HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

WP 18/21

External Medical Review in the Disability Determination Process

Helge Liebert

August 2018

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

External Medical Review in the Disability Determination Process

Helge Liebert*, †

April 23, 2018

Abstract

This paper investigates the effects of introducing mandatory external medical review for disability insurance (DI) applications in a system relying on treating physician testimony. Using a unique policy change and administrative data from Switzerland, I show that external medical review reduces DI incidence by up to 23%. Incidence reductions are closely tied to difficult-to-diagnose conditions, suggesting inaccurate assessments by treating physicians. Due to a partial benefit system, reductions in full pension benefit awards are partly offset by increases in partial benefits. More intense screening also increases labor market participation, indicating either moral hazard or substantial income effects on the side of applicants. Existing benefit recipients are downgraded and lose part of their pension when scheduled medical reviews occur. Back-of-the-envelope calculations indicate that external medical review is highly cost-effective.

*Center for Disability and Integration, Department of Economics, University of St. Gallen, Rosenbergstr. 51, 9000 St. Gallen, Switzerland. Email: helge.liebert@unisg.ch.

†I have benefited from discussions with Eva Deuchert, Beatrix Eugster, Per Johansson, Rafael Lalive, Michael Lechner, Nicole Maestas and seminar participants at the University of St. Gallen, the University of Uppsala/IFAU and the 2015 SOLE/EALE meeting in Montreal. All remaining errors are my own. This work was funded by the Swiss National Science Foundation under grant no. 100018_143317/1.

1 Introduction

Targeted social assistance is the most common form of welfare worldwide. Welfare payments are disbursed to groups identified by a common characteristic – families, the unemployed or persons with a work-limiting disability. Among these welfare programs, disability insurance (DI) is by far the most costly. In both the Unites States and the European Union, the number of beneficiaries has been rising for decades, and the average EU country spends about 2.3% of GDP on disability-related benefits alone (OECD 2010). Increases in DI beneficiaries have often been associated with imperfect screening of DI applicants (e.g. Autor and Duggan 2003). One indication for this is that the relative prevalence of difficult-to-diagnose health conditions like musculoskeletal or mental problems on the DI rolls has increased at a higher rate than prevalence in the general population (Campolieti 2002, OECD 2010). Across OECD countries, 60% of DI inflow can be attributed to muscoloskeletal conditions or mental health claims (OECD 2009).

In many countries, disability benefit decisions are made based on assessments of individuals' residual functional capacity, i.e. their remaining ability to work. For eligibility determination, the disability insurance agency consults the documentation of applicants' medical impairments from physicians. The first gatekeeper to the DI system is the treating physician, who diagnoses the health condition and forwards documentation to the insurance provider. Whether treating primary care physicians or clinical specialists should assess the residual functional capacity of DI applicants remains an open question. Treating physicians are considered to have an informational advantage, hence their recommendation is often influential in award decisions. The United States Social Security Administration (SSA) even adopted a 'treating physician rule' in 1991, giving 'controlling weight' to the treating physician's opinion. At the same time, treating physicians are known to diagnose clients favorably in the context of sick-listing, possibly to prevent harming a long-standing physician-patient relationship (e.g. Englund et al. 2000, Kankaanpää et al. 2012). Moreover, it is unclear whether complex or multidisciplinary disabling conditions can be accurately diagnosed by primary care physicians. In January 2017, the SSA rescinded the treating physician rule. Evidence from treating sources is now weighed equally along with evidence from DI physicians and other sources.

This paper investigates the introduction of systematic external medical review for DI applications. Medical review in this context means both file-based review and personal examinations by third-party physicians. In the main analysis, I focus primarily on the effectiveness of medical review for reducing DI inflow and the incidence of difficult-to-diagnose conditions in particular. Supplementary analyses trace rejected applicants' labor market response, the effects of medical review on benefit revisions in the recipient stock, and demonstrate that medical review is highly cost effective.

Relying on the introduction of external medical review in Switzerland, where treating

physicians' assessments are subjected to scrutiny by clinical specialists, I show that external medical review reduces DI admissions by about 23%. Reductions are closely tied to psychological and musculoskeletal conditions, diseases which are more prone to inaccurate diagnoses. Review screening also increases labor market participation, indicating either that DI applications are afflicted by moral hazard or the presence of substantial income effects. Existing benefit recipients are downgraded and lose part of their pension when scheduled medical reviews occur.

Identification relies on quasi-experimental policy variation generated by the introduction of a medical gatekeeper system in Switzerland. A 2002 reform made external medical review mandatory for all DI applications in certain regions. In doing so, the reform substantially increased the medical staff and funding directed towards reviewing DI applicants' cases. Previously, only few selected cases were reviewed. The measures aimed at improving screening quality by substantially reducing the individual DI physicians' caseload and by directing cases to physicians' specialized in the relevant field. The physicians are mandated to review all DI applications, to conduct medical checks if required and to provide the responsible DI caseworker with better information about applicants' health. Before the policy change, caseworkers relied on information provided by applicants' general practitioners (GPs) for their decision, as the DI offices had insufficient resources to screen individuals. The reform also abolished legal obstacles that prevented DI physicians from examining applicants in person. In the analysis, I use the sequential spatial implementation of the reform for identification in a difference-in-differences design based on matching local labor markets. This design is embedded in an age-based duration analysis framework for estimation.

Previous studies investigating the influence of screening on insurance inflow and enrollment have reported mixed results. De Jong et al. (2011) evaluate changes in caseworker stringency in the Netherlands and do not not find a reduction in incidence due to stricter screening but provide evidence for lower sickness absence and DI applications in the short-run, possibly due to self-screening (cf. Parsons 1991). A related study by Staubli (2011) for Austria shows that stricter eligibility requirements reduce insurance prevalence and affect labor supply. Given these results, this paper expands the available evidence. Since *prevalence* (enrollment stock) is likely to be inertial in many contexts, I focus on insurance *incidence* (inflow) in the analysis. While previous studies have often focused on prevalence, research has shown that incidence is the relevant policy parameter with regard to DI, since policies to induce work take-up among recipients are costly and often unsucessful (e.g. Bütler et al. 2015). In addition, I focus on medical screening measures, abstracting from mechanical inflow effects which arise due to implicit eligibility requirement changes.

Moreover, the paper contributes to three distinct strands of literature. First, it contributes to the economic literature on screening, difficult-to-diagnose conditions and

moral hazard in DI. The DI literature has traditionally focused on investigating moral hazard by evaluating the effect of program entry, benefit or stringency changes on labor supply (e.g. Gruber 2000, Autor and Duggan 2003, Mitra 2009, Maestas et al. 2013). However, medical screening and (mis-) representation of health status to the insurance provider (dubbed ex post moral hazard in the worker compensation literature, e.g. Staten and Umbeck 1982, Bolduc et al. 2002) is an aspect that has received less attention. Other studies have observed that individuals out of the labor market tend to overstate health limitations and self-reports differ from objective measures of functional limitations (Butler et al. 1987, Kreider 1999, Kreider and Pepper 2007, 2008), but actual evidence regarding DI inflow is scarce. Campolieti (2006) notes that stricter DI entry requirements are associated with less self-reports of difficult-to-diagnose conditions in the general population. Exaggeration and malingering of health limitations by patients in anticipation of insurance benefits has also been documented in medical studies (e.g. Frueh et al. 2003). Using administrative records, I provide evidence that medical screening both reduces insurance inflow of difficult-to-diagnose conditions and increases labor market participation.

Second, the results relate to the medical literature on physicians' conflict of interest in their role as care givers and social security gatekeepers (e.g. Carey and Hadler 1986, Duddleston et al. 2002, Bogardus et al. 2004). Research shows that physicians, especially GPs, tend to side with their patients when facing a trade-off (e.g. Englund et al. 2000, Kankaanpää et al. 2012). Surveys also indicate a general willingness to deceive insurance providers among a substantial minority of physicians if misrepresentation is deemed to be in the patient's best interest (e.g. Novack et al. 1989, Zinn and Furutani 1996, Freeman et al. 1999, Everett et al. 2011). However, these issues have never been directly tied to insurance take-up. I show that reviews by independent clinical specialists reduce benefit awards, indicating that incaccurate or lenient diagnoses by GPs result in a sizeable and costly number of disability benefit awards.

Third, an extensive theoretical literature investigates the implications of imperfect tagging in social insurances. The seminal work by Akerlof (1978) has been extended to include two-sided classification errors and applied to the DI context by Sheshinski (1978), Parsons (1996) and Kleven and Kopczuk (2011), among others. Few empirical studies have attempted to estimate the size of award errors. Given auxiliary assumptions, the results in this paper provide a tentative lower bound estimate of the false positive classification error rate in these models. Although not an exact quantification, this result, unlike earlier studies, does not rely on small sample expert reviews and the assumption of subsample perfect classification (Nagi 1969, Smith and Lilienfeld 1971, Benitez-Silva et al. 2004).

In sum, the paper provides three distinct contributions. First, I show that medical review by clinical specialists is cost-effective in reducing both inflow and stock of DI recipients. Second, I demonstrate that reductions are exclusively tied to difficult-to-diagnose conditions. Together with the fact that screening increases labor supply, this

suggests a combination of inaccurate diagnoses by treating physicians and the the presence income effects or moral hazard on the side of applicants. Third, I provide explicit conditions under which the inflow reduction implied by the reduced-form reform effect can be interpreted as a net reduction in DI award errors.

The paper proceeds as follows: The next section discusses the institutional setting and the role of medical screening in DI, section 4 covers identification and estimation methods, section 3 introduces the data, sections 5 and 6 present and discuss the results, section 7 concludes.

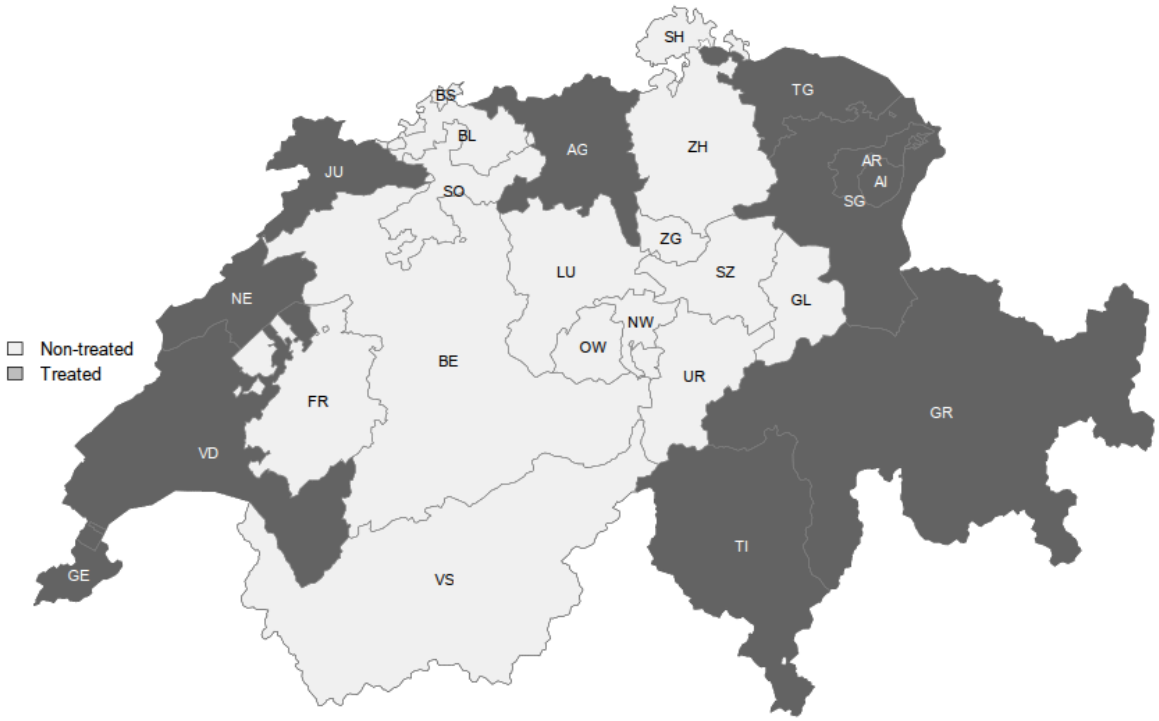
2 Institutional background

The Swiss DI system is characterized by generous benefits. Individuals can expect to receive between 60% and 95% of their final wage in benefits from the main insurance schemes if fully disabled.[1] Exact replacement rates are based on an individual's *disability degree*, a measure of work incapacity calculated as one minus the ratio of potential labor market income with disability to the potential income without disability (typically prior earnings). The determination of potential income is directly tied to a medical assessment of individuals' *residual work capacity*. If granted, benefits are paid indefinitely, and are only revised if applicants' health or earnings change substantially, or they become eligible for retirement pay. Unlike unemployment insurance, DI benefits are not attached to return-to-work measures. The Swiss system allows for partial disability benefits in quarterly increments.

