ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Liebert, Helge

Conference Paper

Medical Screening and Award Errors in Disability Insurance

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Social Policy, No. G14-V2

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Liebert, Helge (2015) : Medical Screening and Award Errors in Disability Insurance, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Social Policy, No. G14-V2, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft

This Version is available at:
https://hdl.handle.net/10419/113224

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Medical Screening and Award Errors in Disability Insurance

Helge Liebert*, †

May 21, 2015

Work in progress. Do not cite.
Comments are very welcome!

Abstract

This paper investigates the impact of medical screening on individual disability insurance benefit receipt. Using a unique policy change in Switzerland, I assess the size of award errors in disability insurance and show that improvements in medical screening can reduce insurance incidence by between 14% and 23%. Misclassification is tied to difficult-to-diagnose conditions, indicating inaccurate assessments by treating physicians or health status misrepresentation by applicants. Reductions in full pension benefit awards are partly offset by increases in partial benefits. The overall reduction in inflow suggests that wrongful admissions dominate rejection errors in a setting where explicit state medical screening is absent.

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

†I have benefited from discussions with Eva Deuchert, Beatrix Eugster, Michael Lechner, Per Johansson and seminar participants at the University of St. Gallen, Switzerland and IFAU Uppsala. 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 to imperfect screening of DI applicants (e.g. Autor and Duggan 2003). The efficiency of any targeting system crucially relies on how accurate it is in identifying deserving beneficiaries. Separating meritorious from non-meritorious claims is especially crucial for DI, where substantial indefinite benefits are frequently awarded to individuals with difficult-to-diagnose medical conditions.

Disability benefit decisions are typically made based on assessments of individuals' residual functional capacity, i.e. their remaining ability to work. The first gatekeeper in most countries is the treating physician, who diagnoses the health condition and passes documentation to the insurance provider. These physicians are supposed to have an informational advantage, hence their recommendation is often influential in award decisions. The United States Social Security Administration even adopted a 'treating physician rule' in 1991, giving 'controlling weight' to the treating physician's opinion under certain conditions. However, the treating doctor faces a conflict of interest when evaluating a long-term client. Furthermore, since the treating physician is typically a general practitioner, their supposed informational advantage has come into question with the advent of complex or multidisciplinary disabling conditions, which can often only be diagnosed accurately by medical specialists. Physicians also lack any knowledge of actuarial requirements.

This paper investigates changes in screening quality and their implications for misclassification of DI applicants. Using a massive screening expansion in Switzerland, where treating physicians' assessments are subjected to scrutiny by specialists, I show that better medical screening reduces DI admissions by between 14% and 23%. Misclassification is closely tied to psychological and musculoskeletal conditions, diseases which are more prone to inaccurate diagnoses. Furthermore, the results provide novel insights on the incidence and size of award errors in DI. The reduction in beneficiary incidence suggests that award errors dominate rejection errors and account for large parts of insurance inflow when public medical screening is superficial.

Identification relies on quasi-experimental policy variation generated by the introduction of a medical gatekeeper institution in Switzerland. Generous benefits and low-intensity medical screening render the Swiss system especially liable to moral hazard. However, a 2002 reform instituted new medical screening offices which effectively amounted to a

substantial increase in the medical staff and funding directed towards reviewing applicants' cases. The measures where aimed at improving screening quality, foremost by substantially reducing the individual DI physicians' caseload and also by directing cases to physicians' specializing in the relevant field. The medical offices 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. Previously, DI offices had insufficient resources to screen individuals thoroughly and had to rely on information provided by applicants' general practitioner. Due to legal obstacles, they were also unable to examine applicants in person. The sequential spatial implementation of the reform is exploited in a difference-in-differences design focusing on local labor markets. An age-based duration framework is used for estimation.

A central result of this paper is that better quality screening can substantially reduce DI incidence. Previous evidence regarding the effect of screening on insurance inflow is ambiguous. A paper by de Jong et al. (2011) does not find a reduction in incidence due to stricter screening but provides evidence for lower sickness absence and DI applications, possibly due to self-screening (cf. Parsons 1991). Another related paper by Staubli (2011) shows that stricter screening reduces insurance inflow and affects labor supply. Unlike these papers, I can abstract from mechanical inflow effects which arise due to eligibility requirement changes.

