Does external medical review reduce disability insurance inflow?

Helge Liebert[*, †]

March 2019

Journal of Health Economics, Vol. 64, 108–128

Abstract

This paper investigates the effects of introducing external medical review for disability insurance (DI) in a system relying on treating physician testimony for eligibility determination. Using a unique policy change and administrative data from Switzerland, I show that medical review reduces DI incidence by 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 benefit awards are partly offset by increases in partial benefits. More intense screening also increases labor market participation. Existing benefit recipients are downgraded and lose part of their benefit income when scheduled medical reviews occur. Back-of-the-envelope calculations indicate that external medical review is highly cost-effective. Under additional assumptions, the results provide a lower bound of the effect on the false positive award error rate.

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

[†] I thank the editor and three anonymous referees for their valuable comments. The paper benefited from discussions with Simone Balestra, Eva Deuchert, Beatrix Eugster, Per Johansson, Rafael Lalive, Michael Lechner, Nicole Maestas, Beatrice Mäder and seminar participants at the University of St. Gallen, the University of Uppsala/IFAU, the 2015 SOLE/EALE meeting in Montreal and the 2018 European Workshop on Health Economics and Econometrics in Groningen. 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 programs constitute the most common form of social protection worldwide. Benefit payments are disbursed to groups identified by a common characteristic – families, the unemployed or persons with a work-limiting disability. Among the different social programs, disability insurance (DI) is by far the most costly. The average OECD country spends about 2.3% of GDP on disability-related benefits (OECD 2010). In both the United States and Europe, the number of DI beneficiaries has been rising throughout the late 20th and early 21st century and recently stabilized on a high level—on average about 6% of the working age population in OECD countries receive disability benefits (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 health 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).

Disability benefit decisions are made based on medical assessments of individuals' residual functional capacity, i.e., their remaining ability to work. However, the medical assessment process required for eligibility determination differs across countries. In 40% of the OECD countries surveyed in OECD (2003), the first gatekeeper to the DI system is the treating physician. In Norway, Switzerland and the United States—countries which are characterized by high rates of DI prevalence—treating physician testimony has historically often been decisive for claims decisions. Treating physicians also hold an influential role in the DI determination process in Australia, Denmark, Germany, Sweden, and the United Kingdom. In these DI systems, the treating physician submits the medical documentation of applicants' diagnosis and treatment history to the DI administration. After submission, the documentation is reviewed by caseworkers—and potentially also by DI physicians.

Whether treating physicians or DI-appointed physicians alone should assess 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. Zinn and Furutani 1996, Englund et al. 2000, Kankaanpää et al. 2012). Moreover, treating physicians are often general practitioners and not clinical specialists, and it is unclear whether complex disabling conditions can be accurately diagnosed by treating physicians. For these reasons, treating physicians' assessments are commonly subjected to *medical review* by DI physicians, who are often clinical specialists.

This paper evaluates the effectiveness of external medical review and its implications. Identification relies on quasi-experimental policy variation generated by an extensive pilot program that preceded the nationwide introduction of mandatory medical review in Switzerland. For the analysis, I develop a combined difference-in-differences and spatial matching approach, embedded in an age-based duration analysis framework for estimation. The results indicate that introducing medical review reduces DI admissions by 23%. Reductions are closely tied to psychological and musculoskeletal conditions, diseases which are more prone to inaccurate diagnoses. Medical review also increases labor market participation. In an extension to the main analysis, I provide explicit identifying conditions under which the inflow reduction can be interpreted as a bound on the reduction in DI award errors. Looking at the stock, I find that existing benefit recipients are downgraded and lose part of their benefit income when scheduled medical reviews occur. Finally, I demonstrate that medical review is highly cost effective.

In 2005, external medical review became mandatory for all DI applications in Switzerland. This reform was preceded by a pilot, which introduced mandatory medical review in several Swiss cantons already in 2002. Medical review in this context means file-based review, exchange with treating physicians and personal examinations by official DI and other third-party physicians. The reform had three major components. First, it substantially increased the medical staff and funding directed towards reviewing DI applicants' cases, more than doubling the number of full-time equivalent staff positions. Screening quality was improved by substantially reducing the individual DI physicians' caseload and by directing cases to physicians' specialized in the relevant field. Second, 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' treating physicians for their decision, as the DI offices had insufficient resources to screen individuals. Third, the policy also abolished legal obstacles that prevented DI physicians from examining applicants in person or requesting further documentation. Meanwhile, the decision structure remains unchanged, the final eligibility decision remains with the responsible DI caseworker.

This paper contributes to the literature on screening in DI by investigating medical review, a form of screening which has so far been largely neglected. DI screening involves two distinct aspects: *stringency* and *quality*. Interestingly, while screening has received considerable attention in the literature on DI, studies on screening in DI almost exclusively focus on variations in screening stringency and use them to obtain a control group to identify the disincentive effect of DI on labor supply (e.g. Karlström et al. 2008, Mitra 2009, de Jong et al. 2011, Staubli 2011, Maestas et al. 2013, French and Song 2014). These studies rely on either explicit or implicit changes to eligibility criteria and the admittance threshold for identification and generally find positive labor supply effects of screening.