The Swiss parliament passed a reform of the DI system in 2003 (*4. Revision des Bundesgesetzes über die Invalidenversicherung*). Prior to this, medical review occured infrequently and caseworkers made their decisions based on medical assessments submitted by the applicants' GP. The GP-based screening procedure had been in place unrevised since 1973. The reform created regional medical screening institutions, tasked to conduct (re-) appraisals of benefit claims and authorized to carry out medical examinations. This resulted in a large expansion of the medical staff available for review of insurance applications and substantially extended their legal competences. To assess the effect of the institutional changes, regional insurance offices could already hire new staff as part of a voluntary early adopter pilot project in 2002. 11 out of 26 Swiss cantons (states) chose to participate in this test scheme. In the remaining cantons, operation began in 2005 as scheduled by the

[1] The distribution of potential replacement rates is strongly left-skewed, even without children the average working-age individual can expect to receive upwards of 80% of his last wage (OECD 2010). Depending on the prior level of income, minimum benefits for a full pension amount to 1,160 CHF, maximum benefits to 2,320 CHF per month before taxes from the main public scheme alone. On top of this, people receive substantial additional payouts from occupational pension plans, family-contingent benefits for spouses and children, means-tested supplementary benefits or additional private insurance. Eligibility for payouts is determined by the local disability office and binding for all other insurance providers.

Figure 1: Cantons with medical screening offices



Note: Pilot cantons shaded gray. Legend: ZH: Zürich, BE: Bern, LU: Lucerne, UR: Uri, SZ: Schwyz, OW: Obwalden, NW: Nidwalden, GL: Glarus, ZG: Zug, FR: Fribourg, SO: Solothurn, BS: Basel-Stadt, BL: Basel-Landschaft, SH: Schaffhausen, AR: Appenzell A.-Rh., AI: Appenzell I.-Rh., SG: St. Gallen, GR: Graubünden, AG: Aargau, TG: Thurgau, TI: Ticino, VD: Vaud, VS: Valais, NE: Neuchâtel, GE: Geneva, JU: Jura.

reform proposal. Following the nationwide implementation in 2005, staff funding was then expanded further.

The variation in timing and spatial scope of the reform is exploited for identification in the remainder of the paper; using a combined spatial matching and difference-in-differences approach focusing on individuals in ex ante comparable subregions and comparing their development over time. The cantons that introduced screening institutions in 2002 are shown in Figure 1. The cantonal DI offices operate autonomously, but hold a yearly joint conference. Participation in the early adopter program was decided during one of the conferences; the head managers of the cantonal DI offices could choose to opt-in (endogenous self-selection is addressed in more detail in section 4). The program was fully funded by the federal ministry.

To become eligible for DI, individuals have to register with their local DI office. When filing a benefit claim, applicants have to submit the medical documentation of their condition and their previous earnings records. The earnings loss induced by the condition must span at least twelve months to qualify for benefits. The disability insurance office then assesses the individual earnings loss based on the severity of the condition and its impact on work capability. Based on the assessment, the caseworker makes a decision

whether the person qualifies for benefits.

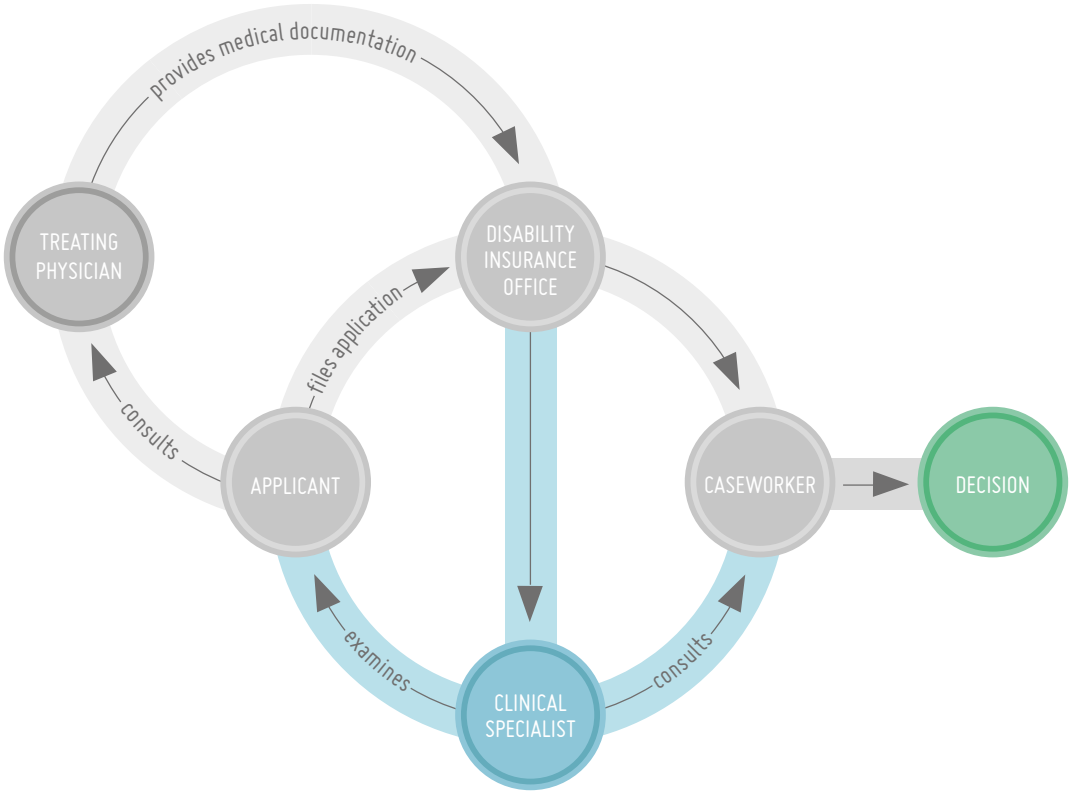
Prior to 2002, the insurance office could only assess eligibility from the medical certificates issued by the applicant's chosen doctor, typically the applicant's GP. DI offices were legally not allowed to examine the applicant, even when in doubt about the credibility or severity of the impediment. The DI caseworkers deciding on the application have no medical training themselves, but can consult with public health officers for clarification if they deem it necessary. However, the DI offices were notoriously understaffed with physicians. In one office, two doctors were reviewing about 40 applications each per day. This implies that on average, each application had to be dealt with in less than twelve minutes. For this reason, only a small subset of selected dossiers were passed to the physicians for inspection. Public officials were reliant on the medical assessment provided by the treating physician when awarding benefits.

This situation changed with the introduction of the screening institutions. The reform essentially strengthened the role of physicians in the application process by increasing the independent medical staff working at the DI offices and by substantially extending their competencies. In the abovementioned office, the medical staff increased fivefold due to the additional funding during the pilot period. Increases in other regions were similarly drastic. The physicians already working at the offices were reassigned to the new screening offices. The new staff consists of clinical specialists, which are then purposely trained in actuarial regulations. New physicians are selected to have specialized in fields relevant to diagnose difficult cases (e.g. chronic pain, musculoskeletal conditions or mental problems). In addition, physicians were given the power to screen people in person and order further examinations with other specialists. Before, reviews were legally restricted to file-based review. The staff is instructed to focus on new DI applicants and aid with scheduled revisions of existing beneficiaries claim status.

A schematic overview of the application process and the additional processes is depicted in Figure 2. Under the new system, the responsible screening office always receives a complete copy of an individual's insurance application, including the medical documentation of potential limitations. The office then provides an evaluation of the applicant's eligibility for the DI caseworker. If the documentation is considered insufficient, additional information can be requested. Furthermore, if the physicians notice inconsistencies in the application or deem it to be invalid, they have the authority to conduct further examinations or order specialist consultations.[2] The screening offices frequently use the available channels to gather additional information: Aggregate figures suggest that in-house examinations occur in up to 10% of cases, specialist consultations are decreed in up to 12% of cases and special multidisciplinary reports when multiple conditions are present are requested in up to 6%

[2] Examples for inconsistencies are an applicant claiming benefits on grounds of depression without a sufficiently documented history of therapy or medication, or an individual with moderate chronic pain claiming full work incapacity.

Figure 2: The DI application and decision process



(Wapf and Peters 2007).

In addition to the direct medical assessment, the audit offices are supposed to help public officials to better assess the actual implications of diagnoses and their impact on an individual's ability to work in relation to the insurance requirements. The treating physician typically supplies a diagnosis, suggests that his client is disabled and provides an approximation of the patients functional limitations. However, disability is a legal status determined by whether the applicants residual functional capacity meets the insurance requirements. The caseworker who knows the actuarial requirements lacks the judgement to assess capacity for work from medical diagnoses. The dual training of the employees at the screening services and the non-technical report provided for the DI office is supposed to reduce communication deficiencies.[3] Furthermore, the new legal framework also allows for the possibility to simply request additional information from the treating physician or other specialists if required.

The screening physicians' eligibility evaluation is not binding for the DI office. The final decision on whether benefits are granted remains with the responsible insurance caseworker and the actuarial requirements are the same. This implies that the regulatory

[3] A qualitative evaluation of the screening offices' work during the pilot program commissioned by the Federal Ministry of Social Insurances mentions that the institution has improved internal communication and reduced knowledge disparities between physicians and insurance offices (Wapf and Peters 2007).

framework remains unchanged, only the provision of information about the subjects' eligibility regarding health limitations is affected by the reform.

3 Data

The main analysis regarding insurance inflow and the analysis of the labor market response are both based on the SESAM (*Syntheserhebung soziale Sicherheit und Arbeitsmarkt*) data set provided by the Swiss Federal Statistical Office. The SESAM data link the official Swiss labor force survey to administrative public insurance records. The sample period ranges from 1999–2011. SESAM is a rotating panel which tracks individuals for five years until they drop out and each year 20% of individuals are resampled. Due to the small incidence of disability insurance in the population (around 0.5% per year) and the limited number of individuals that can be tracked over several years, the longitudinal dimension cannot be used for the analysis. Instead, the most recent observation for each individual is used, resulting in a large dataset of repeated cross-sections (the choice of observation does not influence the results in the paper). Given the survey weights, the data is representative of the Swiss population.

The treatment region is defined as the cantons participating in the pilot project and the treatment period comprises the years 2002–2004. The main outcome in the analysis is DI inflow. Inflow is measured using the age of first disability benefit receipt as registered in the data. The data also contains information about the specific health limitations that ultimately lead to the DI award. In addition, the data provides a rich set of information about income, labor market history, current welfare receipt, education, family background and a wealth of other socio-economic characteristics.

The SESAM data also contain information about individuals' municipality of residence. The empirical strategy outlined in section 4 partly relies on a local estimation approach due to the spatial treatment assignment and requires detailed geospatial information. I use the location information in the data to identify individuals in the vicinity of administrative borders. I augment the location data with information about distances between municipal centroids obtained from www.search.ch. For each municipality, I compute the distance to the nearest treated/non-treated counterpart that was sampled in the same year.

Based on this, I construct two estimation samples from the SESAM data, a *global* sample (containing all individuals in all regions) and a *local* sample (containing only individuals near the border between treated and control regions). Distance information is available as both actual travel distance and travel time by car. I choose a travel distance of 20 kilometers between municipalities as the threshold for the local estimation sample. Microcensus data on mobility show that 80% of commuters stay within this distance limit, and it corresponds approximately to the average commuting distance and time in

Switzerland (BSV 2012, Eugster and Parchet 2011).

Following this procedure, I obtain two distinct samples. The unrestricted global sample comprises 259,323 individuals in all Swiss regions. The local sample is restricted to individuals in local labor markets near border regions between treated and control cantons and comprises 133,549 individuals. Descriptive statistics for both estimation samples are given in Table A1 in the appendix. All results in the paper are robust to the choice of distance measure, variations in the threshold level and whether weights are applied. The sample composition is mapped in Figure A1 in the appendix.

A follow-up analysis investigates the effects of medical review on existing beneficiaries. For this analysis, I use a second dataset provided by the Swiss Federal Ministry of Social Insurances. The administrative data tracks the stock of all existing DI recipients from 2001 onwards. Since the data is selected conditional on benefit award, it is unsuitable for analysis of the DI population hazard. However, it can be used to investigate disability degree classification or benefit payment changes in the beneficiary stock. For each individual, I also observe the age of entry and the time spent on the DI rolls. In addition, the data register the actual disability degree, the benefit amount paid out by the state insurance and the health limitations the person suffers from, among other socio-economic variables. However, the stock data only register the region of residence, rendering localized analyses impossible. All stock analyses condition on individuals with benefit receipt prior to treatment in 2001, such that results are unconfounded by new entries to the DI payroll.

4 Empirical strategy

4.1 Identification approach and estimation method

The main quantity of interest is the change in the population DI hazard induced by external medical review, i.e. the change in the rate of newly awarded benefits among previously non-receiving working-age individuals. However, due to an opaque political decision process and self-selection into the early adopter scheme, treatment assignment cannot be assumed to be fully random. The cantons participating in the pilot program are a mixture of high and low prevalence regions, and regional cooperation considerations were relevant in the assignment process.

A difference-in-differences identification approach is used to evaluate the impact of the medical review institutions. Differencing removes time-invariant influences on potential outcomes. This removes bias due to selection into the program based on fixed or inertial aggregate regional differences. However, identification still requires a common development of DI incidence in the absence of the expansion of medical review. This assumption raises concerns related to regional heterogeneity and selection.

As Autor and Duggan (2003) illustrate, people rarely transition directly from employ-

ment into DI, but typically apply conditional on job loss. One concern in the present context is that labor markets may be less resilient in some regions, or that regions with strong industrial and commercial hubs are more affected by common economic shocks. Similar arguments are likely to have been relevant for the decision to participate in the early adopter scheme. If screening is imperfect and disability insurance is used as an extension to unemployment insurance or an early retirement vehicle in case of job loss, differential labor market trends can confound the results. Since Switzerland is a country with historically tight labor markets, such concerns are alleviated to some degree. Nevertheless, there may also be other underlying differences between regions based on the self-selection into the pilot program that cause time-variant divergence. Remaining time-variant heterogeneity among Swiss regions may raise concerns about biased treatment effect estimates.