Screening involves two distinct aspects: *Stringency* and *quality*. Many DI policy reforms involve changes to screening stringency, i.e. either explicit or implicit changes to eligibility criteria. This is equivalent to redefining the disability insurance eligibility threshold. Changes in the eligibility criteria may cause a reduction in insurance inflow, but do not improve targeting efficiency. Furthermore, they cannot be used to identify misclassification rates, as a share of rejected applicants may have been deserving under the previous regime, but is now declared fit for work. Given a fixed entry requirement, only changes to screening quality can effectively reduce the amount of misclassification. The main effect found in this paper can be traced to a reduction in the information asymmetry between caseworker and applicant due to the proliferation of better information about the applicant's true health state.

The paper contributes to three main strands of literature. First, 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. The results in this paper provide a lower bound estimate of the false positive classification error rate in these models. Since better screening quality theoretically reduces both the number of type-I (award) and type-II (rejection) errors and screening is observed to reduce the overall amount of benefit awards, this suggests that

award errors occur more frequently in low-intensity screening environments. Although not an exact quantification, this result, unlike earlier studies, does not rely on small sample expert reviews and the assumption of perfect classification in some subsample (Nagi 1969, Smith and Lilienfeld 1971, Benitez-Silva et al. 2004).

Second, it contributes to the economic literature on moral hazard in DI, screening and difficult-to-diagnose conditions. The DI literature has focused largely on investigating moral hazard by evaluating the effect of benefit or screening stringency changes on labor supply (e.g. Gruber 2000, Autor and Duggan 2003, Mitra 2009). However, misrepresentation 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. I provide evidence that award errors are directly tied to such conditions using actual insurance records.

Third, issues with reported health are related 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 general practitioners, 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, if it is deemed to be in the patients best interest, among a substantial minority of physicians (e.g. Novack et al. 1989, Zinn and Furutani 1996, Freeman et al. 1999, Everett et al. 2011). Furthermore, exaggeration and malingering of health limitations by patients in anticipation of insurance benefits has been documented in some studies. However, these issues have never been directly tied to insurance take-up. I show that reviews by clinical specialists without extensive patient contact reduce benefit awards, indicating that lenient diagnoses by treating physicians result in a sizeable and costly number of award errors.

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

2 Institutional Background

2.1 The Swiss DI system and the 2002 reform

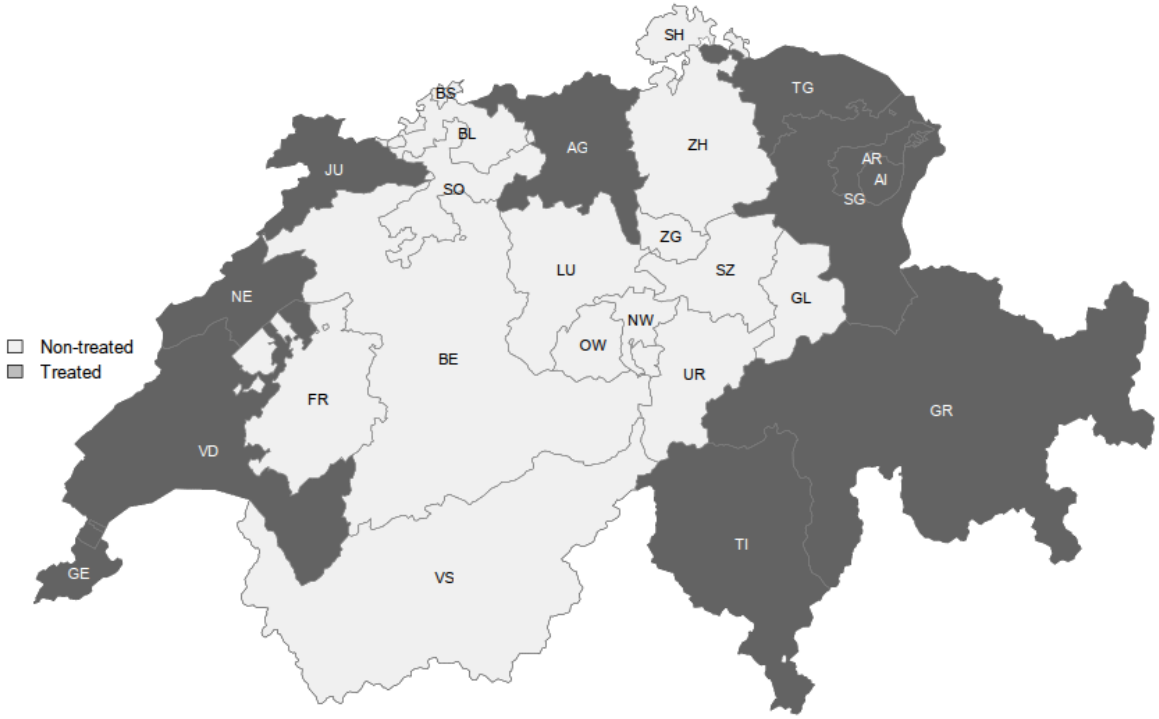
A combination of generous benefits and (previously) low-intensity medical screening render the Swiss DI system an attractive alternative to work. Individuals can expect to receive between 60% and 95% of their final wage in benefits from the main insurance schemes if fully disabled.[1] Exact replacement 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 development of the insurance rolls in Switzerland closely resembles the situation in other developed economies. The number of DI beneficiaries increased substantially during the 1990s, with about 65% of DI inflow due to musculoskeletal conditions or mental health problems. Mounting financial pressure then led to a sequence of revisions of the DI system. The Swiss parliament passed the first major reform in 2003 (4. Revision des Bundesgesetzes über die Invalidenversicherung). The reform created several regional medical screening institutions tasked to conduct (re-)appraisals of benefit claims and authorized to carry out medical examinations. This resulted in a massive expansion of the medical staff who reviewed the insurance applications and extended legal competences for public physicians. To gauge the effect of the institutional changes, regional insurance offices could already hire new staff as part of a voluntary early adopter scheme in 2002. 11 out of 26 Swiss regions chose to participate in this pilot scheme. In the remaining regions, operation began in 2005 as scheduled by the reform proposal. Following the nationwide implementation in 2005, staff funding was then expanded further. This quasi-experimental variation is exploited in the remainder of the paper. The cantons which introduced screening institutions in 2002 are shown in Figure 1.