For example, de Jong et al. (2011), Maestas et al. (2013) and French and Song (2014) rely on variations in adjudicator stringency, while Karlström et al. (2008) and Staubli (2011) rely on explicit policy reforms that limited eligibility for certain groups. Looking at DI in Austria, Staubli (2011) shows that stricter eligibility requirements both reduce insurance prevalence and increase labor supply. Naturally, these studies also often find lower take-up rates of DI because individuals become mechanically ineligible for DI due to changes in the admittance criteria.

In this paper, I focus on the implications of medical review, an intervention that influences screening quality by providing more information on individuals' underlying capacity to work. Looking at medical review allows abstracting from mechanical inflow effects which arise due to implicit eligibility requirement changes. Unlike stringency changes, medical review does not inherently involve a trade-off between false positive and false negative decision errors (e.g. Kleven and Kopczuk 2011, Low and Pistaferri 2015). Since medical review is primarily targeting new DI applicants, I focus explicitly on insurance *incidence* (inflow) in the analysis, since *prevalence* (stock) is likely to be more inert. In addition, research has shown that inducing work take-up among long-term beneficiaries can be difficult and results regarding the employment capabilities of this group are mixed (e.g. Kornfeld and Rupp 2000, Adam et al. 2010, Borghans et al. 2014, Bütler et al. 2015, Moore 2015, Garcia Mandico et al. 2018).

Moreover, the results in this paper also relate to the findings of health condition-dependent effect heterogeneity in the literature on disincentive effects of DI and the literature on misreporting of health status. In a seminal paper, Bound (1989) finds that up to half of DI recipients in the US would be working in the absence of DI. Newer studies have confirmed Bound's (1989) main result, but also show that there is considerable effect heterogeneity (e.g. Chen and van der Klaauw 2008, von Wachter et al. 2011, Maestas et al. 2013, French and Song 2014). Results by von Wachter et al. (2011) indicate that especially employment of younger individuals and those who applied based on mental health and muscoloskeletal conditions would be non-negligible in the absence of DI. Related to this, Campolieti (2006) notes that stricter DI entry requirements cause fewer reports of these difficult-to-diagnose conditions among older males.[1] Using administrative records, I show that medical screening reduces insurance inflow of difficult-to-diagnose conditions and increases labor market participation. This effectively ties excess inflow of individuals capable of working to certain conditions and suggests that medical review is a cost-effective policy to reduce it.

[1] Other studies have observed that self-reports of disability differ from objective measures of functional limitations and that individuals out of the labor market tend to overstate health limitations (Butler et al. 1987, Kreider 1999, Kreider and Pepper 2007, 2008). 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) and the literature on worker compensation schemes (e.g. Staten and Umbeck 1982, Bolduc et al. 2002).

Finally, an extensive theoretical literature investigates the implications of imperfect tagging in social insurances. Since disability status is private information, it is inferred by the insurance with error. 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 classification errors directly. Given auxiliary assumptions, the results in this paper provide a tentative lower bound estimate of the effect of medical review on the false positive classification error rate in these models. In addition, the results suggest that award errors most likely exceed rejection errors, a finding that diverges from the results for the US. Although not an exact quantification, these results, unlike earlier studies, do not rely on small sample expert reviews and the assumption of subsample perfect classification (Nagi 1969, Smith and Lilienfeld 1971) or a comparison with self-reported disability status (Benitez-Silva et al. 2004).

Taken together, many results in the paper are also closely related to the findings by Low and Pistaferri (2015), who analyze the trade-off between incentives and insurance in DI using a life-cycle model. Among other results, they find that false acceptances exist especially among individuals with moderate limitations, which can be related to the result that medical review is especially effective for soft, difficult-to-diagnose health conditions, which are only partially work limiting. Since welfare effects in their model are dominated by coverage for the severely work-limited, they pose whether allowing for partial disability and partial benefits may be a way to reduce incentive costs. While I cannot make a statement about costs relative to a binary DI system, my results are obtained within a partial benefit system, indicating that incentive costs still matter with partial classifications and that misclassification is a question of degree.

In sum, the paper provides three distinct contributions. First, I show that medical review is cost-effective in reducing and downgrading inflow 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 possible moral hazard on the side of applicants. Third, I provide explicit conditions under which the inflow reduction implied by the reduced-form estimate 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 3 introduces the data, section 4 covers identification and estimation methods, section 5 discusses the results and section 6 concludes.