To address this issue, I follow a twofold approach. A first set of results is based on the full sample of individuals across all regions. A more narrow identification approach focuses on individuals within the same local labor markets in border regions between treated and control areas. Focusing on these regions generates samples that are balanced in observable characteristics ex ante and increases the credibility of the common trend assumption. Similar strategies are used by Frölich and Lechner (2010) and Campolieti and Riddell (2012). However, local estimation approaches relying on sampling based on the distance to a border can suffer from problems due to spatial clustering on different sides along the border (cf. Keele and Titiunik 2016). To alleviate these concerns, I compute weights corresponding to nearest-neighbor pairwise differences and use them in the estimations.

For estimation, I exploit the spell format of the data and model insurance take-up as a duration problem. The main specification uses a stratified Cox (1972) proportional hazard model to estimate the impact of the reform on DI incidence. The hazard rate is modeled as

$$h(t, P, D|X < \bar{x}) = h_{0g}(t) \exp\Big(\beta_0 P + \beta_1 D + \beta_2 PD\Big) \,, \tag{1}$$

where $h_{0g}(t)$ is the non-parametric baseline hazard within birth cohort stratum g, t denotes time in years, $D \in \Big\{0, 1\Big\}$ is a binary treatment group indicator and $P \in \Big\{0, 1\Big\}$ is a binary time-varying indicator for the pilot period during $t \in \Big\{2002, 2003, 2004\Big\}$. Samples are restricted to local labor markets in border municipalities between treated and control regions within an absolute distance threshold $\bar{x}$ (20 km in the main specification), where individuals are similar in observables and remaining differences can credibly be assumed to be time-constant.

The model is specified using age as the time scale. This is preferable to using time-on-study as analysis time due to the age-dependent nature of the disability hazard, the rich cohort data available and the interest in the effect of a time-varying covariate (Kom et al. 1997, Thiébaut and Bénichou 2004). All models are stratified by five-year birth cohorts

to account for cohort-specific differences in health environments. Individuals become at risk when they are eligible for insurance at age 18. Censoring occurs at the sampling date or when individuals reach the retirement age, whichever occurs first. Disability benefit receipt constitutes failure. Due to data limitations, the analysis is restricted to single spells and disability insurance is assumed to be an absorbing state. However, this is not much of an abstraction. Actual outflow rates due to reasons other than death or moving to the old-age pension system amount to less than 1% of the stock per year (BSV 2012). Previous research has shown that DI recipients are unlikely to give up safe benefits even when faced with strong financial incentives to do so (e.g. Bütler et al. 2015).

A duration approach has a number of advantages compared to a linear difference-in-differences framework in this setting. It corresponds naturally to the spell format of the available cross-sectional data and the fact that DI entry is essentially a survival outcome. Data issues also limit the feasibility of the standard difference-in-differences approach. DI receipt is observed retrospectively as year of entry and only repeated cross-sections of a representative sample of the population are available. Since total DI incidence in the population is low, actual DI entry observed in each sampling year is low and insufficient for the analysis. Note that DI entry year and sampling year can be distinct. As the DI entry year is observed for each recipient, irrespective of the sampling date, pooling all data increases power substantially. This is due to the fact that all information on DI entry in any given year that is available from all subsequent years in which data was sampled can be utilized.

Pooling all cross-sectional data and conducting the analysis by age instead of sampling year (time-on-study) also limits the possibility of implicit sampling bias. With inflow observed retrospectively, relying on absolute sampling time would require creating a pseudo-panel structure by inferring past incidence figures from a post-treatment cross-section and adjusting for past eligibility. Since the disability risk is concentrated at older ages near the official retirement age, extrapolating past incidence causes bias due to intermittent entry into the retirement scheme. A non-negligible share of those in the old-age pension system at the sampling date may have received DI previously, but are not observed to do so any more when they are sampled. This share will increase the further past incidence figures are inferred retrospectively. Incidence figures inferred this way will be artificially low and the cross-sectional data ceases to be representative.

Finally, estimation of effects on incidence rates in a standard difference-in-differences framework would require modifying the standard common-trend assumption in a way which prohibits a more detailed analysis. Since incidence is defined as new benefit awards among previously non-receiving working-age individuals, it is necessary to condition on the absence of benefit receipt in the previous period when calculating the incidence rate for each period. Since the pilot program spans three years, only incidence rates within this time frame can effectively be compared without biasing results by conditioning on an

outcome. In contrast, a model built around the hazard as the parameter of interest lends itself naturally for this purpose.[4]

In follow-up analyses, I investigate possible labor market responses and how existing beneficiaries react to the medical review process. Unlike the inflow setting above, these measures can be analyzed in a linear model framework. In the analysis, I estimate a linear difference-in-differences specification with canton and year fixed effects and the interaction of the treated cantons with the pilot period.

4.2 Identification: Difference-in-differences for duration analysis

The standard assumptions for difference-in-differences estimation have to be restated for proportional hazard models. The exponentiated coefficient on the interaction between treatment time and region represents a ratio of hazard ratios

$$\exp\left(\beta_2\right) = \frac{h(t, D=1, P=1)/h(t, D=1, P=0)}{h(t, D=0, P=1)/h(t, D=0, P=0)} \ . \tag{2}$$

The distance condition has been dropped to ease notation. The effect of interest is the relative change in the hazard for the treated, i.e. a relative average treatment effect on the treated (rATT),

$$\text{rATT} = \frac{h^1(t, D=1, P=1)}{h^0(t, D=1, P=1)} \ , \tag{3}$$

where h^D denotes potential hazard rates. I assume SUTVA (Rubin 1977) holds, i.e. either of the two potential treatment states is observed. As disability insurance applicants are a small fraction of the population, it is credible that general equilibrium effects are absent. Identification then requires the two usual conditions in restated form

$$h^1(t, D=1, P=0) = h^0(t, D=1, P=0) \ , \qquad (\textit{no anticipation}, 4)$$

and

$$\frac{h^0(t, D=1, P=1)}{h^0(t, D=1, P=0)} = \frac{h^0(t, D=0, P=1)}{h^0(t, D=0, P=0)} \ . \qquad (\textit{common trend}, 5)$$

The main identifying assumption is that in the absence of mandatory medical review, incidence for individuals in both pilot and non-pilot (border) regions would have changed

[4] Another possibility would be to analyze prevalence, i.e. the effect of medical review on the stock of disability insurance beneficiaries. This is unappealing for two reasons. In a standard difference-in-differences framework, the common trend assumption implies that first differences between periods are equal for treated and control regions prior to treatment. Using prevalence as an outcome, this would imply *equal* incidence rates across regions, an assumption which seems unlikely to be fulfilled in the present context. Furthermore, medical review is more likely to affect the rates of newly awarded benefits immediately before effects on the stock of recipients eventually materialize.

proportionally. The common trend assumption is not invariant to the scaling of the dependent variable (e.g. Lechner 2010) and is modified accordingly. Instead of assuming a common trend between regions over time in differences, I am assuming a constant hazard ratio, i.e. a common relative change or a common absolute change in logs. In addition, I assume that anticipation effects are absent. Given these assumptions, the coefficient of the interaction identifies the hazard ratio of interest, the relative ATT.

4.3 Identification: Misclassification rates by eligibility type

The assumptions outlined in the previous section are sufficient to identify the reduced form effect of introducing mandatory external medical review on the DI inflow rate. This section explores additional conditions under which the reduced form effect can be interpreted as a mixture of effects on the false positive and false negative DI misclassification rates given individuals' latent eligiblity status.

It is the main duty of the insurance office to separate meritorious from non-meritorious claims ('tag' the eligible). Given the null hypothesis of 'no disability', two types of classification errors can occur in this situation: (1) Award errors (type-I) and (2) Rejection errors (type-II). If medical review is imperfect, benefits may be awarded to persons who are ineligible, and deserving applicants may be denied benefits.

Hence, medical review may not necessarily reduce insurance inflow unambiguously. Suppose that introducing mandatory medical review increases the probability to detect applicants' true type. This implies that medical review can reduce *both* type-I and type-II misclassification, resulting in opposing effects on the incidence rate. The net effect on inflow is undetermined and depends on the relative prevalence and likelihood of benefit receipt for eligible and ineligible applicants.

To illustrate the relative prevalence of type-I and type-II errors, decompose the effect in (3) by latent eligibility status $E = \Big\{0, 1\Big\}$,

$$\text{rATT} = \frac{\begin{array}{r} h^1(t, D=1, P=1|E=1) \cdot p^1(E=1, D=1, P=1) \\ + h^1(t, D=1, P=1|E=0) \cdot \Big[1 - p^1(E=1, D=1, P=1)\Big] \end{array}}{\begin{array}{r} h^0(t, D=1, P=1|E=1) \cdot p^0(E=1, D=1, P=1) \\ + h^0(t, D=1, P=1|E=0) \cdot \Big[1 - p^0(E=1, D=1, P=1)\Big] \end{array}} \ . \tag{6}$$

This underscores that the identified effect is a mixture of changes in the hazard for both eligible and ineligible types. Using this expression, it is possible to explore the conditions for a negative treatment effect (i.e. an inflow reduction, corresponding to a hazard ratio smaller than one) depending on the effect for each type separately. Assuming continuity in composition

$$p^0(E=1, D=1, P=1) = p^1(E=1, D=1, P=1) \ , \qquad (\textit{no self-screening}, 7)$$

and simplifying notation, propose

$$\text{rATT} = \frac{h^1(t|E=1)\cdot p(E=1) + h^1(t|E=0)\cdot\Big[1-p(E=1)\Big]}{h^0(t|E=1)\cdot p(E=1) + h^0(t|E=0)\cdot\Big[1-p(E=1)\Big]} \overset{!}{\leq} 1 \ . \tag{8}$$

Rearranging leads to

$$\Big[h^1(t|E=1) - h^0(t|E=1)\Big]\, p(E=1) \leq -\Big[h^1(t|E=0) - h^0(t|E=0)\Big]\Big[1-p(E=1)\Big], \tag{9}$$

i.e., the absolute value of the population-weighted treatment effect for the ineligible must exceed the population-weighted treatment effect for the eligible to observe an aggregate reduction in inflow. This implies the relative reduction in award errors (type-I, r.h.s.) must exceed the relative reduction in rejection errors (type-II, l.h.s) for the effect to be negative. This is consistent with the interpretation of the effect in (3) as a net effect.

Assuming the treatment does not decrease inflow of eligible types,

$$h^1(t|E=1) - h^0(t|E=1) \geq 0 \ , \quad (\textit{monotone treatment response for eligible types}, 10)$$

the left hand side of condition (9) is greater or equal zero. If medical review actually decreases the chances of the ineligible to get insurance benefits, the weighted decrease in the hazard for the ineligible must be less in absolute value than the weighted increase in the hazard for the eligible for the condition to be fulfilled. In this case, any observed inflow reduction can be interpreted as a net reduction in DI award errors. Alternatively, the condition in (9) is trivially fulfilled if (*monotone treatment response for eligible types*, 10) is violated and medical review actually has the perverse effect of worsening the chances of the truly eligible to get insurance, reducing their DI inflow hazard.

4.4 Potential threats

The two main threats to identification are a violation of the no anticipation condition and the common trend assumption. Prospective or ongoing reform changes may induce some individuals to change their behavior in anticipation of future loss or gain. The main confounding mechanisms are mobility (individuals move to treated regions to apply for DI), the timing of applications (early application in anticipation of medical review) and self-screening (individuals are discouraged from applying, see Parsons 1991).

The implementation and chronology of the reform alleviate these concerns. The first draft of the reform which included the institutional changes introducing medical review was proposed in parliament in February 2001, and underwent some revisions until being approved by popular vote in March 2003. The pilot project began already in January 2002. The early adopter scheme was scheduled immediately after the reform proposal

was publicised and began only ten months afterwards. Importantly, the pilot scheme was never publicly announced. Communication only occurred internally between the Federal Ministry of Social Insurances and the DI offices and was never publicised. A systematic news search on newspaper databases Factiva, LexisNexis, Pressreader and Swissdox does not reveal a single hit mentioning the early adopter program. Overall, the medical review changes implied by the reform proposal received little public attention and were only scheduled to be implemented in 2005.[5]

Considering the one-year earnings loss restriction required for DI eligibility, these time frames leave limited scope for the strategic timing of applications in both treated and control regions, even if public knowledge of the program were available. In the treated regions, the project started ten months after the first reform proposal, effectively leaving too little time for the strategic timing of applications in treated regions. Similarly, there is only a relatively short time period between the reforms definite approval in March 2003 and its nationwide implementation in January 2005 that allows for strategic behavior given the one-year restriction.

Anticipation effects can also manifest in increased mobility. Individuals considering to apply for disability benefits may anticipate the reform and move to regions where external medical review is not implemented, generating higher inflow in control regions and biased results. However, as described above, the medical review changes were not announced publicly at the time and the time frames are relatively narrow. In addition, the amount of people moving to another region who can be identified by tracking panel cases in the data is negligible. Between 1999 and 2011 about 3.1% of the people for whom time series information is available move to another canton, and less than 0.8% percent move from a non-treated to a treated region. About 0.5% of those sampled during the pilot period do so. Mobility in Switzerland is generally low compared to other countries.