To apply for benefits, an applicant has to submit the medical documentation of his condition and his previous earnings records. The earnings loss induced by the condition

[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.

must span at least twelve months to qualify for benefits. The disability insurance office then has to assess the individual earnings loss based on the severity of the condition and its impact on work capability. Based on this, 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 this is the applicant's general practitioner, who may be partial towards evaluation in favor of a long-term client. It also allows applicants to 'shop around' for a doctor who writes a more favorable review. Medical studies show that doctors are inclined to conform to their patients' interests when facing a conflict between the patients' welfare and their gatekeeper role, even if this implies deception of the insurance provider (e.g. Novack et al. 1989, Zinn and Furutani 1996, Freeman et al. 1999, Wynia et al. 2000, Everett et al. 2011). Furthermore, there is evidence that general practitioners are more likely to prescribe sick-leave and to succumb to patients' pressure to declare them as sick compared to clinical specialists (e.g. Englund et al. 2000, Kankaanpää et al. 2012). Such effects are potentially exacerbated because prior to 2002, the disability insurance was legally not allowed to examine the applicant, even when in doubt about the credibility or severity of the impediment. The caseworkers deciding on the application also have no medical training, although they could consult with public health officers for clarification. However, the DI offices were notoriously understaffed with physicians. In one office, two doctors were reviewing approximately 40 applications each per day. This implies that on average, each application had to be dealt with in about ten minutes. Due to this, public officials often relied on the medical assessment provided by the treating physician when awarding benefits.

This situation changed with the introduction of the screening institutions. Although the reform branded the screening services as a new institution, it essentially strengthened the role of physicians in the application process by increasing the independent medical staff working at the DI offices and extending their competencies. In the abovementioned office, the medical staff increased fivefold due to the additional funding. Increases in other regions were similarly drastic. The few general practitioners which were already working at the offices were reassigned to the new screening offices. The new staff consists of clinical specialists which are then trained in the 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. The staff is instructed to focus on new DI applicants and aid with scheduled revisions of existing beneficiaries claim status.

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 likelihood that inconsistencies are noticed can be expected to increase due to the substantial reduction in caseload per physician. Furthermore, the offices are expected to reduce asymmetric information by providing a more qualified judgement about eligibility due to a more competent and impartial medical assessment. Although the treating physician may know his client well, he may not be qualified to diagnose specific complex muscoloskeletal or psychological limitations accurately, or assess their implications for residual functional capacity correctly. 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%.

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 reduces communication deficiencies, i.e. the evaluation is framed in terms the local insurance office can better comprehend.[3]

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 is a crucial issue: The regulatory setting remains unchanged, only the provision of information about the subjects' eligibility regarding health limitations is affected by the reform such that the award decision is made on more qualified grounds.