2 Institutional background

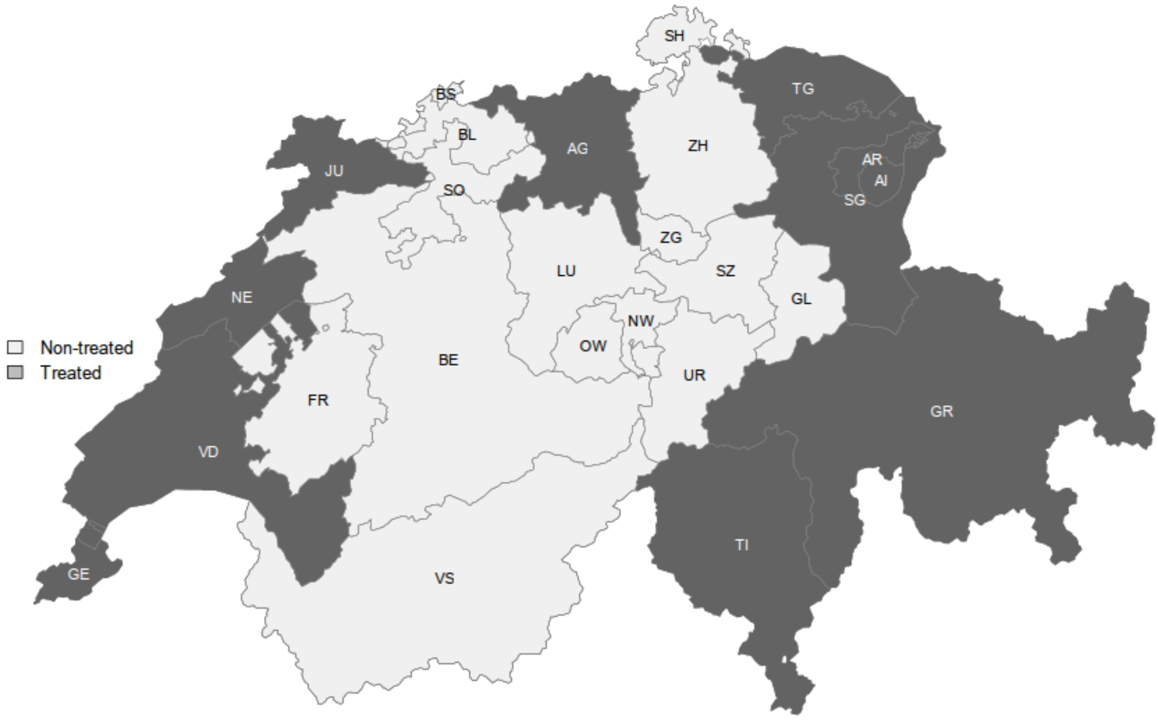
The Swiss DI system is characterized by generous benefits. Individuals can receive benefits from three main benefit schemes: mandatory public DI, mandatory employer-provided occupational pensions and optional private DI. Eligibility for benefits is determined by the local public DI office responsible for the main mandatory public scheme and binding for all other benefit providers. Replacement rates are based on an individual's previous income, contribution history, whether the individual receives full or partial benefits and the family situation. The full benefit amount from the mandatory public DI scheme is capped between 1,175 CHF and 2,350 CHF per month before taxes, depending on prior income, marriage and contribution history. Individuals with children receive an additional 40% of this amount for each dependent child. In addition, there are income-contingent benefits for spouses and means-tested supplementary benefits for recipients who fall below the subsistence earnings threshold. The additional payouts from the mandatory occupational pension scheme vary based on the contribution length and the employers contract terms. Focusing only on the two mandatory schemes, a 40 year old adult with full contribution history and average wage can expect a replacement rate of 70% if single, 80% if married, and 100% if married with two children. At earnings below the average wage, the replacement rate increases sharply up to 120%, exceeding the prior earnings level (OECD 2006, 2010).

Eligibility status and the benefit amount from the main public DI scheme are determined 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 (UI), 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 DI caseworkers made their decisions based on medical assessments submitted by the applicants' treating physician. The treating physician-based screening procedure had been in place unrevised since 1973. The reform resulted in a large expansion of the medical staff available for review of insurance applications and substantially extended their legal competences. Physicians were tasked to conduct (re-) appraisals of benefit claims and authorized to carry out medical examinations.

To assess the effect of the institutional changes, the Federal Ministry of Social Insurances devised a pilot scheme. Beginning in 2002, 11 out of 26 cantons could already hire new staff and conduct medical review. In the remaining cantons, operation began in 2005

Figure 1: Cantons with medical review during the pilot



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.

as scheduled by the reform proposal. Following the nationwide implementation in 2005, staff funding was expanded further. The cantons that introduced medical review in 2002 are shown in Figure 1. The cantonal DI offices operate autonomously, but hold a yearly joint conference, during which participation in the early adopter program was decided (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. Applicants must register with the DI office corresponding to their place of residence and cannot file for benefits elsewhere. When filing a benefit claim, applicants have their treating physician 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 treating physician, typically the applicant's general practitioner. 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 could consult with physicians working at the DI offices if they deemed it necessary. However, the DI offices were notoriously understaffed with physicians. In 2006, the average DI physician reviewed about 612 dossiers per year. Considering the changes in manpower, this figure would have to be 2.25 times as high prior to the reform to ensure the same coverage given that application numbers remained constant (Appendix Figure A4). For this reason, only a subset of selected dossiers were passed to the DI physicians for inspection. Caseworkers were reliant on the medical assessment provided by the treating physician when awarding benefits.