Another potential concern is that results are confounded by changes in self-screening (Parsons 1991). Again, this behavior is unlikely since information about the pilot program did not transpire to the public. Unfortunately, the data on DI applications is very limited. It is only available after 2002, only at the aggregate level and unreliable due to a flawed electronic reporting system in some cantons during the early 2000s. However, looking at the limited aggregate data available, it appears that application rates evolve similarly across both groups of cantons over time (cf. Figure A4).

Regarding the common trend assumption, I first perform a balancing test to ensure that regions are comparable ex ante. Although balance in observables is not strictly required for identification, the common trend assumption is more credible if the comparison regions are

[5] Other reform measures scheduled to come into effect at a later time included the introduction of a three-quarter pension and the abolishment of additional pensions for spouses. These measures received the bulk of public attention. The changes were adopted nationwide and only became effective in late 2004. There were no further reforms to DI or other social insurances during the introduction period.

similar. This exercise reduces concerns about remaining regional heterogeneity (e.g. due to self-selection) that may induce common trend violations. Table A2 shows differences in selected covariates between treatment and control regions, separately for both the local and the unrestricted sample for a representative subset of data sampled prior to the pilot period. In the full sample there are significant differences with regard to age, the share of foreigners, education, marriage status and family size, characteristics which influence the propensity to receive DI. Among DI beneficiaries, musculoskeletal conditions are more prevalent in treated regions. In the local sample, balance improves considerably. Differences are small in magnitude and mostly insignificant. People in treated regions are on average more likely to be from a foreign country; there are about 2% more people with primary education in treated regions, and correspondingly less with secondary and university-level education. There is also a small difference in the unemployment rate of about 0.8 percentage points. These remaining differences in observables are small in economic terms and will not affect the estimates unless trends between treatment and control regions differ.

The typical diagnostic graph to inspect the validity of the common trend assumption in difference-in-differences designs are trend plots that show how treated and control units evolve prior to the treatment period. Figure Figure A2 shows trends in the aggregate recipient stock for all cantons based on statistics published by the Federal Ministry for Social Insurances. Prior to 2002, the number of DI recipients evolves similarly in treated and control regions. Afterwards, the numbers start to diverge and inflow is higher in control regions. The trend breaks after the reform is implemented on the federal level in 2005 and the DI rolls evolve similarly again.

While an indication of comparability between regions, strictly seen, this trend plot does not correspond to the dependent variable used in the estimations. Generating an equivalent plot in an age-based duration framework is hindered by the fact that the treatment occurs for every individual at a different time in life, i.e., the age at which they experience the reform being implemented. One possibility to investigate the assumption is to look at the log cumulative hazard by age as shown in Figure A3 (referred to as a 'log-log plot' in biostatistics). Since individuals are randomly sampled across regions and have the same age distributions, the log cumulative hazard estimates for both groups should be parallel. At the same time, the log-log plot is a common diagnostic to assess the validity of the proportional hazards assumption in the Cox model, with non-parallel or crossing lines seen as an indication that the proportionality assumption is violated. Figure A3 does not indicate a non-parallel hazard.

5 Results

5.1 Disability incidence and award errors

The main results are presented in Table 1, separately for the unrestricted and the local sample. The first column for each sample considers only spells which are censored or result in failure before the end of the pilot period in 2005, the remaining columns use all recorded spells and control for the post-treatment period in which the intervention was extended nationwide. The last column adds individual control variables, including gender, education, marital status, number of children and foreign citizenship. All specifications stratify the baseline hazard by five year birth cohort intervals to account for cohort specific differences in health environment. Survey weights are applied in the full sample such that estimates are representative of the Swiss population. Observations in the local sample are weighted for pairwise nearest-neighbor estimation. All tables report hazard ratios, i.e. exponentiated coefficients and corresponding standard errors.

All estimates of the effect of the reform are negative (i.e. hazard ratio less than one) and significant at conventional levels, indicating that third-party medical review significantly reduced insurance inflow. The estimate for the full sample implies a 14% reduction. The magnitude for the local sample is slightly higher and corresponds to a 23% lower inflow rate. Both estimates are stable in magnitude across specifications. The post coefficient estimates are negative as well, reflecting the fact that the reform was extended to the federal level after 2004 and funding increased even further. However, the post estimates for the local sample are imprecise as the failure density in the local sample is not dense enough in later years, when many observations are censored at the sampling date.

The preferred specification for the remainder of the paper is given in column (5), since adding covariates does not affect the results in a notable way. The remaining analysis focuses on the local sample. Results for the main sample are qualitatively similar.

External medical review is also likely to affect the classification of the severity of health impediments for new awards. I analyse whether medical review changes the relative incidence of partial and full benefit awards. Indeed, results in Table 2 show that incidence reductions occur only for full benefit awards (columns (2) and (3)) and those due to limitations classified as very serious (disability degree of 70% or larger, columns (4) and (5)). Estimates for partial pension awards and those classified as less serious are too imprecisely estimated to draw a clear conclusion, but may be unaffected. One possible explanation is that incidence reductions occur mainly for full benefit applicants. However, it is unlikely that only applicants claiming 100% work incapability constitute the affected marginal cases. A more likely scenario is that DI incidence reductions occur at all latent health levels. After introducing medical review, some individuals who would have received a full pension previously are now downgraded, resulting in a zero net effect for partial DI

Table 1: Disability incidence

	(a) Full sample			(b) Local sample (within 20 km)		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	1.322***	1.322***	1.236***	1.150***	1.151***	1.148***
	(0.041)	(0.041)	(0.039)	(0.061)	(0.061)	(0.061)
Pilot time	1.083	1.088	1.110	1.257*	1.267**	1.298**
	(0.089)	(0.089)	(0.090)	(0.148)	(0.148)	(0.152)
Treat x pilot	0.856**	0.856**	0.860*	0.770**	0.771**	0.766**
	(0.067)	(0.067)	(0.068)	(0.087)	(0.087)	(0.086)
Post time		0.690***	0.731***		0.867	0.918
		(0.068)	(0.072)		(0.151)	(0.160)
Treat x post		0.971	0.970		0.841	0.829
		(0.078)	(0.078)		(0.105)	(0.104)
Other controls	-	-	✓	-	-	✓
N municipalities	2,337	2,338	2,338	1,086	1,087	1,087
N individuals	249,750	259,323	259,323	128,536	133,549	133,549
N failures	7,877	9,204	9,204	3,985	4,693	4,693
N failures during pilot	1,713	1,713	1,713	885	885	885

Note: Cox Proportional Hazard estimates for individuals in treated and control regions based on SESAM individual-level survey and administrative data sampled during 1999–2011. Estimations separately for a complete representative sample of the Swiss population and only for individuals in the vicinity of the border between treated and non-treated regions. Baseline hazard for all regressions stratified by 5-year birth cohorts. Survey weights applied for the full sample. Observations in the local sample are weighted for nearest-neighbor pairwise differences. Results are reported in exponentiated form as hazard ratios. The hazard ratio for 'Treat x pilot' corresponds to the relative average treatment effect on the treated as defined in section 4. Standard errors clustered at the individual level in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

benefits.

5.2 Incidence of difficult-to-diagnose conditions

The main analysis indicates that DI awards declined substantially due external medical review, most likely due to a reduction in false positive benefit awards. If the effect is driven by more accurate health and functional capacity diagnoses, then incidence reductions are more likely to occur for diseases which are difficult to diagnose and verify for conventional physicians, the first DI gatekeeper. The reduction will be most pronounced for illnesses which are both difficult to diagnose and whose functional capacity implications are more likely to be misjudged.

Table 3 investigates this by differentiating between health impairments leading to benefit awards. The results confirm that reductions occur most frequently for difficult-to-diagnose conditions, while conditions which can typically be diagnosed unambiguously are not affected. Looking at column (3) and (4), the effect is pronounced for psychological diseases and illnesses related to nerve problems. Benefit awards due to mental health problems are reduced by 30%. Nerve-related handicaps are reduced by over 60%, but incidence in this group is generally very low. Column (5) looks at the incidence of musculoskeletal (MSC) diseases. This category also includes a variety of conditions which are difficult to verify (e.g. whiplash injuries, back pain). The hazard ratio suggest a substantial reduction in incidence as well. The specification in column (6) looks at disability benefit awards due to

Table 2: Disability classification

	All	Partial	Full	DD < 70	DD ≥ 70
	(1)	(2)	(3)	(4)	(5)
Treated region	1.151***	1.071	1.169**	1.043	1.219***
	(0.061)	(0.115)	(0.073)	(0.104)	(0.081)
Pilot period	1.267**	1.541**	1.118	1.509**	1.166
	(0.148)	(0.305)	(0.165)	(0.290)	(0.183)
Treat x pilot	0.771**	0.925	0.710**	0.981	0.646***
	(0.087)	(0.178)	(0.102)	(0.181)	(0.099)
Post time	0.867	1.446	0.584**	1.423	0.633*
	(0.151)	(0.400)	(0.133)	(0.382)	(0.151)
Treat x post	0.841	0.722	1.003	0.717*	0.992
	(0.105)	(0.147)	(0.164)	(0.141)	(0.169)
N municipalities	1,087	1,087	1,087	1,087	1,087
N individuals	133,549	133,549	133,549	133,549	133,549
N failures	4,693	1,352	3,283	1,481	2,879
N failures during pilot	885	338	538	357	474

Note: Cox Proportional Hazard estimates for individuals in treated and control regions based on SESAM individual-level survey and administrative data sampled during 1999–2011. Sample is based on individuals living within 20 km of the border between treated and non-treated regions. Columns distinguish between partial/full DI benefit awards and awards due to less serious/serious health limitations (disability degree smaller/greater than 70). Baseline hazard for all regressions stratified by 5-year birth cohorts. Observations are weighted for nearest-neighbor pairwise differences. Results are reported in exponentiated form as hazard ratios. The hazard ratio for 'Treat x pilot' corresponds to the relative average treatment effect on the treated as defined in section 4. Standard errors clustered at the individual level in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

Table 3: Disability types

	All	Illness	Illness: Psych.	Illness: Nerve	Illness: MSC	Accident	Congenital/ Other
	(1)	(2)	(3)	(4)	(5)	(7)	(8)
Treatment region	1.151***	1.229***	1.185*	1.100	1.245**	0.843	1.293**
	(0.061)	(0.072)	(0.106)	(0.216)	(0.136)	(0.148)	(0.162)
Pilot period	1.267**	1.384**	1.450*	2.373*	1.412	0.900	0.795
	(0.148)	(0.178)	(0.282)	(1.185)	(0.330)	(0.362)	(0.201)
Treat x pilot	0.771**	0.683**›	0.699*	0.377**	0.633**	1.729	1.150
	(0.087)	(0.084)	(0.129)	(0.167)	(0.145)	(0.656)	(0.290)
Post time	0.867	0.974	0.667	1.737	1.285	0.175***	1.220
	(0.151)	(0.183)	(0.188)	(1.211)	(0.460)	(0.102)	(0.441)
Treat x post	0.841	0.733**	0.897	0.607	0.596**	6.436***	0.748
	(0.105)	(0.097)	(0.176)	(0.272)	(0.156)	(2.942)	(0.197)
N municipalities	1,087	1,087	1,087	1,087	1,087	1,087	1,087
N individuals	133,549	133,549	133,549	133,549	133,549	133,549	133,549
N failures	4,693	3,827	1,685	339	1,090	409	835
N failures during pilot	885	753	352	61	210	59	149

Note: Cox Proportional Hazard estimates for individuals in treated and control regions based on SESAM individual-level survey and administrative data sampled during 1999–2011. Sample is based on individuals living within 20 km of the border between treated and non-treated regions. Columns distinguish between DI awards due to different health impairments. Baseline hazard for all regressions stratified by 5-year birth cohorts. Observations are weighted for nearest-neighbor pairwise differences. Results are reported in exponentiated form as hazard ratios. The hazard ratio for 'Treat x pilot' corresponds to the relative average treatment effect on the treated as defined in section 4. Standard errors clustered at the municipality level in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

handicaps incurred in accidents; the last column considers disabilities due to congenital defects and other diseases. These conditions are unlikely to be subject to award errors, as there is rarely any ambiguity and they are typically well-documented. Indeed, there is no effect on conditions which are unaffected by intensified medical review measures.

5.3 Labor market responses to medical review

This section investigates the labor market reaction in response to external medical review. In case reductions in DI incidence are driven by rejections of individuals capable of returning to the labor market, medical review should also have a positive effect on labor market participation. Conversely, if the reduction is largely driven by rejections of individuals incapable of working, medical review should not have an effect on employment, but possibly on the inflow into other social security programs (e.g. Inderbitzin et al. 2013). Table 4 uses the pooled cross-sectional administrative SESAM data to estimate a differences-in-differences specification using a linear model.

The results in Table 4 show that the share of individuals in registered employment increases, both in the full and the local sample. Similarly, the share of individuals with positive (non-benefit) earnings increases as well ($p = 0.137$ in the local sample). In the full sample the share of individuals registered with the employment office as job seekers also decreases. One explanation for these results is that DI applications are partly made by people capable of gainful employment and driven by moral hazard. An alternative explanation is that they are partly due to income effects (e.g. Autor and Duggan 2007, Gelber et al. 2017).

