nationally implemented field experiment. In comparison to the old system the new protocol implied higher application costs and more information about the true disability status at the point of the award decision. This will affect application behaviour and award decisions. To relate the behavioral responses to changes in targeting efficiency, we first estimate a Regression Discontinuity in Time (RDiT) regressions on the effect of the start of the GKP on application rates and on the composition of the pool of applicants and non-applicants.

We find that the introduction of the reform reduced application rates with about 40%. These self-screening effects are likely to reflect the effect of increased work resumption and/or the decision of the worker to pull out of the application process during the waiting period. The largest declines occurred in conditions that are difficult-to-verify, primarily musculo-skeletal and mental impairments. These impairments account for about two-third of all applications. We furthermore find differential application responses by gender, income and age. This suggests substantial changes in the pool of applicants.

We next implement a simple method to uncover and characterize individuals that continued to apply after the reform (i.e. non-compliers) and those that do not apply but would have done so without the reform (i.e. compliers). We find that the health of the non-compliers is much worse than the health of compliers, implying that the pool of applicants becomes more deserving and that targeting efficiency has improved. Interestingly, women are strongly over represented among the group of compliers. These women are generally prime aged and have a high fraction impairments that are difficult-to-verify. However, a worsening of the pool of applicants does not rule out that some of the deserving workers decide not to apply for DI benefits; these are Type 1b errors in the definition of Kleven and Kopczuk (2011). To investigate the relevance of such effects, we compare the characteristics of compliers with the characteristics of those of non-applicants in the old system. We find that the compliers are in worse health than this group of non-applicants, have substantially lower incomes and are more often unemployed. This indicates that there are spillovers of the DI reform to other social insurance programs (cf Low and Pistaferri (2015) and Borghans et al. (2014)).

While the reform induced substantial self-screening effects, the additional checks and reintegration efforts during the absence period may also provided new information for the examiner at the

point of the award decision. This potentially improved the quality of the award decision. To examine such targeting effects, we show that the ratio of the conditional award rate of workers in bad health and good health can be identified from sample averages of the health of all applicants and awarded applicants. An increase in this relative conditional award rate indicates that the medical examiner manages to screen out additional healthy applicants, leading to (further) improvements in targeting efficiency. In doing so, we find that the reform did not change the relative conditional award rates. Changes in the health condition of the pool of awarded applicants thus are fully driven by self-screening of (potential) applicants during the waiting period.

Bearing in mind that the GKP formulates minimum standards for employers and workers in the absence period, a pertaining question is whether additional increases in screening stringency contribute to targeting efficiency. To investigate such 'intensive margin' effects of increased screening efforts, we exploit a nationally implemented field experiment that was set up in the year the GKP came into effect. In this field experiment, two out of the 26 Dutch regions implemented a stricter screening than in other regions. The results point at further reductions in DI applications for workers with mental impairments only. This suggests that intensified contacts provide value added for this group with difficult-to-verify impairments. We also detailed the intensive margin effect by treatment region. These additional analyses indicate that focusing the screening efforts on the worker is much more effective than on the employer.

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 the application process. The procedure examined here involves both the worker and employer, but leaves the main responsibility to comply to the requirements as laid out by the protocol to the employer. This set-up leads to strong self-screening effects and shows that important improvements in targeting can be obtained in the very first months of the sickness period. What is more, early work resumption (83% of those who are screened out are in work one year later) is a first guard against possible scarring effects 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 increased 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*.

Autor, D. and M. Duggan (2014): "Moral Hazard and Claims Deterrence in Private Disaibility 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. 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*.

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.

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, forthcoming.*

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., 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 sicklisted 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 (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*.