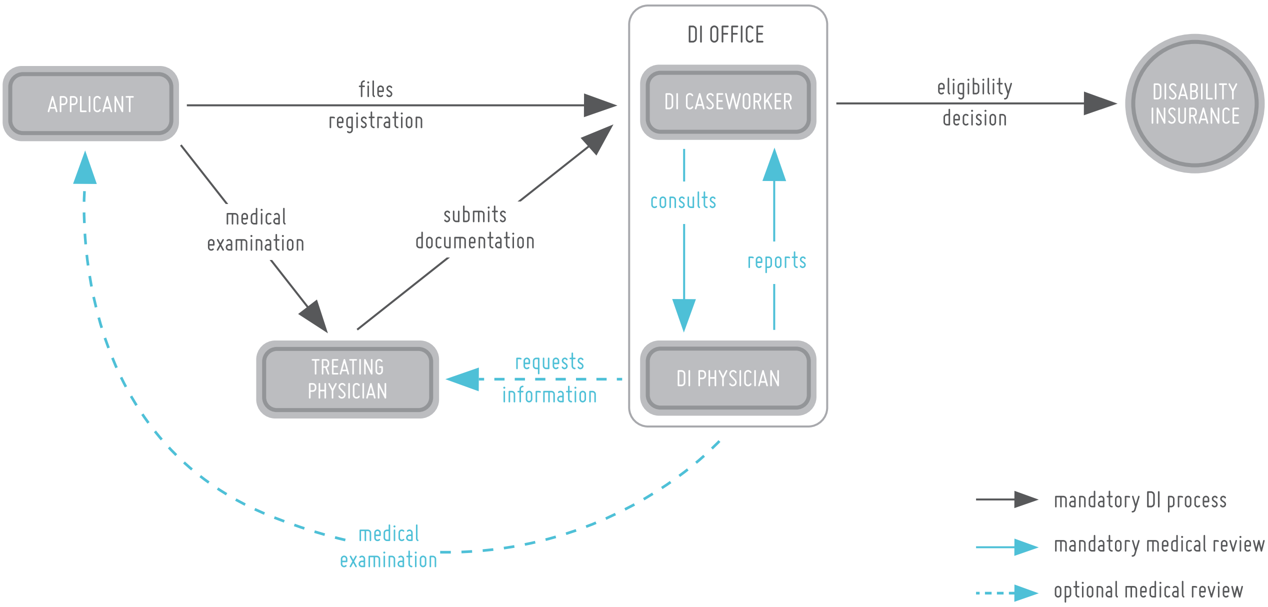
This situation changed with the reform, which essentially strengthened the role of independent DI physicians in the application process. There are three major changes attached to the policy. First, the reform substantially increased the medical staff working for the DI offices. Aggregate figures indicate that the number of full-time equivalent positions increased by 125%. Nationwide, the number of staff positions increased from 105 to 235 due to the reform. Positions are distributed among cantons proportional to the insured population, implying that the relative increase is the same for every region. Pilot cantons experienced this increase three years earlier (see Appendix Figure A2).[2] New physicians are selected to have specialized in fields relevant to diagnose difficult cases (e.g. rheumatology, orthopedics or psychiatry) and are trained in actuarial regulation. Second, medical review became mandatory for DI claims. Every applicants' medical history is reviewed and summarized in a non-technical report for the DI caseworker. Third, physicians were given the authority to screen people in person, to consult with treating physicians 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 DI physician always receives a complete copy of an individual's insurance application, including the medical documentation of potential limitations. The DI physician then provides an evaluation of the applicant's eligibility for the DI caseworker. If the documentation is considered insufficient, additional information can be requested from treating physicians. Furthermore, if the physicians notice inconsistencies in the application or deem it to be invalid, they have the authority to consult with the treating physician, to conduct further examinations or request visits to other clinical specialist.[3] The DI frequently uses the available channels to gather

[2] Since the reform more than doubled the number of physicians working at the DI offices, there is concern about delays in hiring staff and filling positions. However, comparing the average share of vacancies filled in 2006 between offices in pilot and late adopter regions does not indicate that such delays did occur.

[3] Examples for inconsistencies are an applicant claiming benefits on grounds of depression without a

Figure 2: The DI application and decision process



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% (Wapf and Peters 2007).

The DI physicians' eligibility evaluation is not binding. 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 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 (SAKE, *Schweizerische Arbeitskräfteerhebung*) to administrative records. The sample period ranges from 1999–2011. I rely on the SESAM data to analyze the DI hazard because they are the largest representative administrative data source available which combines different social security and labor market registers and has sufficient coverage over time. Given the survey weights, the data is representative of the

sufficiently documented history of therapy or medication, or an individual with moderate chronic pain claiming full work incapacity.

Swiss population.

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 (at most 0.5% per year) and the limited number of individuals that can be tracked over several years, the longitudinal sample dimension cannot be used for the analysis. Instead, the most recent observation for each individual is used (the choice of observation does not influence the results in the paper). This sample restriction results in a large dataset of repeated cross-sections of individual spells. The subsequent analysis relies on the longitudinal information contained within the dates of each spell.