In columns (4) and (5), I consider other pathways from unemployment and reasons for not being registered with the employment office anymore. I find no effect on dismissal from the employment office (and the associated return-to-work measures) due to being considered 'unemployable' by the caseworker. Similarly, I find no effect on the receipt of social assisstance, the minimum social security provision. If rejected DI applicants were incapable of working, we would expect to see an increase in these measures. However, the results do not provide evidence for this channel.

5.4 Disability degree and benefit revisions in the recipient stock

Although the primary task of the medical staff is to screen applicants, they also aid with reviews of recipients' disability degree classification. While scheduled by law to occur regularly, revisions seldom resulted in actual disability degree or benefit cuts and typically involved going over beneficiaries files without personal contact. Revisions also commonly take place if applicants have submitted new medical information, typically documenting deteriorating health, and often result in pension increases. With the new regime in place, files that are scheduled for review are now also passed to the physicians in charge of the

Table 4: Labor market responses

(a) Full sample

	Work registered	Positive labor income	Employment office registration	Employment office dismissal	Social asssistance
Treat x pilot	0.009***	0.008***	−0.007***	−0.002	0.000
	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)
Individual covariates	✓	✓	✓	✓	✓
Canton FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
N	556,540	557,270	411,461	411,461	411,461

(b) Local sample (within 20 km)

	Work registered	Positive labor income	Employment office registration	Employment office dismissal	Social asssistance
Treat x pilot	0.007*	0.006	−0.003	0.000	0.001
	(0.004)	(0.004)	(0.003)	(0.002)	(0.002)
Individual covariates	✓	✓	✓	✓	✓
Canton FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
N	282,858	283,111	208,340	208,340	208,340

Note: Linear model estimates for individuals in treated and control regions based on SESAM individual-level survey and administrative data sampled during 1999–2011. Estimations separately for a complete representative sample of the Swiss population (panel a) and only for individuals in the vicinity of the border between treated and non-treated regions (panel b). All models include cantonal and year specific effects and control for gender, age and native status. Standard errors given in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

medical assessments.

To assess whether stock reclassifications occur, I estimate a linear difference-in-difference model using data for the stock of all DI beneficiaries in Switzerland in 2001. I condition on benefit receipt prior to treatment and track the changes to the disability degree and the effective benefit payments of existing beneficiaries over time. Results are given in Table 5. The sample is again stratified by disease groups. The outcome in panel (a) is the individual disability degree, panel (b) looks at the benefit amount. On average, recipients are classified less disabled by 0.35 percentage points and lose about 17 CHF in monthly benefits. The effect magnitudes are small since reclassification remains a rare event. Summary statistics indicate that only 9.3% of individuals of the 2001 stock are reclassified during the three years of the pilot period. Complete denial of benefits after a revision occurs only in exceptional cases.[6] Upward revisions are far more common, downward changes only account for 2.3 percentage points. Still, introducing mandatory medical review appear to lead to revisions of the disability status of beneficiaries whose documentation is deemed insufficient, suspicious or whose health has improved. Both the disability classification and payouts are again only adjusted for those beneficiaries with illnesses which are more difficult to screen. Again, cuts are most pronounced for

[6] Complete benefit denial is legally difficult, unless fraud or malingering are proven beyond reasonable doubt. These cases also require high-up front investment from DI offices, e.g. investigators for surveillance, and are initiated only in extreme cases.

Table 5: Stock reclassification and pension cuts

(a) Disability degree

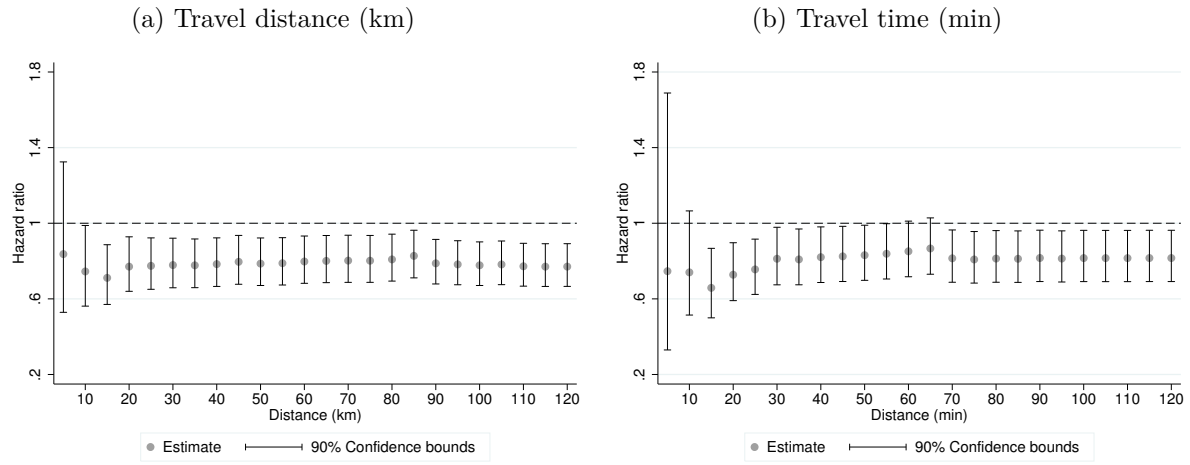
	All	All illnesses	Psychological	MSC	Accident	Congenital
Treated region	3.04***	3.41***	3.66***	2.78***	1.67***	1.19***
	(0.08)	(0.09)	(0.13)	(0.18)	(0.26)	(0.18)
Pilot	0.49***	0.60***	0.60***	0.39***	0.46***	0.30**
	(0.06)	(0.07)	(0.09)	(0.13)	(0.17)	(0.13)
Treat x pilot	−0.35***	−0.42***	−0.58***	−0.39*	−0.24	−0.10
	(0.09)	(0.11)	(0.15)	(0.20)	(0.30)	(0.21)
Post	1.63***	1.80***	1.44***	0.87***	0.89***	1.39***
	(0.05)	(0.06)	(0.09)	(0.12)	(0.16)	(0.12)
Treat x post	−0.52***	−0.61***	−0.83***	−0.47**	−0.39	−0.28
	(0.09)	(0.10)	(0.14)	(0.19)	(0.28)	(0.19)
Constant	78.54***	77.98***	82.75***	72.31***	74.53***	86.07***
	(0.05)	(0.06)	(0.08)	(0.11)	(0.15)	(0.11)

(b) Pension amount

	All	All illnesses	Psychological	MSC	Accident	Congenital
Treated region	124.68***	141.87***	101.22***	133.26***	88.41***	8.36***
	(2.20)	(2.61)	(3.67)	(4.94)	(7.25)	(3.14)
Pilot	37.04***	39.92***	33.12***	39.02***	33.29***	28.43***
	(1.55)	(1.85)	(2.62)	(3.56)	(4.83)	(2.25)
Treat x pilot	−17.25***	−21.77***	−17.79***	−21.90***	−8.10	−0.45
	(2.52)	(2.98)	(4.17)	(5.66)	(8.33)	(3.64)
Post	143.12***	148.37***	126.96***	139.82***	122.50***	126.26***
	(1.46)	(1.75)	(2.46)	(3.38)	(4.54)	(2.10)
Treat x post	−41.25***	−48.74***	−33.50***	−50.51***	−19.17**	−1.61
	(2.38)	(2.82)	(3.92)	(5.38)	(7.84)	(3.39)
Constant	1232.08***	1221.01***	1311.78***	1134.61***	1199.86***	1343.85***
	(1.35)	(1.61)	(2.30)	(3.10)	(4.20)	(1.95)
N	2,489,323	1,884,876	887,604	537,191	282,224	274,918

Note: Estimates from a linear model. Outcomes are the disability degree in percent (panel a) and the effective benefit amount paid to recipients in panel (b). The reference group are individuals in the non-treated regions in 2001. Based on administrative panel data provided by the Swiss Federal Ministry of Social insurances which tracks the complete stock of Swiss DI benefit recipients in 2001 until 2011. Standard errors in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

Figure 3: Distance windows



Note: Treatment effect estimates and 90% confidence bounds from the main specification for different distance windows measured using actual travel distance and travel time.

those who receive DI due to mental health problems or musculoskeletal conditions, while beneficiaries with congenital diseases or handicaps incurred in accidents are unaffected. Unlike previously, nerve-related diseases are not declared in this data.

5.5 Robustness checks

To assess the validity of the main identifying assumption, I test the effect of a placebo reform prior to the treatment period and assume a pseudo-treatment to be effective during 1999–2001. Results are shown in Table A3. Hazard ratio estimates across all specifications are close to one, precisely estimated and insignificant at conventional levels, supporting the validity of the identification strategy.

Another potential concern is that the results are sensitive to the choice of distance window. Figure 3 addresses this issue by plotting treatment effect estimates across a large set of bandwidths, using both actual travel distance and travel time as distance measures. The coefficient of interest remains stable in size and significant across a large set of distances. The estimates consistently suggest about a 20% reduction in incidence in the treatment group during the pilot program. More detailed estimates over selected distances are provided in Table A4 in the appendix.

Another potential issue pertains to bias incurred by selective sampling due to DI outflow. Previous benefit receipt is not observed in the data. Only persons who are still on the DI rolls and those who are not are observed in each sampling year. If the reform affected DI outflow as well, sampling may be biased, as those who were barred from receiving insurance due to treatment are not observed in later years. This may result in a selected sample with artificially lower inflow in treatment regions. An actual outflow effect

would be mistaken for an inflow effect due to unobserved dropout. I can test for such outflow effects using the stock data. A duration model similar to the main specification is estimated for those who are beneficiaries prior to treatment in 2001. Exit from the DI rolls is considered failure, individuals are censored at the sampling limit in 2011 or when they exit at the relevant pension age. Variable measurements are less clean-cut in this case. Exit due to work or expulsion cannot be separated. However, there is no explicit reason why trends in work take-up by insurees (a similarly rare event) should differ between regions. Results are given in Table A5, separately for all individuals and those below age 50 in 2001, an age requirement which prohibits early retirement within the analysis horizon and selects a younger and possibly healthier group more likely to exit. All estimates are consistently indistinguishable from zero and precisely estimated.

A final concern is that external medical review might simply prolong the decision process and delay benefit approval. Note that the DI entry measurement effectively precludes this possibility. Entry is observed for those who effectively enter the insurance system at the time when they register with the insurance office and file their application, not when they are finally granted benefits. This is due to the fact that benefits are paid retroactively from the filing date after approval.

As illustrated, the main results are robust to a series of checks and very stable in magnitude. The results for both samples are also robust to model changes. Estimations using a piecewise constant exponential model or a flexible parametric model using splines instead of the stratified Cox model return similar results. Similarly, the results are not dependent on the application of weights, stratification or the stratification level.[7]

Equally persistent through variations is the approximately 7% difference in magnitude between estimates for the global and the local sample. It is illustrative to trace where the difference in results may arise from. To shed light on the differences between the local and the full sample, I estimate a Probit model for the probability to be included in the local sample, separately for treated and control regions. Table A6 presents the results. The local treated sample closely resembles the rest of the treated region. However, the local control sample differs from the rest of the control population. It has a higher share of foreigners (about 10% at the mean), more women and more well-educated individuals – all factors which contribute to a lower overall incidence and are likely to drive the difference in results.

6 Discussion

The main results indicate that external medical review can reduce insurance inflow substantially, especially for difficult-to-diagnose conditions. The most conservative estimate

[7] Not reported. All results available on request.

suggests that in absence of the reform, an additional 16% of actual inflow would have been approved for benefits.

Results from the local approach (sample restricted to commuting distance around borders) have the same sign and are comparable in magnitude to the global approach (using the full sample). The distance variations in Figure 3 consistently suggest a reduction in the hazard of about 20%. Considering the sizeable effect of medical review on the DI hazard, it is illustrative to assess how large the absolute effects induced by introduction of the medical reviews are. Looking at the main specification, without treatment, the baseline DI hazard in the treated regions is about 0.38%, i.e. on average 3.8 persons per thousand enter DI. The medical review process reduces this by about 23% to 0.29%, implying that approximately one person less per thousand enters DI due to a second medical assessment.

The results also cast doubt on relying on the evaluations of GPs for the DI decision, an established practice in many OECD countries. Considering that inflow reductions are restricted to difficult-to-diagnose conditions and the results indicate that work take-up increases when medical review is done by clinical specialists, GPs may not be well suited to serve as the main gatekeeper to DI. This result corroborates medical studies which posit that specialists may be better suited to judge social insurance eligibility than personal physicians (e.g. Novack et al. 1989, Zinn and Furutani 1996, Freeman et al. 1999, Wynia et al. 2000, Everett et al. 2011). In addition, treating physicians have often voiced discomfort with being both care-takers of patients and gatekeepers to public insurance systems. In surveys, physicians are overwhelmingly in favor of designating independent third-party physicians to determine disability status to prevent damaging physician-patient relations (e.g. Zinn and Furutani 1996).

Considering the substantial present-discounted value of DI pensions, it is interesting to examine whether external medical review is a cost-effective policy. Simple back-of-the-envelope calculations indicate that outlays for hiring physicians are more than offset by reductions in the beneficiary payload if inflow reductions are permanent. Using the average benefit amount and remaining spell duration until retirement for calculation, yearly savings are likely to be above 1.5 billion Swiss Francs (approximately 1.5 billion US$). Even if all rejected applicants never reenter the labor market and immediately receive social assistance, estimated savings are upwards of 650 million Swiss Francs yearly. These calculations disregard the fact that benefit decisions are tied to additional occupational benefits and private pension schemes, which are substantially more generous than the main state pension and would result in further savings. Nevertheless, the yearly savings far exceed potential outlays for the medical personnel that was hired. Introducing external medical review is a highly cost-effective tool to reduce insurance inflow.