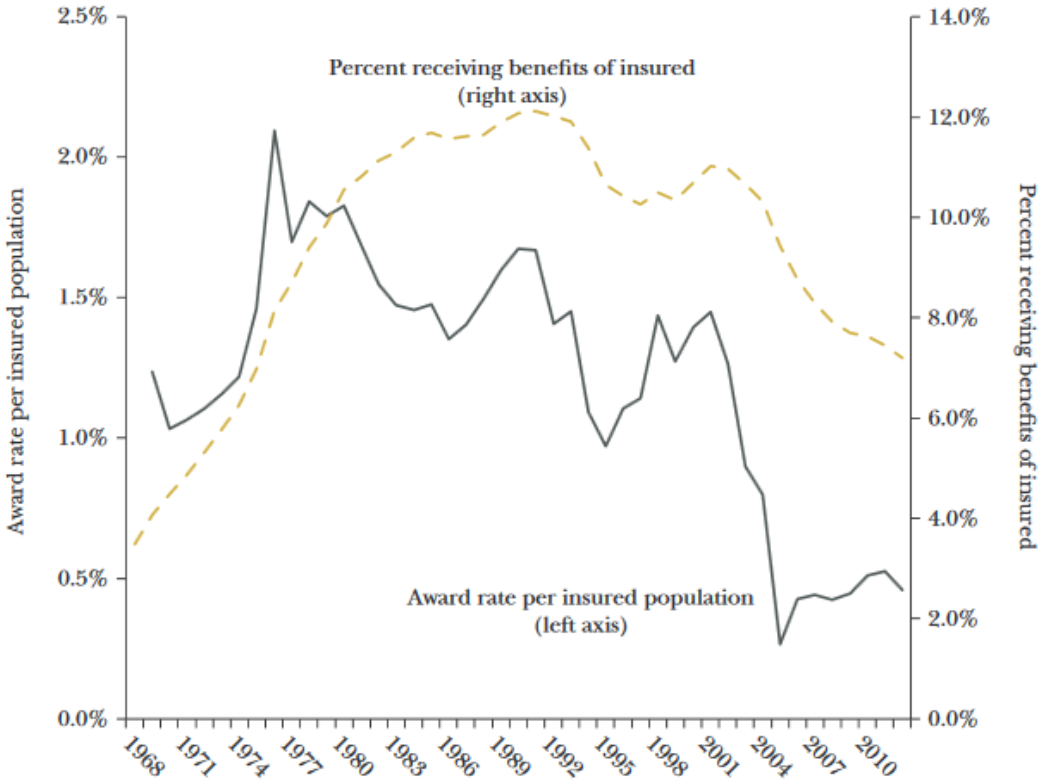
Staubli, S. (2011): "The impact of stricter criteria for disability insurance on labor force participation," *Journal of Public Economics*, 95, 1223–1235.

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.

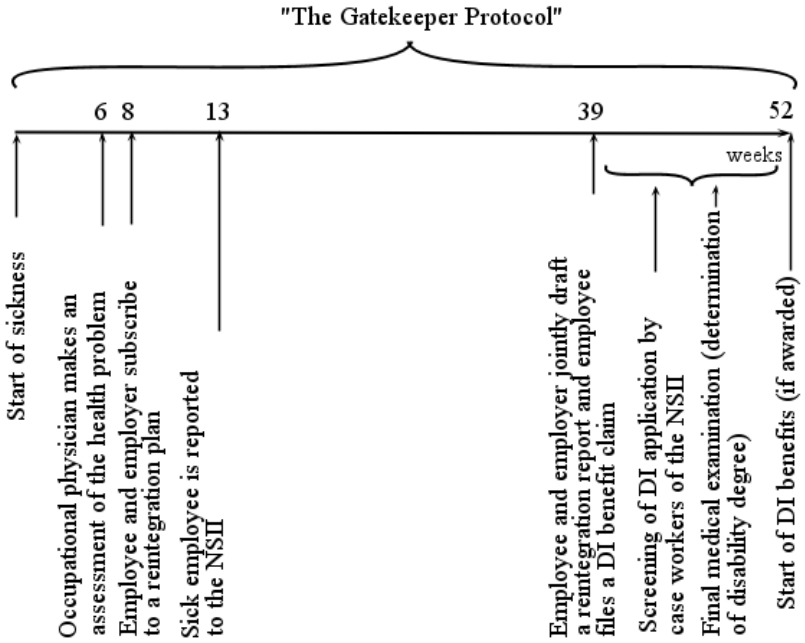
von Wachter, T., J. Song, and J. Manchester (2011): "Trends in Employment and Earnings of Allowed and Rejected Applicants to the Social Security Disability Insurance Program," *The American Economic Review*, 101, 3308–3329.

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



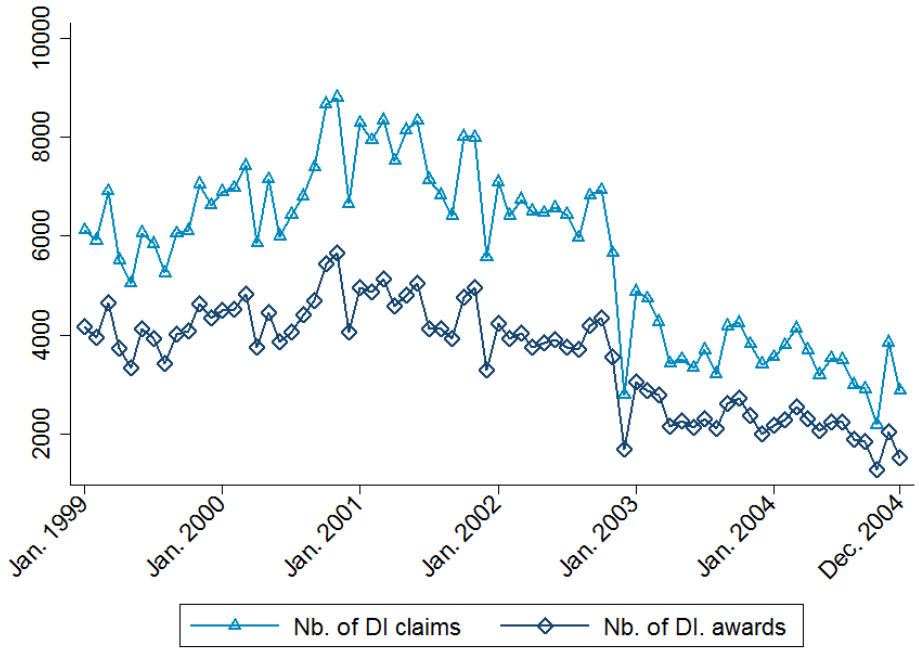
Note : The Disability Insurance award rate is the share of the insured population that started to receive disability payments in a given year. Source : UWV (2012). Borrowed from Koning and Lindeboom (2015).