DI receipt is measured in the data with the year in which the individual became eligible for benefits. Since benefits are paid retrospectively, this date usually coincides with the date the claim was filed. DI receipt is only observed for those who receive benefits at the time of sampling. The main outcome in the analysis is DI inflow, measured using the age of disability benefit receipt. The treatment region is defined as the cantons participating in the pilot project and the treatment period comprises the years 2002–2004. 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.

Although the panel dimension in the data does not provide a sufficiently large enough sample to analyze inflow, when pooled across the whole observation period it does provide some insight into the dynamics surrounding the time of DI receipt (see Appendix Table A1). The statistics show that even though individuals reduce their labor supply before filing for benefits, a non-negligible share of beneficiaries continues to work. This reflects the fact that DI insures earnings losses and that the benefit system is graduated. Individuals might still work part-time or be absent from work with a sickness note from a physician. Still, the share of individuals engaging in work drops from 62% to 36% before filing for benefits. About 30% of recipients continue to work two years after filing for benefits. Starting from a lower baseline rate, there is a similarly sharp drop in the share of individuals actively looking for work before filing for benefits. Very few persons look for a job two years after filing for benefits. In addition, almost half of all individuals report having been absent at work despite having a valid contract one year prior to filing for benefits. This figure decreases after filing, suggesting that these individuals either leave employment or find a more stable work arrangement. The income loss that is a requirement for DI eligibility can also be seen in the data. Income one year after filing for benefits is only about one third of the income one year prior to filing. The share of people reporting dismissal from the employment office due to exhausting UI benefits increases one year prior to filing but is generally low, indicating that individuals transition to DI either smoothly from UI or directly from work. Finally, individuals self-reported health declines sharply when filing for benefits. Two years prior to filing, 23% of future beneficiaries report having physical or psychological problems (about twice the unconditional population rate). This share increases to 84% after filing for benefits.

The empirical strategy outlined in section 4 partly relies on a local estimation approach and requires geospatial information to identify municipalities in the vicinity of administrative borders. The SESAM data contain information about individuals' municipality of residence. I augment the 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 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 in municipalities 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 sample.[4] I then compute nearest-neighbor estimation weights for this sample. The unrestricted global sample comprises 259,323 individuals, the local sample is restricted to 133,549 individuals. (descriptive statistics are given in Appendix Table A2, the sample composition is mapped in Appendix Figure A1). In the estimations, I use the survey weights for the global sample and nearest-neighbor weights for the local sample. All results in the paper are robust to the choice of distance measure, variations in the threshold level and whether weights are applied.

As discussed in section 2, medical review also applies to scheduled reassessments of existing beneficiaries' claim status. In the second part of the analysis, I investigate the effects of medical review on existing beneficiaries. For this analysis, I use a second administrative dataset provided by the Swiss Federal Ministry of Social Insurances. I use the data to estimate the effects of medical review on the disability degree classification and benefit payment in the beneficiary stock. Moreover, I rely on this data to investigate potential outflow effects in the beneficiary stock which could confound the main results (see section 4).

The stock data tracks the stock of all existing DI recipients from 2001 onwards. For each individual, I 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] 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 2018).

4 Empirical strategy

In this section, I develop the empirical approach used in the remainder of the paper. Section 4.1 discusses identification and introduces the duration model used in the main analysis. Section 4.2 provides explicit identifying conditions for difference-in-differences in a Cox (1972) proportional hazards model. Section 4.3 discusses potential mechanisms that could violate these conditions and provides evidence to support their validity. Finally, section 4.4 explores and discusses additional identifying conditions which tighten the interpretation of the reduced-form estimate, bounding the effect of medical review on the false positive award error rate.

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 inert 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. The remainder of this section introduces the modeling approach, the following sections present the identifying assumptions and discuss potential threats to their validity.

As Autor and Duggan (2003) illustrate, people rarely transition directly from employment 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. 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 in border regions within commuting distance 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. This weighting approach is equivalent to spatial matching. The main advantage of weighting is that it creates a sample that is well-balanced in observables and increases the credibility of the identifying assumptions introduced in the next section. Weighting reduces the bias of the estimator by restricting comparisons to a more similar control group. The bias reduction potentially comes at the cost of an increase in variance, since the estimator may not use all available data. In the context of matching, this bias-variance trade-off is often favorable, as the gain from finding good matches dominates the loss due to higher variance.

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 individuals 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.[5]

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

[5] All estimates are robust across a large set of bandwidths and whether travel distance or travel time is chosen as the distance metric. Moreover, the results are also robust to replacing (1) with a more flexible specification containing cantonal fixed effects.

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 for Switzerland has shown that DI recipients are loath to give up safe benefits even when faced with strong financial incentives to do so (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 which is available from 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 as the time measure for the analysis 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.[6]

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

[6] Comparisons with aggregate data indicate that the reported aggregate rates are underestimated by about 20% going back five years. Inferring incidence further retrospectively, inferred inflow continues to decrease as attrition caused by moving to the old age pension system and mortality increase. Going back 30 years, inferred incidence converges to zero and is almost exclusively driven by small-sample variation of individuals who were awarded DI when they were very young.

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.

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, 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

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

and

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

The main identifying assumption (5) is that in the absence of mandatory medical review, 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. 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 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 untreated regions to apply for DI) and the timing of applications (early application in anticipation of medical review).