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

[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). Due to data limitations, the authors are not able to make a decisive causal statement about whether reduced inflow rates can be associated to the introduction of the screening offices.

2.2 Misclassification errors and medical screening

It is the main duty of the insurance office to separate meritorious from non-meritorious claims ('tag' the eligible). Two types of classification errors can occur in this situation: (1) Award errors (Type-I) and (2) Rejection errors (Type-II).[4] If screening is imperfect, benefits may be awarded to persons who are actually ineligible, and deserving applicants may be denied benefits.

Hence, better screening may not necessarily reduce insurance inflow unambiguously. Improvements in screening quality as described above increase the probability to detect applicants' true type. As such, better screening most likely reduces *both* type-I and type-II misclassification if the effect on the detection probability is symmetric, resulting in opposing effects on the incidence rate. The net effect on inflow is undetermined. If type-I errors dominate type-II errors and the likelihood of detection for both is affected equally by the reform, a reduction in the inflow rate will be observed. In case type-II errors are completely absent, the identified effect is the change in the amount of type-I misclassification. Otherwise the estimate provides a lower bound. Similarly, unless screening is perfect ex-post, it provides a lower bound for the total amount of award/rejection errors.

I expect award errors to be larger in size than rejection errors and the net effect of screening on inflow to be negative for several reasons. Although a study by Benitez-Silva et al. (2004) for the US suggests that rejection errors are comparably larger, the Swiss setting differs in key aspects. The situation prior to the reform is characterized by extremely low screening intensity and high benefits, rendering benefit receipt both comparably more desirable and attainable. Furthermore, there is no lengthy appeals process similar to the US, where the majority of applications is initially rejected by default. Although legal appeal is possible, it occurs infrequently and is rarely successful. Less than 20% of appeals result in an award. Since the award decision is essentially one-shot, caseworkers are tend to be more well-meaning and cautiously err on the side of awards rather than rejections.

3 Empirical Strategy

3.1 Identification and Estimation

The main quantity of interest is the change in the population DI hazard induced by the screening measures, 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

[4] The terminology partly follows Kleven and Kopczuk (2011), but in accord with the rest of the literature, the no disability case constitutes the null hypothesis.

to be random. The cantons participating in the pilot program are a mixture of high and low prevalence regions, and regional considerations were relevant in the assignment process. A difference-in-differences identification approach is used to evaluate the impact of the medical screening institutions. Differencing removes time-invariant influences on potential outcomes. However, identification still requires a common development of DI incidence in the absence of the screening expansion. This assumption raises some concerns.

As Autor and Duggan (2003) illustrate, people rarely transition directly from employment into DI, but typically apply conditional on job loss. The main concern 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. 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, remaining heterogeneity among Swiss regions may raise concerns about biased treatment effect estimates.

To address this issue, I pursue 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. Similar strategies are used by Frölich and Lechner (2010) and Campolieti and Riddell (2012).

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 $h(t, P, D|X < \bar{x}) = h_{0g}(t) \exp\Big(\beta_0 P + \beta_1 D + \beta_2 PD\Big)$, where $h_{0g}(t)$ is the non-parametric baseline hazard within 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). As recommended by Kom et al. (1997) and 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. 2014).

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. As the DI entry year is observed for each recipient, irrespective of the sampling date, pooling all data increases power substantially. Doing so also limits the possibility of implicit sampling bias. With inflow observed retrospectively, other approaches would require creating a pseudo-panel structure by inferring past incidence figures from a chosen post-treatment cross-section and adjusting for past eligibility. Since the disability risk is concentrated at older ages near the official retirement age, bias due to intermittent drop out is a real concern when extrapolating past incidence. 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.[5]

The standard assumptions for difference-in-differences estimation have to be restated for proportional hazard models. The common trend assumption is not invariant to the scaling of the dependent variable and is modified accordingly. The main identifying assumption is that in the absence of stricter screening measures, incidence for individuals in both pilot and non-pilot (border) regions would have changed proportionally. Instead of assuming a common trend between regions over time in differences, I am assuming a constant hazard

[5] Another possibility would be to analyze prevalence, i.e. the effect of screening on the stock of disability insurance beneficiaries as in Staubli (2011). 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. 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, the screening measures are much more likely to affect the rates of newly awarded benefits immediately before effects on the stock of recipients eventually materialize.

ratio, i.e. a common relative change or a common absolute change in logs. In addition, I assume that anticipation effects are absent. Explicit identifying conditions are given in the appendix. Given these assumptions, the coefficient of the interaction between treatment time and region identifies the relative change in the hazard for the treated, i.e. a relative average treatment effect on the treated,

$$\text{rATT} = \frac{h^1(t, D = 1, P = 1)}{h^0(t, D = 1, P = 1)} = \exp\left(\beta_2\right) ,$$

where h^D denotes potential hazard rates.