Figure 2: Schematic representation of the process toward entering DI



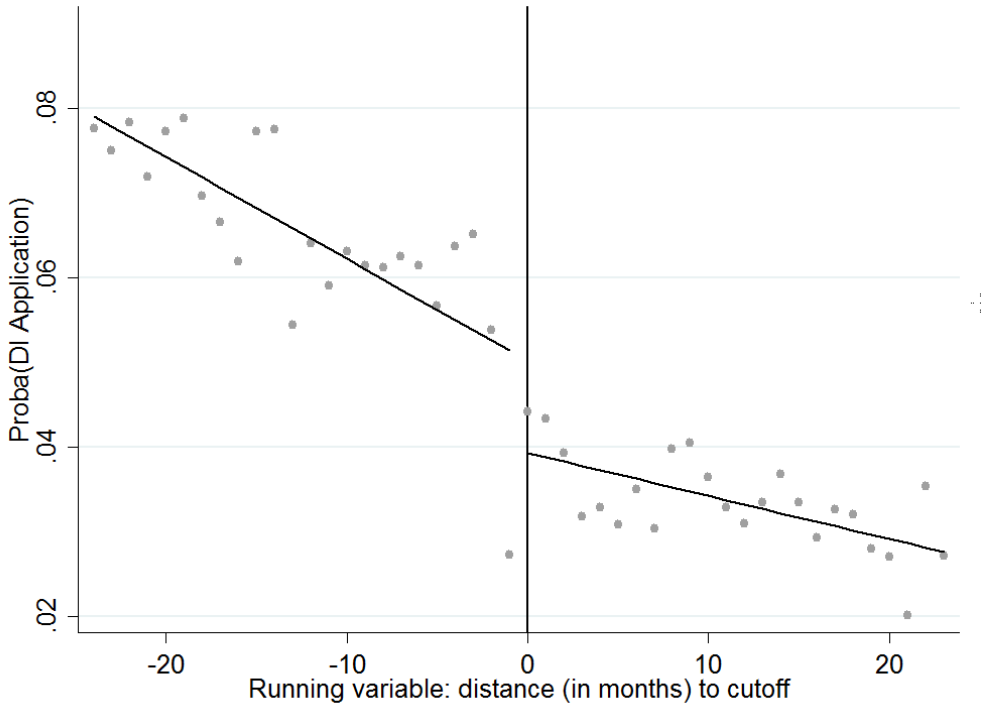
Note : Borrowed (and extended) from De Jong et al. (2011).

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



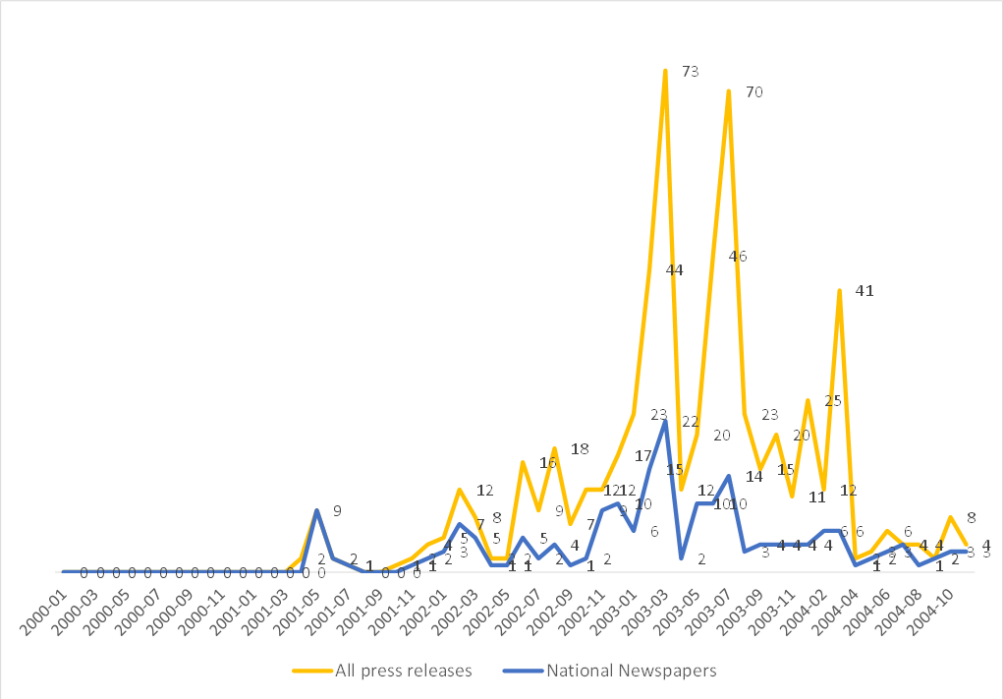
Note : The DI Application Inflow is the number of prime-age individuals who filled a DI benefit claim in a given year; The DI Award Inflow is the number of prime-age individuals that started collecting disability payments in a given year.