The reduction in the DI hazard has further implications. Given the auxiliary assumptions outlined in section 4.3, the result indicate that that DI award errors occur more frequently than rejection errors. However, it is theoretically possible that the inflow reduc-

tion is driven by screening institutions inducing an even larger number of false negative errors (as illustrated in section 4.3). This would imply that insurance offices now reject *more* applicants that are actually deserving than previously. One argument is that if lower DI incidence is politically desired, individual physicians might be tempted to be generally more critical when reviewing new applications due to a fear of being laid off. However, the additional staff at the screening offices were hired on permanent employment contracts and could not have been easily laid off in any case, irrespective of the development of the insurance rolls. It was generally recognized that the insurance offices' structure, last revised 1973, needed to be overhauled and that they were notoriously understaffed with physicians. The physicians responsible for medical review had the explicit mandate to improve the accuracy of medical diagnoses of functional limitations and received specialized training specifically tailored for this purpose.[8] In addition, the institutional structure of the DI offices and the underlying regulation remains unchanged. The final decision still lies with the DI caseworker.

Moreover, the scope of the federal government to influence local public entities is limited due to the decentralized nature of the Swiss political system. The implementation of the medical review process and the DI decision are made on the local level, even though DI pensions are paid out of federal funds. The role of the federal government was limited to providing funding for the program. Even if political pressure were exerted to increase stringency, the trend would have to differ between regions to confound the main result. Moreover, if differential changes in stringency where to occur, these would most likely manifest in higher rates of legal claims regarding DI entitlements. Comprehensive data on legal claims is sparse due to limited reporting coverage. Where available, I collected data on the amount of legal claims for each canton from yearly reports of the cantonal courts. Figure A5 illustrates that both the number of total and rejected lawsuits in treated and control regions evolve very similarly over time. This suggests that it is unlikely that the results are driven by differential changes in stringency.

7 Conclusion

This paper provides a comprehensive evaluation of the introduction of mandatory medical review for DI applications in a setting in which treating physician testimony was decisive. The results indicate that mandatory external medical review reduces DI benefit awards substantially, suggesting that award errors are likely to occur. Effect magnitudes are substantial: Specialist medical assessments reduce DI uptake by between 14% and 23%.

[8] The physicians accompanying the implementation were in fact acutely aware of the possibility that more intense medical review could increase DI incidence. The leading physician in one office stated that in their experience, rejection errors do occur and are sometimes encountered during revisions, but are much less frequent in relation to the amount of award errors uncovered ex post.

Reducations are closely tied to difficult-to-diagnose conditions, suggesting a more accurate assessment of complex or multidisciplinary diseases. This is corroborated by the fact that disability status and benefit revisions in the stock of recipients occur only for individuals with the same types of conditions and the fact that medical review also increases labor market participation. Taken together, the results cast doubt on the established practice to assign 'controlling weight' to the treating physician's opinion in DI insurance decisions.

It is important to note that screening in this setting does not come at the cost of increased program complexity (e.g. as modeled by Kleven and Kopczuk 2011). The additional administrative hassle is low, and there are few visible additional up-front costs borne by the applicant. As such, external medical review is unlikely to discourage take-up strongly in the long-term. Expert medical review boards appear to be an effective tool in curbing inflow rates into disability insurance.

This is strengthened by the fact that the introduction of the screening offices is highly cost-effective. Since autonomous medical gatekeeper institutions appear to be effective in the Swiss setting, they might provide a viable policy option for other countries burdened by high disability insurance costs. However, it is important to bear in mind that prior to the reform, medical review was conducted almost exclusively by treating physicians and public health officers could not decree medical checks. Many countries already have more elaborate review and application mechanisms in place. Both classification errors and the policy impact may well be lower, depending on the initial level of screening intensity.

The mechanisms behind the results in this paper merit further investigation. One possible channel behind the incidence reductions are inaccurate diagnoses by GPs, the first gatekeeper to the DI system. However, whether application behavior suffers from moral hazard (and if so, how much) remains ultimately unclear. Applicants could be largely myopic or actively engage in malingering. In addition, the overall reduction in inflow only provides a tentative suggestion that award errors exceed rejection errors in award decisions. Separating type-I and type-II classification errors more cleanly and examining the mechanisms through which they occur remains a promising pursuit for further research.

References

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(1), 8–19.

Autor, D. and Duggan, M. (2003). The rise in the disability rolls and the decline in unemployment. *The Quarterly Journal of Economics* 118(1), 157–205.

Autor, D. H. and Duggan, M. G. (2007). Distinguishing income from substitution effects in disability insurance. *American Economic Review* 97(2), 119–124.

Benitez-Silva, H., Buchinsky, M., Man Chan, H., Cheidvasser, S. and Rust, J. (2004). How large is the bias in self-reported disability? *Journal of Applied Econometrics* 19(6), 649–670.

Bogardus, S., , Geist, D. and Bradley, E. (2004). Physicians' interactions with third-party payers: Is deception necessary? *Archives of Internal Medicine* 164(17), 1841–1844.

Bolduc, D., Fortin, B., Labrecque, F. and Lanoie, P. (2002). Workers' compensation, moral hazard and the composition of workplace injuries. *The Journal of Human Resources* 37(3), 623–652.

BSV (2012). *Statistiken zur sozialen Sicherheit – IV-Statistik 2011.* Bundesamt für Sozialversicherungen.

Butler, J. S., Burkhauser, R. V., Mitchell, J. M. and Pincus, T. P. (1987). Measurement error in self-reported health variables. *The Review of Economics and Statistics* 69(4), 644–650.

Bütler, M., Deuchert, E., Lechner, M., Staubli, S. and Thiemann, P. (2015). Financial work incentives for disability benefit recipients: Lessons from a randomised field experiment. *IZA Journal of Labor Policy* 4(1), 1–18.

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* 28(3), 419–441.

Campolieti, M. (2006). Disability insurance adjudication criteria and the incidence of hard-to-diagnose medical conditions. *Contributions to Economic Analysis & Policy* 5(1), Article 15.

Campolieti, M. and Riddell, C. (2012). Disability policy and the labor market: Evidence from a natural experiment in Canada, 1998–2006. *Journal of Public Economics* 96(3–4), 306–316.

Carey, T. S. and Hadler, N. M. (1986). The role of the primary physician in disability determination for social security insurance and workers' compensation. *Annals of Internal Medicine* 104(5), 706–710.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2), 187–220.

Duddleston, D. N., Blackston, J. W., Bouldin, M. J. and Brown, C. A. (2002). Disability examinations: A look at the social security disability income system. *The American Journal of the Medical Sciences* 324(4), 220–226.

Englund, L., Tibblin, G. and Svärdsudd, K. (2000). Variations in sick-listing practice among male and female physicians of different specialities based on case vignettes. *Scandinavian Journal of Primary Health Care* 18(1), 48–52.

Eugster, B. and Parchet, R. (2011). Culture and taxes: Towards identifying tax competition. *Cahiers de Recherches Economiques du Département d'Econométrie et d'Economie politique (DEEP) 11.05*, Université de Lausanne, Faculté des HEC, DEEP.

Everett, J. P., Walters, C. A., Stottlemyer, D. L., Knight, C. A., Oppenberg, A. A. and Orr, R. D. (2011). To lie or not to lie: Resident physician attitudes about the use of deception in clinical practice. *Journal of Medical Ethics* 37(6), 333–338.

Freeman, V., Rathore, S., Weinfurt, K., Schulman, K. and Sulmasy, D. (1999). Lying for patients: Physician deception of third-party payers. *Archives of Internal Medicine* 159(19), 2263–2270.

Frölich, M. and Lechner, M. (2010). Exploiting regional treatment intensity for the evaluation of labor market policies. *Journal of the American Statistical Association* 105(491), 1014–1029.

Frueh, B. C., Elhai, J. D., Gold, P. B., Monnier, J., Magruder, K. M., Keane, T. M. and Arana, G. W. (2003). Disability compensation seeking among veterans evaluated for posttraumatic stress disorder. *Psychiatric Services* 54(1), 84–91.

Gelber, A., Moore, T. J. and Strand, A. (2017). The effect of disability insurance payments on beneficiaries' earnings. *American Economic Journal: Economic Policy* 9(3), 229–61.

Gruber, J. (2000). Disability insurance benefits and labor supply. *Journal of Political Economy* 108(6), 1162–1183.

Inderbitzin, L., Staubli, S. and Zweimüller, J. (2013). Extended unemployment benefits and early retirement: Program complementarity and program substitution. *IZA Discussion Papers 7330*, Institute for the Study of Labor (IZA).

de Jong, P., Lindeboom, M. and van der Klaauw, B. (2011). Screening disability insurance applications. *Journal of the European Economic Association* 9(1), 106–129.

Kankaanpää, A. T., Franck, J. K. and Tuominen, R. J. (2012). Variations in primary care physicians' sick leave prescribing practices. *The European Journal of Public Health* 22(1), 92–96.

Keele, L. and Titiunik, R. (2016). Natural experiments based on geography. *Political Science Research and Methods* 4(1), 65–95.

Kleven, H. J. and Kopczuk, W. (2011). Transfer program complexity and the take-up of social benefits. *American Economic Journal: Economic Policy* 3(1), 54–90.

Kom, E. L., Graubard, B. I. and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. *American Journal of Epidemiology* 145(1), 72–80.

Kreider, B. (1999). Latent work disability and reporting bias. *Journal of Human Resources* 34(4), 734–769.

Kreider, B. and Pepper, J. (2007). Disability and employment: Reevaluating the evidence in light of reporting errors. *Journal of the American Statistical Association* 102(478), 432–441.

Kreider, B. and Pepper, J. (2008). Inferring disability status from corrupt data. *Journal of Applied Econometrics* 23(3), 329–349.

Lechner, M. (2010). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics* 4(3), 165–224.

Maestas, N., Mullen, K. J. and Strand, A. (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review* 103(5), 1797–1829.

Mitra, S. (2009). Disability screening and labor supply: Evidence from South Africa. *American Economic Review* 99(2), 512–516.

Nagi, S. Z. (1969). *Disability and rehabilitation: Legal, clinical, and self-concepts and measurement.* Columbus, Ohio State University Press.

Novack, D., Detering, B., Arnold, R., Forrow, L., Ladinsky, M. and Pezzullo, J. (1989). Physicians' attitudes toward using deception to resolve difficult ethical problems. *JAMA* 261(20), 2980–2985.

OECD (2009). Sickness, disability and work: Keeping on track in the economic downturn. *Working paper*, High-Level Forum, Stockholm.

OECD (2010). *Sickness, Disability and Work: Breaking the Barriers.* Paris, OECD Publishing.

Parsons, D. O. (1991). Self-screening in targeted public transfer programs. *Journal of Political Economy* 99(4), 859–876.

Parsons, D. O. (1996). Imperfect 'tagging' in social insurance programs. *Journal of Public Economics* 62(1–2), 183–207.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral Statistics* 2(1), 1–26.

Sheshinski, E. (1978). A model of social security and retirement decisions. *Journal of Public Economics* 10(3), 337–360.

Smith, R. T. and Lilienfeld, A. M. (1971). *The Social Security Disability program: An evaluation study.* 39, US Social Security Administration, Office of Research and Statistics.

Staten, M. E. and Umbeck, J. (1982). Information costs and incentives to shirk: Disability compensation of air traffic controllers. *The American Economic Review* 72(5), 1023–1037.

Staubli, S. (2011). The impact of stricter criteria for disability insurance on labor force participation. *Journal of Public Economics* 95(9-10), 1223–1235.

Thiébaut, A. C. M. and Bénichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: A simulation study. *Statistics in Medicine* 23(24), 3803–3820.

Wapf, B. and Peters, M. (2007). Evaluation der regionalen ärztlichen Dienste. *Beiträge zur Sozialen Sicherheit,* Bericht im Rahmen des mehrjährigen Forschungsprogramms zu Invalidität und Behinderung, Forschungsbericht Nr. 13/07.

Wynia, M., Cummins, D., VanGeest, J. and Wilson, I. (2000). Physician manipulation of reimbursement rules for patients: Between a rock and a hard place. *JAMA* 283(14), 1858–1865.

Zinn, W. and Furutani, N. (1996). Physician perspectives on the ethical aspects of disability determination. *Journal of General Internal Medicine* 11(9), 525–532.