This strategy removes unobservable factors which have a time-invariant effect on log potential outcomes. I assume that conditional on being in the same local labor market, there are no trend-confounding factors. It is worth noting that there are no other region-specific reforms during the relevant time period. Since people living on different sides of the border are subject to different policies, the causal effect of the reform on DI uptake can be identified by comparing hazard rates close to the border across time.

3.2 Potential threats

Since prospective reform changes may induce some individuals to change their behavior in anticipation of losses, the chronology of events is relevant. The early adopter scheme was introduced shortly after the reform proposal got public and it was never publicly announced. Communication only occurred internally between the Federal Ministry of Social Insurances and the DI offices. Overall, the screening changes implied by the reform proposal received little public attention and were only scheduled to be implemented in 2005.[6] The first draft of the reform which included the medical screening institutions was proposed in parliament in February 2001, and underwent some revisions until being approved by popular vote in March 2003. The project began in January 2002. The rapid introduction of the project within ten months alleviates concerns regarding anticipatory behaviour in treated regions, as the operation of the screening institutions began almost immediately after the reform was announced. Similarly, there is only a short time period between the reforms definite approval in March 2003 and its nationwide implementation in January 2005. Given the one-year earnings loss restriction required for eligibility, these time frames leave limited scope for the strategic timing of applications in both treated and control regions.

Anticipation effects can also manifest in increased mobility. Individuals considering to apply for disability benefits may anticipate the reform and move to regions where the

[6] Other reform measures 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 2004. There were no further reforms to DI or other social insurances during the introduction period.

screening expansion is not implemented, generating higher inflow in control regions and biased results. This can be dismissed for similar reasons. The screening measures were not announced publicly at the time. In addition, the amount of people moving to another region who can be identified by tracking panel cases in my data is negligible. Between 1999 and 2011 about 3.1% of the people for whom some 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.

Another potential concern is that results are confounded by changes in self-screening (Parsons 1991). However, due to the hidden nature of the screening offices and the changes during the introduction period, it is unlikely that potential candidates were deterred from applying. Information about the pilot program did not transpire to the media or the general public, and much of the screening offices work is hidden, even for applicants.

4 Data

The main estimations are 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. Given the survey weights, the data is representative of the Swiss population.

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. It also includes information about individuals' municipality of residence which is used to track individuals distance to their nearest treated/control counterpart. I measure DI inflow using the age of first disability receipt as the main outcome in a duration framework. In addition, I also observe the specific limitation that ultimately led to the DI award.

The treatment region is defined as the cantons participating in the pilot project, the treatment period comprises the years 2002–2004. For each individual, the municipality of residence is observed. The spatial treatment assignment is exploited in the estimations. Two different samples are used when evaluating the impact of the reform on the treated population. The unrestricted 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.

To identify individuals in the vicinity of administrative borders, spatial information is required. Data about distances between different municipal centroids is 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. Weights are computed to estimate nearest-neighbor pairwise differences to account for spatial clustering of municipalities. 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 estimation sample. Microcensus data on mobility show that 80% of commuters stay within this distance limit, and it corresponds approximately to the average commuting time in Switzerland of about 25 minutes (BSV 2012, Eugster and Parchet 2011). Results are robust to the choice of distance measure, variations in the threshold level and whether weights are applied. The exact municipalities included in the local sample are mapped in Figure A1 in the appendix.

As a balancing test, 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.

A second dataset used to analyse changes in the stock of beneficiaries was 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

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 3. 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.

on individuals with benefit receipt prior to treatment in 2001, such that results are unconfounded by new entries to the DI payroll.

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, among them 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 highly significant, indicating that screening 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 the reform was extended to the federal level after 2004 and funding increased even further. However, the estimates for the local sample are imprecise as the failure mass density is too thin 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.

Better screening is also likely to affect the classification of the severity of health impediments for new awards. It is interesting to see whether this 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 award errors occur mainly for serious health limitations. However, it is unlikely that only serious limitations constitute the affected marginal cases. A more likely scenario is that applicants across all health levels tend to overstate health limitations and likewise, award errors occur across all disability levels. With better screening, some individuals who would have received a full pension previously are now downgraded, resulting in a zero net effect for partial DI benefits.