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



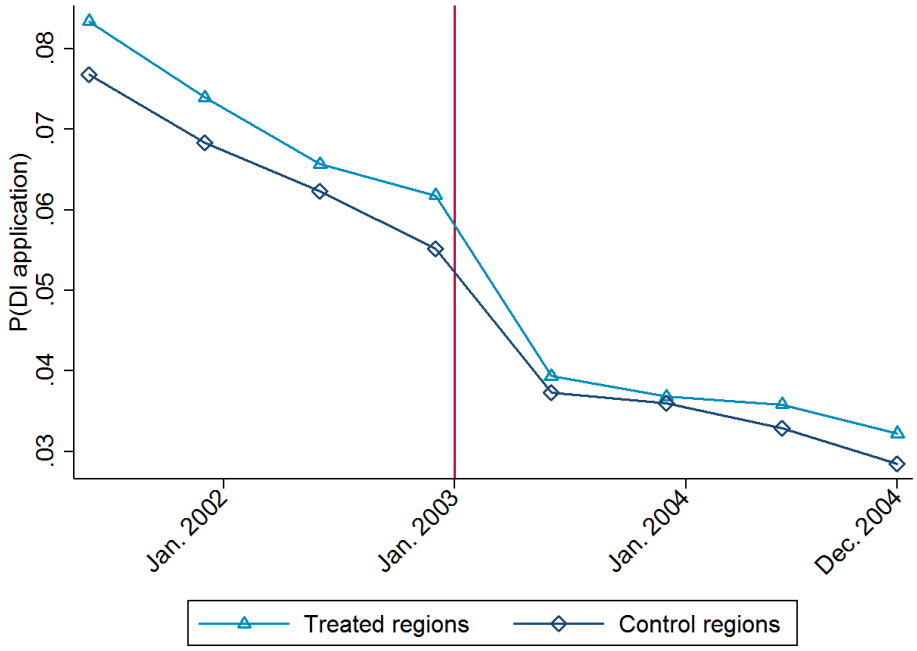
Note: For each month, we take all applicants that month and a 1% random sample of non-applicants that month to compute the monthly DI application rate – see Section 3. Thus, DI monthly application rates should be divided by 100 to reflect the average rate in the population.

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.

Figure 6: Monthly DI application rates (averaged over a 6-month period) in treated and control regions, before and after the field experiment.



Note : The vertical red line represents the start of the experiment (January 1, 2003), when the first DI applications under the new Gatekeeper protocol arrived at the regional offices of the NSII. 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.

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

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
	Applicants	Non-appl.	Applicants	Non-appl.	
	(1)	(2)	(3)	(4)	(5)
Demographics					
Age	43.30	38.37	43.81	39.02	-0.151***
	(9.82)	(9.46)	(9.88)	(9.53)	(0.044)
Male	0.41	0.56	0.48	0.56	0.075***
	(0.49)	(0.50)	(0.50)	(0.50)	(0.002)
Native	0.79	0.82	0.77	0.82	-0.019***
	(0.40)	(0.38)	(0.42)	(0.39)	(0.002)
Labour-market[(b)]					
Employed	0.90	0.95	0.83	0.95	-0.063***
	(0.30)	(0.21)	(0.37)	(0.22)	(0.001)
Gross earnings (2010 euros)	26,267	30,894	20,778	27,391	-1,703.6***
	(22,319)	(22,263)	(18,263)	(23,772)	(104.7)
UI recipient	0.11	0.03	0.18	0.05	0.044***
	(0.32)	(0.18)	(0.38)	(0.22)	(0.001)
Welfare recipient	0.03	0.02	0.04	0.02	0.006***
	(0.18)	(0.13)	(0.19)	(0.13)	(0.001)
Health status					
Health index	0.18	0.03	0.27	0.03	0.089***
	(0.88)	(0.34)	(1.14)	(0.34)	(0.002)
Hosp. past three years[(c)]	0.34	0.16	0.40	0.16	0.048***
	(0.47)	(0.36)	(0.49)	(0.37)	(0.002)
Dead within five years	0.03	0.01	0.04	0.01	0.011***
	(0.17)	(0.10)	(0.20)	(0.10)	(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 $(Y_{applicant,after} - Y_{applicant,before}) - (Y_{non-applicant,after} - Y_{non-applicant,before})$.