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, before the changes were approved. 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. The person responsible for the yearly committee meeting confirmed that the medical review pilot was never publicly communicated to outsiders. Pilots are published only since 2007, and the medical review pilot was one of the first pilots launched by the ministry. To be certain, I conducted a systematic news search on newspaper databases Factiva, LexisNexis, Pressreader and Swissdox. These do not list a single record 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.[7]

Moreover, considering the one-year earnings loss restriction required for DI eligibility, the time frame until implementation leaves 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

[7] Other reform measures scheduled to come into effect at a later time included the introduction of a three-quarter benefit and the abolishment of additional benefits 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.

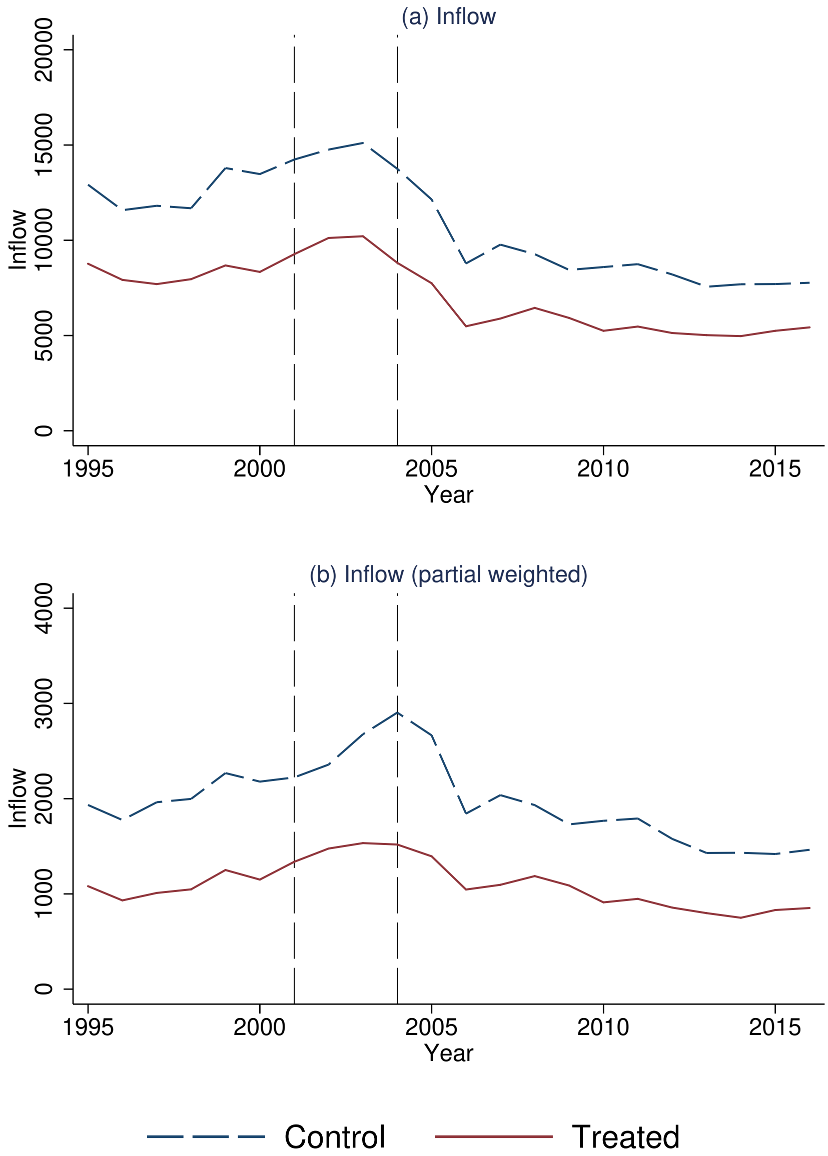
approval in March 2003 and its nationwide implementation in January 2005 that allows for strategic behavior given the one-year restriction.

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 biasing 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.

Regarding the common trend assumption, I first perform a balancing test to ensure that regions are comparable ex ante (Appendix Table A3). Although balance in observables is not strictly required for identification, the common trend assumption is more credible if the comparison regions are similar. This exercise reduces concerns about remaining regional heterogeneity (e.g. due to self-selection) that may induce common trend violations. 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 weighted 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, correspondingly less with secondary and university-level education; and a small difference in the unemployment rate. These remaining differences in observables are small in economic terms and will not affect the estimates unless trends between treatment and control regions differ.