5.2 Incidence of difficult-to-diagnose conditions

The main analysis indicates that DI awards declined substantially due to better screening measures, 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 hard 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%, although this may be an obscurity due to the especially low incidence for this group. Column (5)

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 3. 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: MSK	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 3. 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.

looks at the incidence of musculoskeletal 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 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 improved screening measures.

5.3 Further evidence: Disability degree and benefit revisions

Although the primary task of the medical staff is to screen applicants, they also aid with reviews of recipients' disability degree classification. Although scheduled by law to occur regularly, such revisions until then rarely resulted in actual disability degree or benefit cuts and typically constituted going over beneficiaries files without personal contact. Files which are scheduled for review are now also passed to the screening physicians. Total denial of benefits after a revision occurs only in exceptional cases. Revisions more commonly take place if applicants have submitted new medical information, typically documenting deteriorating health, and result in pension increases.

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 4. 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 benefits. The effect magnitudes are small since reclassification remains a rare event. Only 9.3% of individuals of the 2001 stock are reclassified during the pilot period. Upward revisions are far more common, downward changes only account for 2.3 percentage points. Still, the screening offices appear to aid in revising 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. Cuts are most pronounced for those who receive DI due to mental health problems or musculoskeletal conditions, while beneficiaries with congenital diseases or handicaps incurred in accidents are unaffected. Unfortunately, nerve-related diseases are not tagged in this data, unlike previously.

Table 4: Stock reclassification and pension cuts

(a) Disability degree

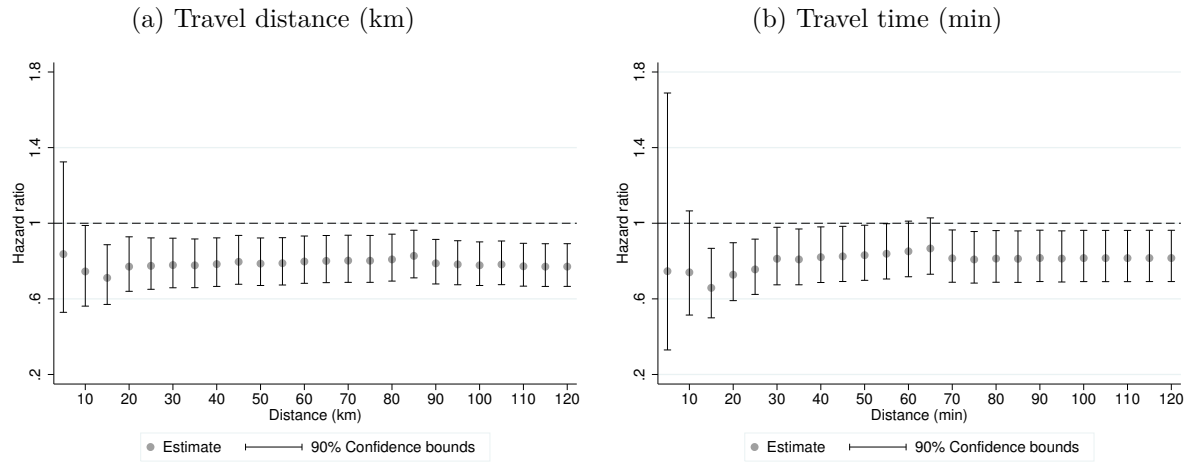
	All	Illness	Psychological	MSK	Accident	Congenital
Treated region	3.04***	3.41***	3.66***	2.78***	1.67***	1.19***
	(0.08)	(0.09)	(0.13)	(0.18)	(0.26)	(0.18)
Pilot	0.49***	0.60***	0.60***	0.39***	0.46***	0.30**
	(0.06)	(0.07)	(0.09)	(0.13)	(0.17)	(0.13)
Treat x pilot	−0.35***	−0.42***	−0.58***	−0.39*	−0.24	−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	Illness	Psychological	MSK	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. The coefficient for 'Treat x pilot' corresponds to the absolute average treatment effect on the treated as defined in section 3. Standard errors in parentheses, number of observations given below. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

Figure 2: 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.

5.4 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 2 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 issue that can be raised pertains to bias incurred by selective sampling. 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. While theoretically possible, this issue is unlikely. The primary focus of the screening offices are new applications. Withdrawing the right to a pension is connected to severe legal obstacles and different from adjusting benefits. Expulsions are an extremely rare occurrence. Nevertheless, 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 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.