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

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

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

	DEPENDENT VARIABLE
	DI application
	Coeff. [% change]
	(se)
Panel A: All impairments	-0.024*** [-39.3%]
	(0.003)
Panel B: Difficult-to-verify impairments	-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)
Panel C: Easy-to-verify impairments	-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 specific impairment types). (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.

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

	(i)	(ii)	(iii)	(iv)
	Change average [% change]	**Non-compliers**	**Compliers**	**Difference (ii) and (iii)**
Panel A: Health and Mortality				
Health (Charlson) index	0.036*** [+18.6%]	0.27	0.18	0.087***
	(0.002)			(0.024)
Death rate within five years	0.002 [+6.0%]	0.042	0.037	0.004
	(0.002)			(0.005)
Panel B: Socio-Demographics				
Male	0.068*** [+16.5%]	0.48	0.32	0.166***
	(0.007)			(0.023)
Native	-0.103 [-13.0%]	0.77	0.79	-0.025
	(0.007)			(0.016)
Age				
Young (25-34)	0.027*** [+11.7%]	0.22	0.15	0.066***
	(0.004)			(0.013)
Prime age (35-49)	-0.026*** [-5.7%]	0.45	0.61	-0.063***
	(0.007)			(0.017)
Senior (50+)	-0.001 [-0.32%]	0.33	0.33	-0.003
	(0.005)			(0.013)
One-year lag earnings	658** [+2.9%]	23,833	22,227	1,606**
(gross, in 2010 euros)	(294)			(755)
Panel C: Share of impairment types				
Hard-to-verify impairments	-0.024*** [-2.7%]	0.86	0.92	-0.058***
	(0.008)			(0.008)
Easy-to-verify impairments	0.024***[+20.0%]	0.14	0.081	0.058***
	(0.008)			(0.008)
Nb. of individuals	107,297	38,886	29,525	107,297

Notes : (1) Each line in column (1) presents the estimated impact of the Gatekeeper on the average value of a characteristic (e.g. average health index) or the proportion of applicants with a given characteristic. Columns (2) and (3) present the average value/proportion for the pool of stayers (x_{stay}) and leavers (x_{leave}), respectively. (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 (1) are clustered both at the individual level and at the month-of-year level. Standard errors in columns (2) and (3) are obtained from seemingly unrelated regression, and clustered at the month-of-year level.

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**	**Compliers**	**Non-compliers**	**Difference**
Panel A: Health and Mortality				
Hospitalization (any type)	0.004*** [+5.9%]	0.14	0.07	0.071***
	(0.001)			(0.010)
Death rate	0.0002*** [+16.2%]	0.008	0.001	0.007***
	(0.0001)			(0.003)
Panel B: Labor-market outcomes				
Earnings	-437.7*** [-1.6%]	17,308	26,751	-9,442.26***
	(54.26)			(895.90)
Employment	-0.004*** [-0.4%]	0.83	0.92	-0.086***
	(0.001)			(0.023)
UI receipt	0.005^{μ} [+9.5%]	0.14	0.05	0.086***
	(0.003)			(0.021)
Welfare receipt	0.001* [+5.0%]	0.016	0.019	-0.003
	(0.000)			(0.006)
Nb. of individuals	2,150,213	29,525	1,061,504	2,150,213

Notes : (1) Each line in column (1) presents the estimated impact of the Gatekeeper on the average value of a characteristic (e.g. average health index) or the proportion of non-applicants with a given characteristic. Columns (2) and (3) present the average value/proportion of that characteristic for leavers from the applicant pool (x_{leave}) and non-applicants in the pre-Gatekeeper period (x_{na}), respectively. (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 (1) are clustered both at the individual level and at the month-of-year level. Standard errors in columns (2) and (3) are obtained from seemingly unrelated regression, and clustered at the month-of-year level.

Table 8: Inference on targeting efficiency from DI award rates: relative changes in averages for applicants ("δ"), before and after the reform.