A1 Appendix: Tables and Figures

Table A1: Descriptive statistics

	Mean	SD	Min	Max	N
(a) Full sample					
All individuals					
Age	50.316	18.033	18.0	104.0	259,323
Female	0.539	0.498	0.0	1.0	259,323
Married	0.552	0.497	0.0	1.0	259,323
Foreign	0.322	0.467	0.0	1.0	259,323
Nr. of children	0.582	0.973	0.0	7.0	259,323
Education: Primary	0.234	0.423	0.0	1.0	259,323
Education: Secondary	0.510	0.500	0.0	1.0	259,323
Education: Tertiary	0.255	0.436	0.0	1.0	259,323
Gross annual earnings	41.450	107.251	0.0	42,317.4	259,323
Travel distance (km)	34.297	31.825	0.2	194.1	259,323
Travel time (min)	31.411	23.167	0.6	169.5	259,323
Unemployed	0.027	0.163	0.0	1.0	259,323
Receives DI	0.035	0.185	0.0	1.0	259,323
Region					
Léman	0.191	0.393	0.0	1.0	259,323
Mittelland	0.194	0.396	0.0	1.0	259,323
Nordwestschweiz	0.136	0.343	0.0	1.0	259,323
Zürich	0.166	0.372	0.0	1.0	259,323
Ostschweiz	0.122	0.328	0.0	1.0	259,323
Zentralschweiz	0.107	0.310	0.0	1.0	259,323
Tessin	0.083	0.275	0.0	1.0	259,323
DI recipients					
Years in DI	9.415	6.847	0.0	48.0	9,204
Disability: Psych. problems	0.341	0.474	0.0	1.0	9,204
Disability: Nerve	0.072	0.259	0.0	1.0	9,204
Disability: Muscoloskeletal cond.	0.235	0.424	0.0	1.0	9,204
Disability: Accident	0.092	0.289	0.0	1.0	9,204
Disability: Congenital disease/other	0.185	0.388	0.0	1.0	9,204
(b) Local sample (within 20 km)					
	Mean	SD	Min	Max	N
All individuals					
Age	49.950	18.019	18.0	104.0	133,549
Female	0.538	0.499	0.0	1.0	133,549
Married	0.546	0.498	0.0	1.0	133,549
Foreign	0.329	0.470	0.0	1.0	133,549
Nr. of children	0.580	0.972	0.0	7.0	133,549
Education: Primary	0.226	0.418	0.0	1.0	133,549
Education: Secondary	0.510	0.500	0.0	1.0	133,549
Education: Tertiary	0.265	0.441	0.0	1.0	133,549
Gross annual earnings	43.252	134.295	0.0	42,317.4	133,549
Travel distance (km)	11.871	4.753	0.2	20.0	133,549
Travel time (min)	14.981	5.170	0.6	30.1	133,549
Unemployed	0.027	0.163	0.0	1.0	133,549
Receives DI	0.035	0.184	0.0	1.0	133,549
Region					
Léman	0.119	0.324	0.0	1.0	133,549
Mittelland	0.156	0.363	0.0	1.0	133,549
Nordwestschweiz	0.260	0.439	0.0	1.0	133,549
Zürich	0.256	0.436	0.0	1.0	133,549
Ostschweiz	0.068	0.252	0.0	1.0	133,549
Zentralschweiz	0.140	0.347	0.0	1.0	133,549
Tessin	0.000	0.003	0.0	1.0	133,549
DI recipients					
Years in DI	9.294	6.779	0.0	47.0	4,693
Disability: Psych. problems	0.359	0.480	0.0	1.0	4,693
Disability: Nerve	0.072	0.259	0.0	1.0	4,693
Disability: Muscoloskeletal cond.	0.232	0.422	0.0	1.0	4,693
Disability: Accident	0.087	0.282	0.0	1.0	4,693
Disability: Congenital disease/other	0.178	0.382	0.0	1.0	4,693

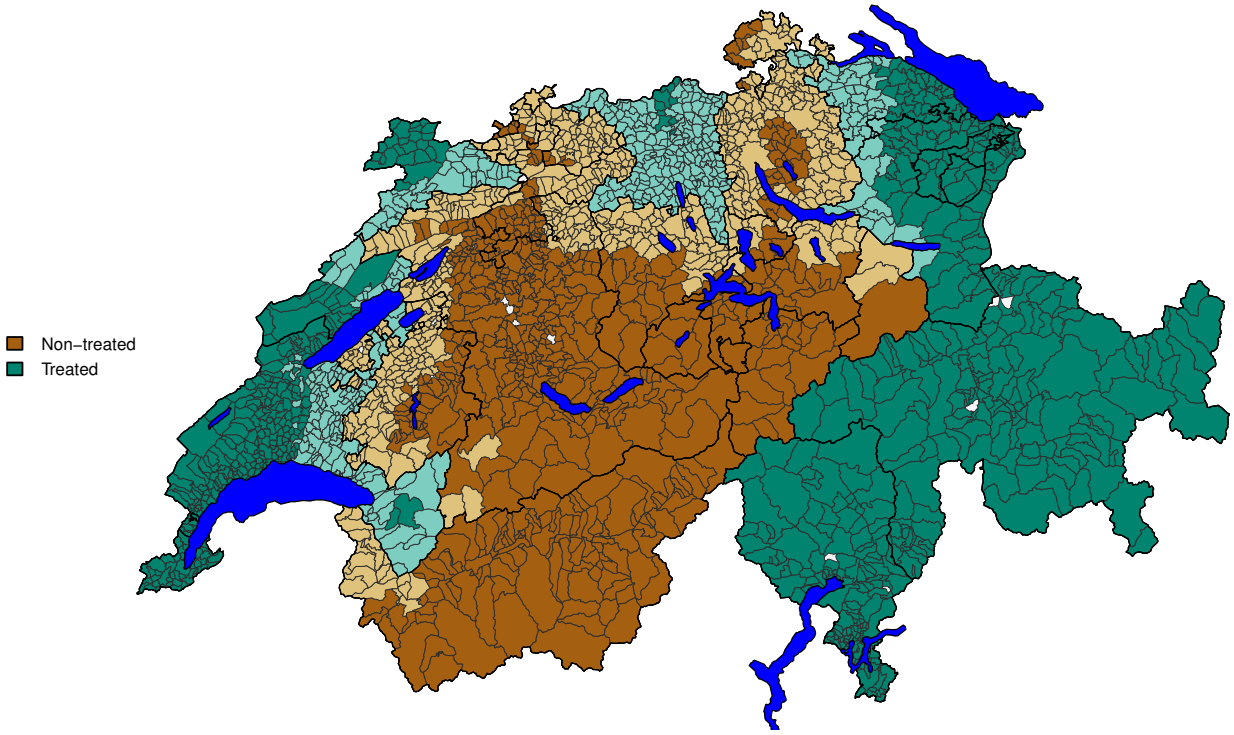
Note: Descriptive statistics for the unrestricted and the local estimation sample. Based on the 1999–2011 SESAM data.

Table A2: Pre-treatment covariate balance

	(a) Full sample				(b) Local sample (within 20 km)			
	Total	Treated	Control	Difference	Total	Treated	Control	Difference
All individuals								
Age	48.34	47.74	48.66	−0.926***	48.55	48.53	48.68	−0.153
	(18.28)	(18.83)	(17.95)	(0.309)	(18.56)	(10.61)	(40.06)	(0.605)
Female	0.54	0.55	0.54	0.009	0.55	0.55	0.54	0.009
	(0.50)	(0.52)	(0.49)	(0.009)	(0.50)	(0.29)	(1.08)	(0.016)
Married	0.52	0.58	0.50	0.078***	0.52	0.53	0.51	0.021
	(0.50)	(0.52)	(0.49)	(0.009)	(0.50)	(0.29)	(1.08)	(0.016)
Foreign	0.09	0.12	0.08	0.043***	0.13	0.14	0.11	0.027***
	(0.29)	(0.34)	(0.26)	(0.005)	(0.34)	(0.20)	(0.67)	(0.010)
Nr. of children	0.56	0.66	0.51	0.142***	0.57	0.57	0.59	−0.023
	(0.98)	(1.08)	(0.91)	(0.018)	(0.98)	(0.56)	(2.15)	(0.035)
Education: Primary	0.21	0.23	0.20	0.028***	0.24	0.24	0.22	0.024*
	(0.41)	(0.44)	(0.39)	(0.007)	(0.43)	(0.25)	(0.89)	(0.014)
Education: Secondary	0.59	0.59	0.60	−0.010	0.58	0.58	0.60	−0.021
	(0.49)	(0.52)	(0.48)	(0.009)	(0.49)	(0.28)	(1.06)	(0.016)
Education: Tertiary	0.20	0.19	0.21	−0.019***	0.18	0.18	0.18	−0.004
	(0.40)	(0.41)	(0.40)	(0.007)	(0.39)	(0.22)	(0.84)	(0.012)
Gross annual earnings	36.09	35.36	36.49	−1.135	34.19	33.93	35.59	−1.658
	(48.35)	(50.57)	(47.10)	(0.877)	(45.81)	(26.26)	(97.48)	(1.444)
Travel distance (km)	28.69	43.02	20.90	22.125***	10.28	10.26	10.42	−0.158
	(27.22)	(37.62)	(15.95)	(0.506)	(4.80)	(2.74)	(10.35)	(0.150)
Travel time (min)	27.80	37.15	22.72	14.434***	13.25	13.22	13.46	−0.240
	(20.27)	(27.79)	(12.98)	(0.378)	(5.24)	(3.00)	(11.23)	(0.165)
Unemployed	0.02	0.02	0.01	0.005	0.02	0.02	0.01	0.008**
	(0.12)	(0.14)	(0.11)	(0.002)	(0.14)	(0.08)	(0.25)	(0.004)
Receives DI in 2001	0.04	0.04	0.04	−0.005	0.04	0.04	0.04	−0.004
	(0.20)	(0.20)	(0.20)	(0.004)	(0.19)	(0.11)	(0.43)	(0.008)
DI recipients								
Years in DI	7.90	7.64	8.03	−0.391	7.41	7.67	6.09	1.582
	(6.94)	(7.48)	(6.62)	(0.646)	(6.68)	(3.71)	(12.70)	(0.967)
Entry age	43.11	44.20	42.55	1.654	45.05	45.26	43.99	1.270
	(11.69)	(13.25)	(10.84)	(1.142)	(11.51)	(6.23)	(24.99)	(2.271)
DI: Psych. problems	0.29	0.27	0.30	−0.028	0.29	0.27	0.35	−0.081
	(0.46)	(0.49)	(0.43)	(0.043)	(0.45)	(0.24)	(1.01)	(0.091)
DI: Nerve	0.11	0.09	0.12	−0.033	0.11	0.11	0.11	0.004
	(0.31)	(0.31)	(0.31)	(0.029)	(0.32)	(0.18)	(0.66)	(0.051)
DI: MSK	0.21	0.27	0.18	0.089**	0.23	0.26	0.12	0.136**
	(0.41)	(0.49)	(0.37)	(0.041)	(0.42)	(0.24)	(0.69)	(0.064)
DI: Other illness	0.21	0.21	0.21	−0.002	0.19	0.20	0.17	0.034
	(0.41)	(0.45)	(0.38)	(0.039)	(0.40)	(0.22)	(0.79)	(0.063)
DI: Accident	0.10	0.09	0.11	−0.025	0.08	0.06	0.19	−0.129
	(0.30)	(0.31)	(0.30)	(0.029)	(0.27)	(0.13)	(0.83)	(0.090)
All individuals	15,522	5,983	9,539		8,570	2,367	6,203	
DI recipients	506	207	299		280	70	210	

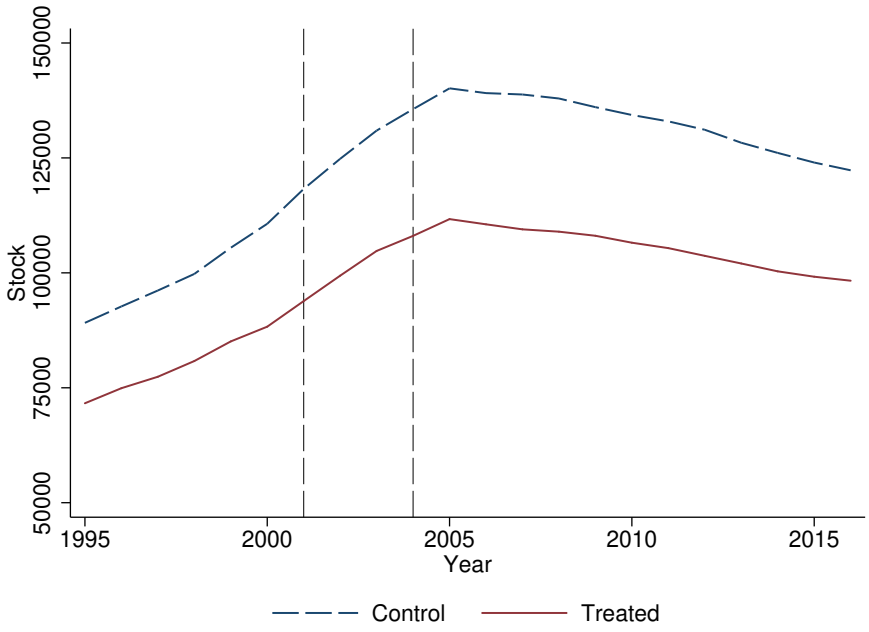
Note: Means of selected covariates for individuals in treated and control regions sampled between 1999–2001, prior to the pilot period. Separate statistics for all individuals and those within a distance of 20 kilometers in border regions. Standard deviation in parentheses. The last column in each block shows the difference between treated and control individuals for each variable, standard error in parentheses. Survey weights applied for the full sample. Observations weighted for pairwise differences in the local sample. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

Figure A1: Sample composition



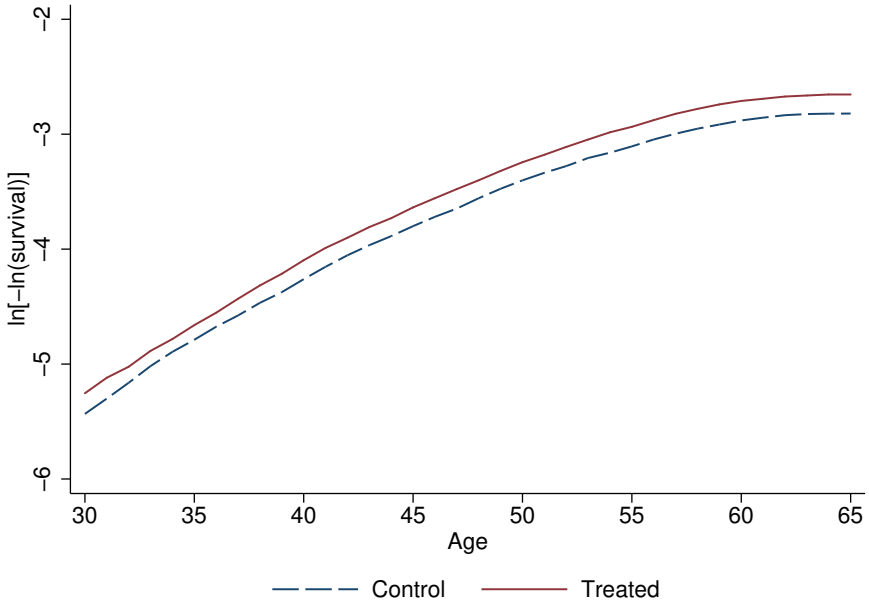
Note: Pilot cantons in green, control in brown. Lighter shades indicate the municipalities that are included in the local sample. The remaining municipalities with darker coloring are included in the full sample. Municipalities in white are never sampled. Lakes shown in blue.