Finally, one concern that has been raised is that the screening procedures 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.

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 more 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 is 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 in absence of the reform at least 14% of inflow would have been misclassified, resulting in erroneous benefit awards. Results from the local approach have the same sign and are comparable in magnitude to the global approach. The distance variations consistently suggest about a reduction in the hazard of about

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

20%. Although it is an advantage of the Cox model that hazard ratios can be estimated without explicitly specifying the baseline hazard, with the effect on the hazard seemingly large, it is illustrative to at least get an idea of how large the absolute effects induced by the screening expansion 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 in a thousand enter DI. Improved screening reduces this by about 23% to 0.29%, implying that approximately one person less in a thousand enters DI due to better medical screening.

The reduction in the DI hazard has further implications. If improved screening decreases the probability of type-I and type-II errors equally, this indicates that DI award errors occur more frequently than rejection errors. However, it is theoretically possible that this finding is driven by screening institutions inducing an even larger number of false negative errors. This would imply that insurance offices now reject more applicants that are actually deserving than previously. The main argument is that in case incidence figures are not reduced as politically desired, individual physicians might be tempted to be generally more critical when reviewing new applications due to a fear of being let off. However, the additional staff at the screening offices was hired on permanent contracts and could not have been let 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 screening physicians had the explicit mandate to improve the accuracy of medical diagnoses of functional limitations and received a specialized training specifically tailored for this purpose. It is unlikely that misguided incentives are driving the effect.[8]

Another explanation for lower incidence could be political pressure to increase stringency. However, the general institutional structure and the regulation remains unchanged. The final decision still lies with the DI caseworker. The implementation of the screening services and the DI decision are made on the local level. Besides providing the funding, federal political influence on local public entities is limited due to the decentralized nature of the Swiss political system. Even if political pressure were exerted to increase stringency, the trend would have to differ between regions to confound the result.

More likely, a substantial share of applications is actually rejected because the applicants are truly undeserving given the legal requirements. Since only the provision of information available to the case workers deciding on the application is improved, this implies that a non-negligible fraction of DI awards are made in violation of the requirements. It also suggests that, unless all applicants are completely myopic regarding their eligibility status, some individuals might exaggerate health limitations to their doctor or treating physicians

[8] The physicians accompanying the implementation were in fact acutely aware of the possibility that more intense screening could possibly 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.

consistently diagnose inaccurately, favoring their clients.

7 Conclusion

The results show that screening improvements reduce DI benefit awards substantially, indicating that award errors occur frequently. Effect magnitudes are substantial: Between 14% and 23% of DI inflow can be attributed to false positive award decisions most likely based on incomplete medical assessments. This is comparable to estimates by Benitez-Silva et al. (2004) and earlier medical review studies for the US (Nagi 1969, Smith and Lilienfeld 1971), who estimate award errors to be about 20% of inflow. Misclassification is closely tied to difficult-to-diagnose conditions, suggesting a more accurate assessment of complex or multidisciplinary diseases. This evidence 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. Without scrutiny, applicants are on average classified as more disabled. The tentative evidence that award errors are larger in size than rejection errors implicitly opposes the finding by Benitez-Silva et al. (2004) for the US, who suggest that about 60% of rejected applicants are disabled. This divergence can be related to key differences in the institutional setting. The institutionalized application, screening and appeals process in the US with its many stages and high initial rejection rates differs from the situation in Switzerland, where explicit screening is of low-intensity and appeals are limited. Separating type-I and type-II classification errors more cleanly remains a promising pursuit for further research.

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 born by the applicant. This results in improved screening quality, but is unlikely to discourage take-up strongly in the long-term. As such, expert medical review boards appear to be an effective tool in curbing inflow rates into disability insurance and improving targeting efficiency. This is strengthened by the fact that the introduction of the screening offices is likely to be cost-effective. Simple back-of-the-envelope calculations indicate that initial outlays of 20 million CHF and variable costs for a moderate number of employees are quickly offset by reductions in the beneficiary payload if inflow reductions are permanent, even if all rejected applicants eventually receive social assistance.

Since institutionally separate 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 the Swiss setting represents a drastic policy change when drawing further policy conclusions. Prior to the reform, screening efforts were only marginal and public health officers could not decree

medical checks. Many countries already have more elaborate screening and application mechanisms in place. Both classification errors and the policy impact may well be lower, depending on the initial level of screening intensity.