	Pre-Gatekeeper Protocol			Post-Gatekeeper Protocol			
Variable:	$X_{app,0}$	$X_{award,0}$	δ_0	$X_{app,1}$	$X_{award,1}$	δ_1	$\frac{\delta_1}{\delta_0}$
Panel A: All impairments							
P(Health (Charlson) Index ≥ 1)	0.083	0.109	1.349	0.109	0.141	1.334	0.989
Dead within five years	0.033	0.045	1.39	0.042	0.056	1.38	0.992
Bottom quartile of earnings (one-year lag)	0.292	0.251	0.814	0.268	0.229	0.811	0.997
Panel B: Difficult-to-verify impairments							
P(Health (Charlson) Index ≥ 1)	0.058	0.075	1,32	0.079	0.102	1.32	1.003
Dead within five years	0.027	0.037	1.39	0.035	0.049	1.39	0.996
Bottom quartile of earnings (one-year lag)	0.299	0.259	0.821	0.274	0.235	0.815	0.993
Panel C: Easy-to-verify impairments							
P(Health (Charlson) Index ≥ 1)	0.269	0.290	1,11	0.297	0.322	1.13	1.016
Dead within five years	0.074	0.085	1.17	0.080	0.094	1.20	1.028
Bottom quartile of earnings (one-year lag)	0.241	0.208	0.830	0.232	0.200	0.828	0.998

Notes : (1) We use data in a bandwidth of 13 months on each side of the cutoff. (2) Reading note: $X_{app,0}$ stands for the average value of X for applicants in the pre-Gatekeeper period.

Table 9: Additional results from a field experiment: The effect of stricter screening of reintegration efforts by NSII caseworkers on DI application rates – DiD estimates for 2003.

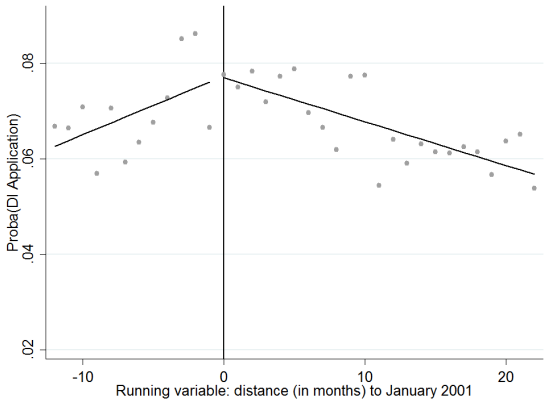
	DEPENDENT VARIABLE: DI application
	Coeff. [% change]
For impairment type:	(se)
All	-0.003* [-4.5%]
	(0.002)
Musculo-skeletal	-0.0005 [-2.6%]
	(0.001)
Mental disorders	-0.004** [-16.7%]
	(0.003)
Cardiovascular diseases	-0.000 [-3.3%]
	(0.000)
Nervous disorders	-0.000 [-0.1%]
	(0.000)
Respiratory disorders	-0.000 [-0.2%]
	(0.000)
Endocrine problems	-0.000 [-17.2%]
	(0.000)
Other	0.001 [+7.3%]
	(0.001)
Month*year dummies	✓
Regional dummies	✓
Background individual characteristics	✓
Nb. of obs.	3,563,624

Notes : (1) Column (1) presents the DiD estimates for 2003 (see Appendix C for more details about the empirical specification, and more specifically Equation (8).). (2) Each line presents the estimated coefficient associated with the treatment for a different outcome, i.e. DI application for a specific impairment type. (3) Estimates in brackets are presented in percentage change. (4) Standard errors are clustered at the regional level.

A Figures

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

(a) Placebo cutoff: January 2001



(b) Placebo cutoff: January 2002

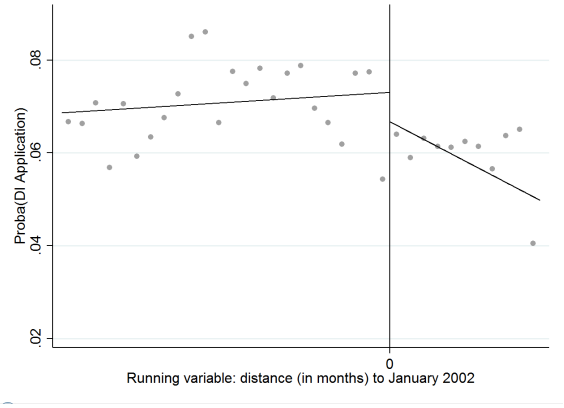
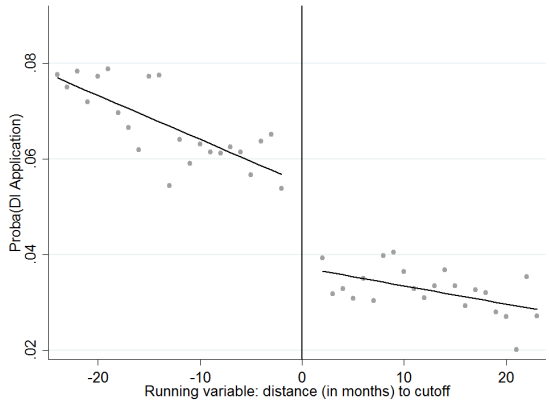
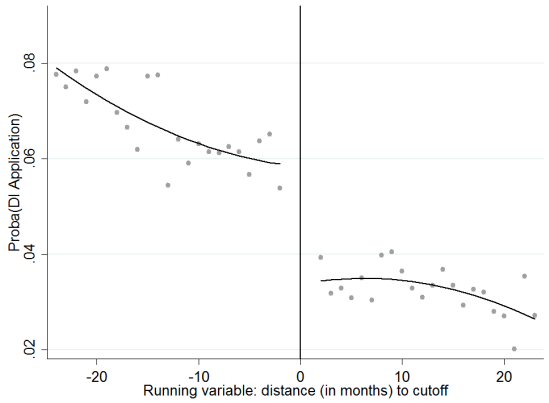


Figure A2: Donut Regression-Discontinuity plots (dropping those within one month of the cutoff) for DI application with various choices of polynomial time control.