Figure A2: Trends for aggregate disability insurance stock and inflow



Note: Disability insurance stock for treated and control regions.

Figure A3: Log cumulative hazard by age and treatment region



Note: Log-log plot showing unconditional log cumulative hazard estimates by age for individuals in treated and control regions.

Table A3: Placebo reform

	(a) Full sample				(b) Local sample (within 20 km)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment region	1.337***	1.337***	1.337***	1.248***	1.150**	1.150**	1.150**	1.148**
	(0.051)	(0.051)	(0.051)	(0.048)	(0.076)	(0.076)	(0.076)	(0.076)
Pre-pilot time	1.235***	1.241***	1.241***	1.274***	1.204	1.213	1.213	1.253*
	(0.082)	(0.082)	(0.082)	(0.084)	(0.146)	(0.146)	(0.146)	(0.150)
Treat x pre	0.970	0.970	0.970	0.975	0.999	0.999	0.999	0.996
	(0.064)	(0.064)	(0.064)	(0.064)	(0.111)	(0.111)	(0.111)	(0.111)
Pilot time		1.320***	1.326***	1.390***		1.514***	1.525***	1.612***
		(0.129)	(0.129)	(0.135)		(0.228)	(0.229)	(0.241)
Treat x pilot		0.847**	0.846**	0.852**		0.770**	0.771**	0.765**
		(0.069)	(0.069)	(0.069)		(0.092)	(0.092)	(0.091)
Post time			0.842	0.917			1.046	1.142
			(0.094)	(0.103)			(0.207)	(0.226)
Treat x post			0.960	0.961			0.841	0.829
			(0.080)	(0.080)			(0.110)	(0.109)
Other controls	-	-	-	✓	-	-	-	✓
N municipalities	2,336	2,337	2,338	2,338	1,086	1,086	1,087	1,087
N individuals	242,531	249,750	259,323	259,323	124,747	128,633	133,648	133,648
N failures	6,164	7,877	9,204	9,204	3,100	3,985	4,693	4,693
N fail during pilot	0	1,713	1,713	1,713	0	885	885	885
N fail during prepilot	1,950	1,950	1,950	1,950	989	989	989	989
N	439,761	631,782	787,954	787,954	226,345	325,321	406,221	406,221

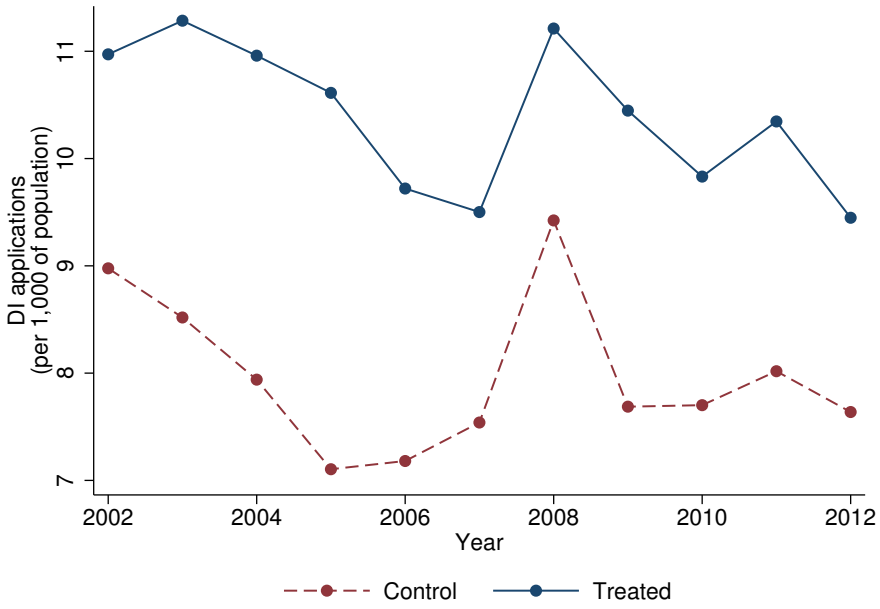
Note: Cox Proportional Hazard estimates for individuals in treated and control regions based on SESAM individual-level survey and administrative data sampled during 1999–2011. Baseline hazard for all regressions stratified by 5-year birth cohorts. Survey weights applied for the full sample. Observations in the local sample are weighted for pairwise estimation. Results are reported in exponentiated form as hazard ratios. The hazard ratio for 'Treat x pilot' corresponds to the relative average treatment effect on the treated as defined in section 4. Standard errors clustered at the individual level in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

Table A4: Distance windows

	(a) Travel distance (km)					(b) Travel time (min)				
	10 km	15 km	20 km	25 km	30 km	10 min	15 min	20 min	25 min	30 min
Treatment region	1.13	1.20***	1.15***	1.16***	1.20***	1.040	1.13	1.18***	1.16***	1.09
	(0.10)	(0.08)	(0.06)	(0.06)	(0.06)	(0.115)	(0.09)	(0.07)	(0.06)	(0.06)
Pilot time	1.29	1.38**	1.27**	1.25**	1.25**	1.469*	1.43**	1.30**	1.32**	1.20
	(0.23)	(0.19)	(0.15)	(0.14)	(0.13)	(0.333)	(0.25)	(0.17)	(0.16)	(0.14)
Treat x pilot	0.75*	0.71**	0.77**	0.78**	0.78**	0.740	0.66**	0.73**	0.76**	0.81*
	(0.13)	(0.10)	(0.09)	(0.08)	(0.08)	(0.164)	(0.11)	(0.09)	(0.09)	(0.09)
Post time	0.92	0.91	0.87	0.82	0.84	1.086	0.87	0.78	0.80	0.80
	(0.24)	(0.18)	(0.15)	(0.14)	(0.13)	(0.337)	(0.21)	(0.15)	(0.14)	(0.13)
Treat x post	0.79	0.83	0.84	0.85	0.85	0.995	0.85	0.86	0.90	0.94
	(0.16)	(0.13)	(0.11)	(0.10)	(0.10)	(0.241)	(0.16)	(0.12)	(0.12)	(0.12)
N municipalities	549	825	1,087	1,286	1,414	372	649	922	1,159	1,371
N individuals	47,403	88,990	133,549	151,215	163,852	26,956	56,609	119,572	143,504	166,486
N failures	1,626	3,230	4,693	5,223	5,690	942	1,948	4,253	5,031	5,752
N failures during pilot	332	612	885	980	1,063	180	379	811	961	1,087
N	107,479	200,431	300,432	340,370	369,235	61,269	128,479	269,155	323,290	375,210

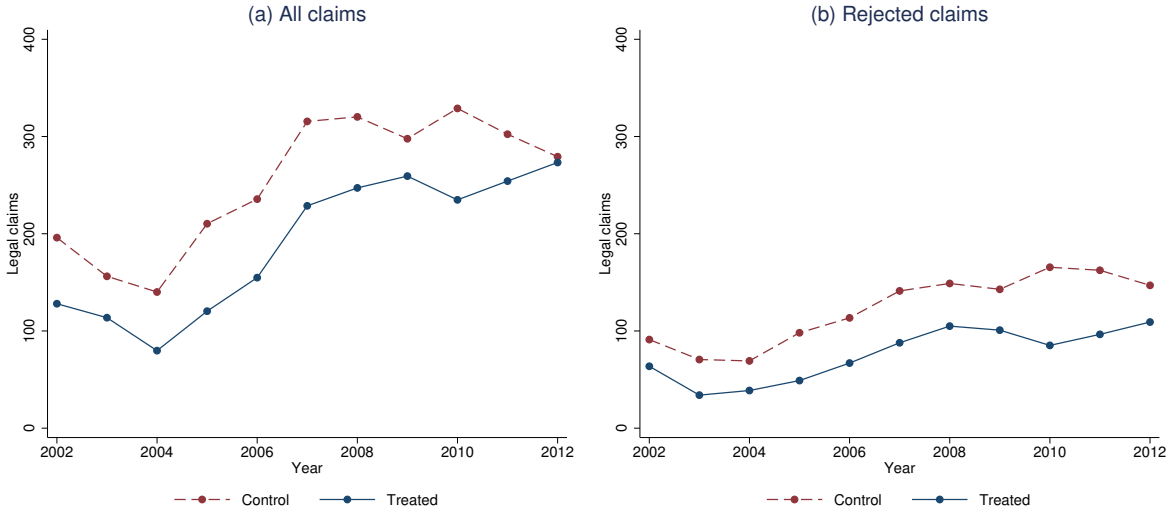
Note: Cox Proportional Hazard estimates for individuals in treated and control regions across various distance windows from the border. Based on SESAM individual-level survey and administrative data sampled during 1999–2011. Observations are weighted for pairwise estimation. Results are reported in exponentiated form as hazard ratios. The hazard ratio for 'Treat x pilot' corresponds to the relative average treatment effect on the treated as defined in section 4. Standard errors clustered at the individual level in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

Figure A4: Disability insurance application rates



Note: Disability insurance application rates per 1,000 of population for the years 2002–2012. Cantons in western Switzerland for which the electronic reporting system is known to have been faulty are omitted from the sample (Fribourg, Genève, Jura, Neuchâtel, Vaud).

Figure A5: Disability insurance court cases



Note: Mean cantonal total and rejected disability insurance legal claims for the years 2002–2012.

Table A5: Stock outflow

	(a) All individuals			(b) Age $\leq$ 50 in 2001		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	0.925***	0.923***	0.911***	0.871***	0.871***	0.872***
	(0.027)	(0.027)	(0.027)	(0.041)	(0.041)	(0.041)
Pilot time	7.698***	7.677***	7.825***	7.515***	7.479***	7.652***
	(0.157)	(0.156)	(0.160)	(0.243)	(0.240)	(0.247)
Treat x pilot	0.985	0.986	0.992	0.995	0.997	0.997
	(0.033)	(0.033)	(0.033)	(0.053)	(0.053)	(0.053)
Post time		7.518***	7.728***		7.676***	7.931***
		(0.152)	(0.157)		(0.236)	(0.246)
Treat x post		1.008	1.014		1.036	1.035
		(0.032)	(0.033)		(0.052)	(0.051)
Other controls	-	-	✓	-	-	✓
N individuals	314,249	327,580	327,580	145,018	154,020	154,020
N failures	20,481	44,529	44,529	8,904	23,547	23,547
N failures during pilot	15,389	15,389	15,389	6,957	6,957	6,957
N	1,032,666	2,489,323	2,489,323	504,801	1,470,137	1,470,137

Note: Cox Proportional Hazard estimates for individuals in treated and control regions based on SESAM individual-level survey and administrative data sampled during 1999–2011. Baseline hazard for all regressions stratified by 5-year birth cohorts. Survey weights applied for the full sample. Observations in the local sample are weighted for nearest-neighbor pairwise differences. Results are reported in exponentiated form as hazard ratios. Standard errors clustered at the municipality level in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

Table A6: Determinants of local sample

	Full sample	Treated	Control
	(1)	(2)	(3)
Age	−0.0004*	−0.0008***	0.0002
	(0.0002)	(0.0002)	(0.0003)
Female	0.0040	−0.0080	0.0159***
	(0.0054)	(0.0063)	(0.0057)
Married	−0.0115	0.0093	−0.0320*
	(0.0165)	(0.0181)	(0.0181)
Foreign	0.0175	−0.0360	0.1117***
	(0.0258)	(0.0270)	(0.0210)
Nr. of children	−0.0030	0.0037	−0.0050
	(0.0041)	(0.0033)	(0.0049)
Education: Secondary	0.0195***	0.0041	0.0189**
	(0.0068)	(0.0066)	(0.0083)
Education: Tertiary	0.0373	0.0008	0.0490**
	(0.0228)	(0.0240)	(0.0239)
N	259,323	117,701	141,622

Note: Probit estimates for the probability to be included in the local sample separately for treated and control regions. Marginal effects at the mean reported. Standard errors clustered at the municipality level in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.