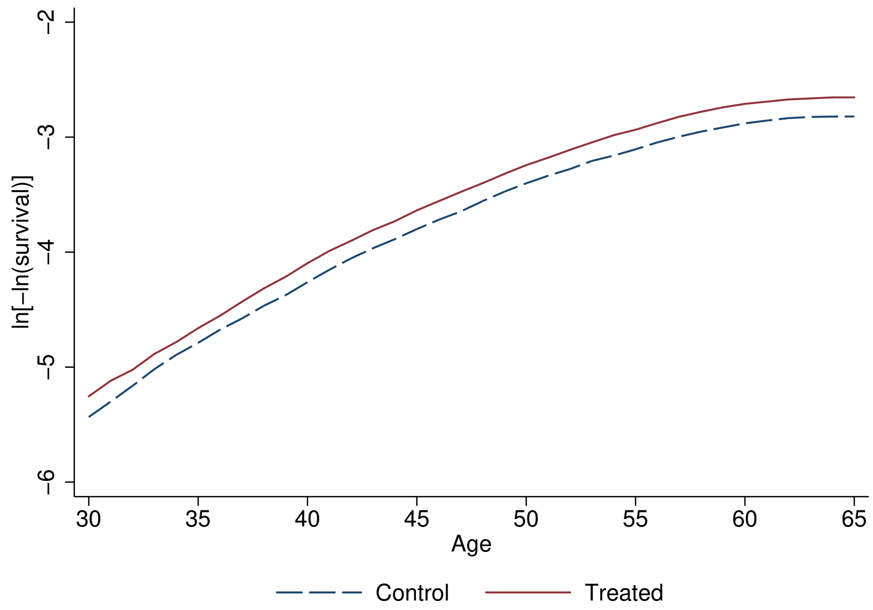
The typical diagnostic graph to inspect the validity of the common trend assumption in difference-in-differences designs are trend plots that show how treated and control units evolve prior to the treatment period. Figure 3, Panel (a) shows trends in the nationwide aggregate DI inflow for all cantons based on statistics published by the Federal Ministry for Social Insurances. Prior to 2002, the number of new DI recipients evolves similarly in treated and control regions. Before the reform, the trends in inflow are comparable across treated and control cantons. During the pilot, the trend breaks and inflow declines in treated regions. This is even more apparent when weighting partial inflow by DI coverage level, indicating that new beneficiaries are downgraded and admitted at lower partial coverage levels (Figure 3, Panel b). After 2005, inflow rates evolve similarly again. These developments are also apparent when looking at trends in the aggregate beneficiary stock (Appendix Figure A3). Throughout the whole time period, applications are largely stable

Figure 3: Trends for aggregate disability insurance inflow



Note: Disability insurance inflow for treated and control regions in panel (a). Insurance inflow partial weighted by pension amount shown in panel (b). Data were provided by the Federal Ministry of Social Insurances.

Figure 4: Log cumulative hazard by age and treatment region



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

and follow the same trend in treated and control regions (Appendix Figure A4). Regarding the treatment, the number of full-time equivalent positions for DI physicians exhibits the expected increase due to the pilot and the nationwide implementation (Appendix Figure A2, Panel a). Correspondingly, the caseload per physician drops substantially (Appendix Figure A2, Panel b). Note that all descriptive graphs are based on data from national statistics or federal social insurance reports and not conditioned on the local sample.

While an indication of comparability between regions, strictly seen, the trend plots in Figure 3 do 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. An alternative test for the common trend assumption are the typical placebo specifications for pre-reform effects (discussed among the robustness checks in section 5.5). These do not indicate that the common trend is violated. Another possibility to investigate the assumption is to look at the log cumulative hazard by age as shown in Figure 4 (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.

In addition, 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 (e.g. Vittinghoff et al. 2011).

Visually assessing the validity of the assumption from the log-minus-log transformation is preferable to comparing survival curves directly, as it is easier to determine whether two curves are apart by a constant difference than to judge whether they are an exponential transformation.[8] The curves in Figure 4 appear parallel and provide no indication that the proportional hazards assumption is violated. The same applies when the cumulative hazard estimate is stratified by five-year birth cohorts as in the analysis (Figure B1, Online Appendix).

Finally, another potential issue pertains to bias incurred by selective sampling due to DI outflow. Previous benefit receipt is not registered in the data—only current benefit receipt at the time of sampling is observed. 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 test for such outflow effects using the stock data and the results do not indicate that outflow is affected. The results for outflow are discussed together with the robustness checks in section 5.5.

4.4 Identification: Bounding the effect of medical review on award errors

The assumptions outlined in section 4.2 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 interpretation of the reduced form effect can be extended. The main estimation results in the paper indicate a reduction in DI inflow. By layering two additional assumption, this reduced form effect can be interpreted as a lower bound of the effect on false positive award errors. Conditioning on individuals' latent eligibility status, the total effect can be decomposed into a mixture of effects on the false positive and false negative DI misclassification rates.

It is the main duty of the insurance office to separate meritorious from non-meritorious claims ('tag' the eligible, Akerlof 1978). Given the null hypothesis of 'no disability', two types of classification errors can occur in this situation: (1) Award errors (type-I, false positive) and (2) Rejection errors (type-II, false negative). 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 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,

[8] Since $h(t|x) = h_0(t)\exp(x\beta)$, the equivalent relation for the survival curve is $S(t|x) = S_0(t)^{\exp(x\beta)}$. Visual inspection requires identifying this exponential relationship. The log-minus-log transformation of the last equation gives $\log(-\log(S(t|x))) = \log(-\log(S_0(t))) + x\beta$, i.e., if the proportional hazards assumption holds, the curves of the treatment groups should be a constant distance apart.