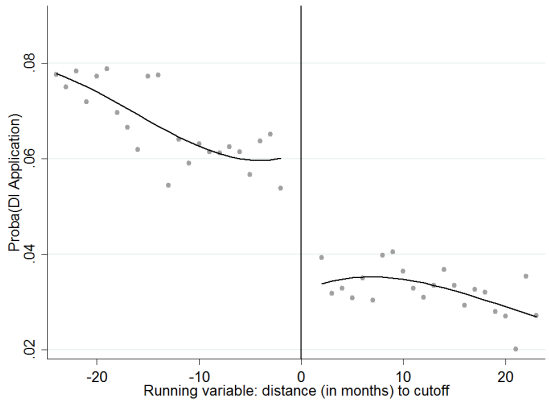
(a) Polynomial fit of order 1



(b) Polynomial fit of order 2



(c) Polynomial fit of order 3



(d) Polynomial fit of order 4

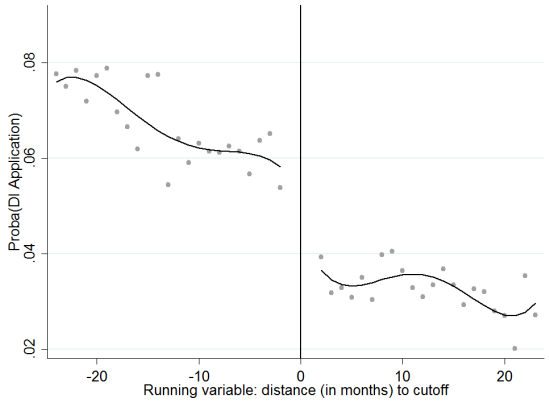
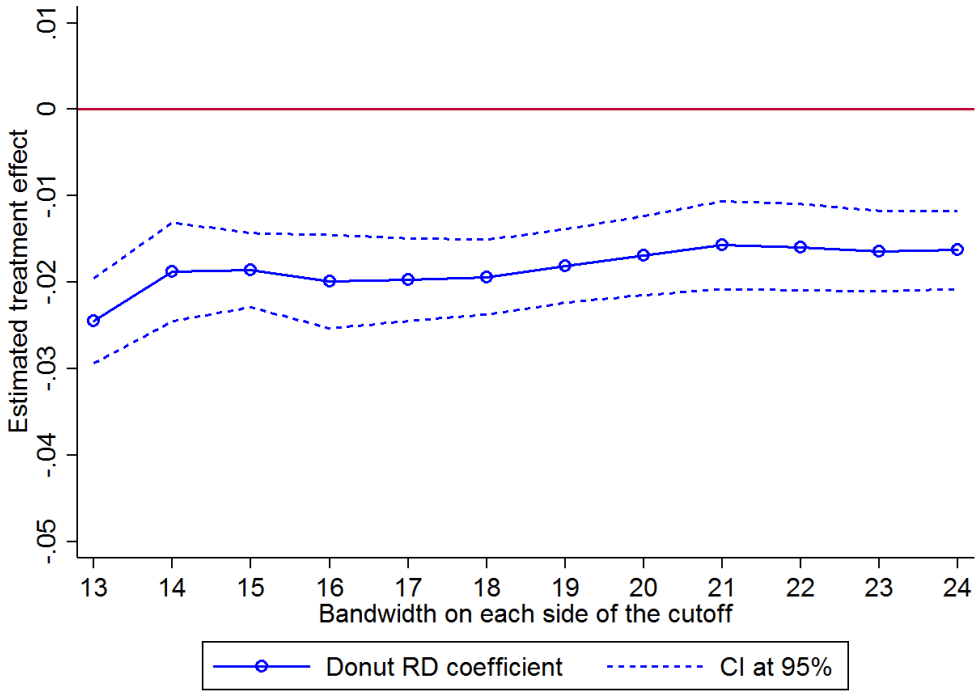
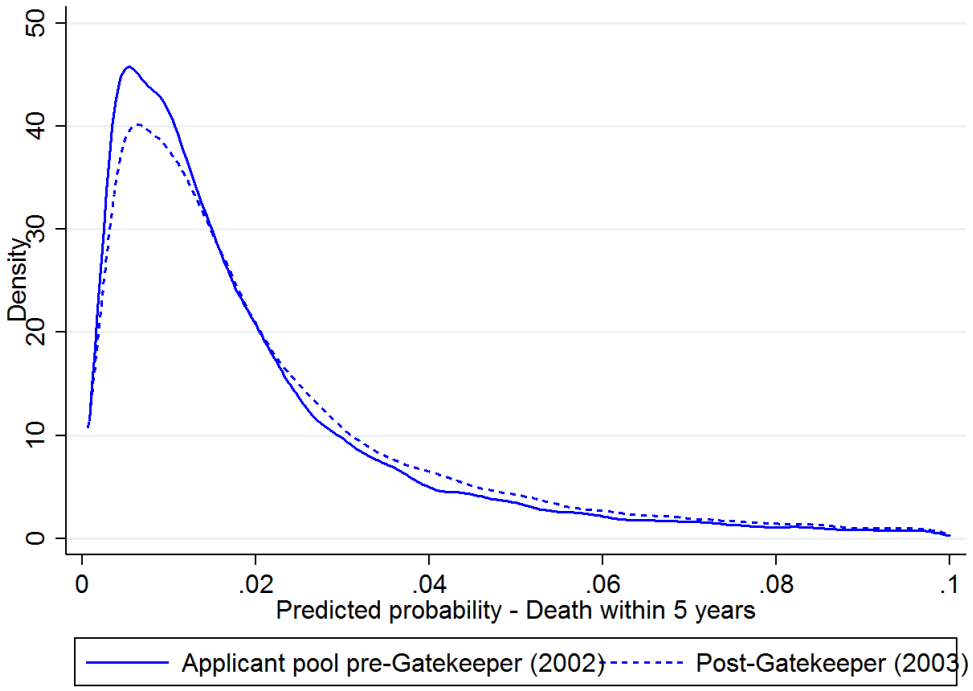