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.

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), pp. 644–650.

Bütler, M., Deuchert, E., Lechner, M., Staubli, S. and Thiemann, P. (2014). Financial work incentives for disability benefit recipients: Lessons from a randomised field experiment. *IZA Discussion Papers 8715*, Institute for the Study of Labor (IZA).

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), –.

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.

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

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.

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.

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 (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.

Appendix: Identification

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

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

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,

$$\text{rATT} = \frac{h^1(t, D=1, P=1)}{h^0(t, D=1, P=1)} \ , \tag{2}$$

where h^D denotes potential hazard rates. I assume the observation rule (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

$$h^1(t, D=1, P=0) = h^0(t, D=1, P=0) \qquad \text{(no anticipation)} \ , \tag{3}$$

$$\frac{h^0(t, D=1, P=1)}{h^0(t, D=1, P=0)} = \frac{h^0(t, D=0, P=1)}{h^0(t, D=0, P=0)} \qquad \text{(common trend)} \ . \tag{4}$$

The main identifying assumption is that in the absence of stricter screening measures, incidence for individuals in both pilot and non-pilot (border) regions would have changed 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. Given these assumptions, the coefficient of the interaction identifies the hazard ratio of interest.

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

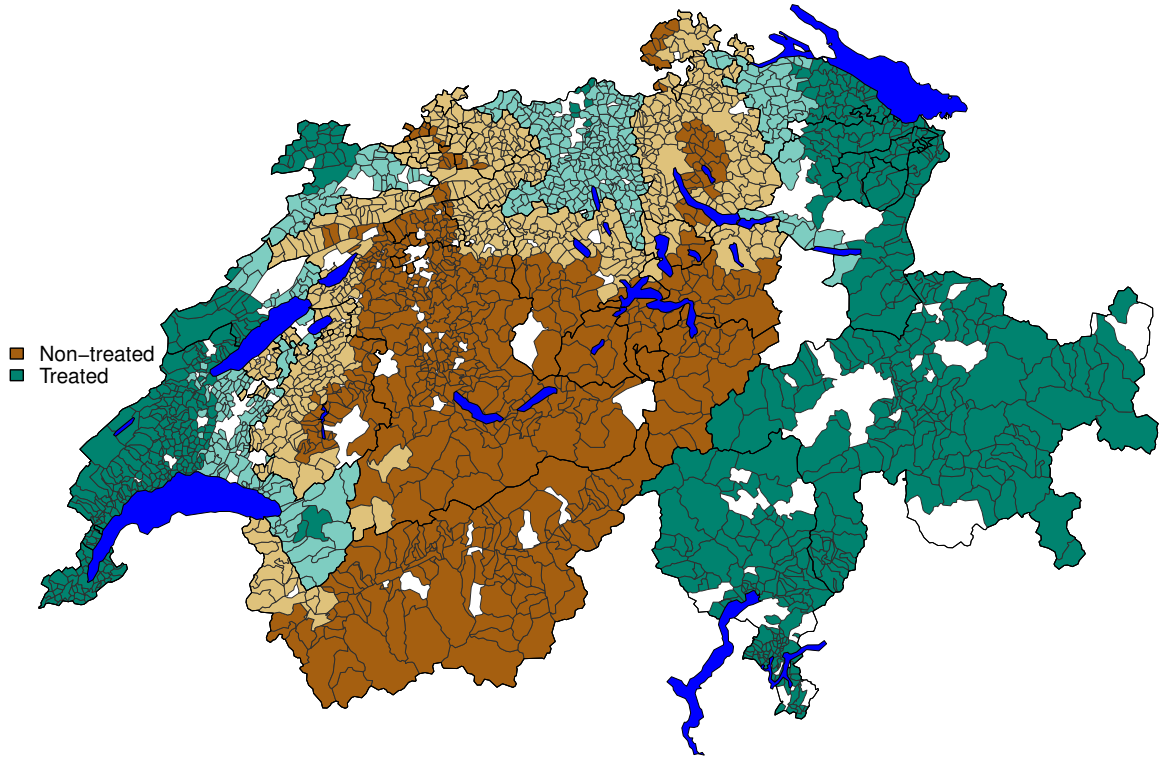
Note: Descriptive statistics for the unrestricted and the local estimation sample. Based on the 1999–2011 SESAM data.

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 across municipalities



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.

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 3. 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 3. 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 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.