Figure A3: The effect of the Gatekeeper Protocol on DI application behaviour. Donut Regression-Discontinuity estimates with varying bandwidth.



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.

Figure A4: Distribution of (predicted) 5-year mortality rate of Di applicants, before and after the Gatekeeper reform.



Note: We predict 5-year mortality rates of DI applicants as follows: (i) we first run a model of DI application using data from the pre-Gatekeeper period. This yields a vector of coefficients β^{pre} (ii) we predict 5-year mortality rates of individuals who applied before the reform using their predictors values and slope parameters in β^{pre} (iii) we predict 5-year mortality rates of individuals who applied after the reform using their predictors values and β^{pre}. The difference in the two distributions yields the part of the mortality differential that is "explained" by group differences in the predictors (here socio-demographic, health, and labor-market characteristics).

B Additional tables

Table B1: Difference in screening stringency between treatment and control regions

	Treatment regions		Control
	Apeldoorn	Hengelo	regions
Only on paper	4%	14%	25%
Telephonic contact with employer	33%	34%	52%
Telephonic contact with worker	14%	14%	23%
Telephonic contact with occupational health agency	3%	12%	32%
Visit to employer	9%	41%	7%
Face-to-face contact with worker	77%	41%	7%
Unknown	4%	2%	

Notes : Borrowed from De Jong et al. (2011). Note that caseworkers can use multiple screening methods on one application, so columns can add up to more than 100%.

C Additional evidence from a field experiment: Empirical strategy

This appendix described the difference-in-difference approach that is followed to estimate the treatment effects of a field experiment that was conducted in two regions of the NSII. Our parameter of interest is the effect of stricter screening of reintegration efforts by NSII caseworkers on DI applications. We estimate the following individual level differences-in-differences model:

$$Y_{it} = \alpha + \delta_t + \lambda_r + \gamma_1 Treatment_{r,t} + \beta X_i + \epsilon_{it} \tag{8}$$

where $Treatment_{r,t}$ is the treatment variable. It is a dummy variable taking value 1 if region r= Apeldoorn/Hengelo *and* if month t is in year 2003, and value 0 else. γ_1 reflects the impact of increased screening stringency in 2003. δ_t and λ_r denote month*year and region fixed effects respectively. Our individual regressors X_i include gender, age (in 8 categories) and whether or not the individual was born in the Netherlands (ethnic background). Standard errors are clustered at the regional level.

The key assumption in the DiD framework is that the outcome in the treatment and control groups would follow the same time trend in the absence of the treatment. A first glance at Figure 6 in Appendix A shows that indeed DI application rates appear to move in parallel prior to the introduction of the GP. In order to test for this assumption more formally, we extend our regression model by interacting the treatment variable with yearly time dummies and take the last pre-treatment year (2002) as the reference category. The results of this regression show that the parallel trend assumption can not be rejected.[31] Furthermore, the parallel trend assumption cannot be rejected for all estimates presented in Table 9.

[31] The coefficient associated with the treatment effect for 2001 is equal to 0.001 (s.e. 0.001).

D 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
Metastasic Cancer	6
AIDS/HIV	6