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 relative average treatment effect in (3) by latent eligibility status $E = \Big\{0, 1\Big\}$,

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

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—an inflow reduction, corresponding to a hazard ratio smaller than one—depending on the effect for each type separately. In the following, I simplify notation by omitting the parameters common to all objects in the conditioning set.

Unlike in other treatment effect settings, population shares in (6) are superscripted by the corresponding counterfactual states. In this setting, distinguishing them is sensible as they can be thought of as shares of applications by eligibility types which might be influenced by the treatment. Ruling this out to ease interpretation, assume that

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

This assumption implies continuity in the composition of applications, effectively ruling out that the propensity to apply for DI is influenced by the pilot. The most likely mechanism to confound this assumption is self-screening, i.e., individuals are selectively discouraged from applying for benefits (Parsons 1991). For the reasons outlined in the previous section, this behavior is unlikely since information about the pilot program did not transpire to the public. Parsons's (1991) original paper on the self-screening mechanism is about how changes in screening stringency and administrative hassle which are perfectly observed by applicants influence the application decision. The medical review process is largely hidden to the applicants and there is no information available to them detailing it. This view is also supported by the data. Looking at the limited aggregate data available, application rates evolve similarly across both groups of cantons, are very stable over time and do not diverge during the pilot (Appendix Figure A4). Due to the non-public introduction of medical review and the common trend in applications, differential variations in application behavior are likely to be negligible.

The estimation results in the next section indicate a reduction in inflow. This in mind and simplifying notation due to (7), the hazard ratio must be smaller than one,

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

Rearranging gives

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

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 reduction in award errors (type-I, RHS) must exceed the reduction in rejection errors (type-II, LHS) for the effect to be negative. This is consistent with the interpretation of the effect in (3) as a net effect.

Finally, 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.

This assumption is not directly testable with the available data. It relies on the fact that medical review is an intervention to improve screening quality and, unlike variations in screening stringency, does not involve a trade-off between false positives and false negatives (Parsons 1991, Kleven and Kopczuk 2011, Low and Pistaferri 2015). Alternatively, the condition in (9) is trivially fulfilled if (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.

I will consider the consequences of violations of these assumptions and how they can be relaxed in turn. Assumption (7) posits that medical review does not change the composition of applications. This assumption could be weakened by assuming that medical review decreases the propensity of ineligible types to apply, i.e., $p^1(E=1) \geq p^0(E=1)$.[9] This coincides with Parsons's (1991) empirical result that self-screening is non-perverse. The finding is also confirmed by Low and Pistaferri (2015), who find that false applications decrease with program stringency. Since medical review extracts information, those at the margin of being discouraged from applying are those that are more likely to be found undeserving. The lower bound interpretation can be retained with non-perverse

[9] I am grateful to an anonymous referee for pointing out this possibility.

self-screening, because a larger share of the effect can be attributed to medical review of the eligible (although the bound will be less informative). Alternatively, perverse self-screening attributes a larger share of the effect to ineligible types, the intended target population, complicating the interpretation (but diminishing the importance of assumption 10). Assuming no effect on self-screening is neutral with regard to composition and eases interpretation.

Regarding the implications of assumption (10), there are four different scenarios to consider (illustrated in the effect matrix in Appendix Figure A5). For each eligibility type, medical review can either be perverse (unintended) or non-perverse (intended). Consider first the cases where screening is perverse for the ineligible, i.e., ineligible types are accepted at higher rates under medical review. These cases can be dismissed. First, if eligible types are accepted at higher rates as well (non-perverse), we would not observe a reduction in inflow. Second, if eligible types would be rejected at higher rates due to medical review, this perverse effect on the eligible would have to exceed the perverse effect on the ineligible. This scenario is implausible.

Consider the remaining two cases where medical review for the ineligible is non-perverse, i.e., ineligible types are rejected more due to medical review as intended. In this case, if (10) is violated and medical review is perverse and reduces the chances of the truly eligible indiscriminately, the effect can still be interpreted as an upper bound. Finally, if medical review has the intended effect of being non-perverse and potentially also increases the chances of the eligible to be allowed benefits, the effect can be interpreted as a lower bound. This is the case implied by assumption (10).

The upper bound interpretation is the most likely case in which assumption (10) would be violated. In this case, medical review has a perverse effect and induces an even larger number of false negative errors. One scenario in which this might occur is that if lower DI incidence is politically desired, individual physicians might be pressured to be generally more critical when reviewing new applications due to a fear of being laid off. However, the additional staff at the DI offices were hired on permanent employment contracts and could not have been easily laid off, irrespective of the development of the insurance rolls. In addition, the scope of the federal government to influence local public entities is limited due to the decentralized nature of the Swiss political system. Hiring, medical review and the DI decision are made on the local level, even though staff and DI benefits are paid out of federal funds. The role of the federal government was limited to providing funding for the program. It was generally recognized that the insurance offices' structure, last revised 1973, needed to be overhauled and that they were insufficiently staffed with physicians. The physicians responsible for medical review had the explicit mandate to improve the accuracy of medical diagnoses of functional limitations.[10] As discussed, the institutional

[10] The leading physician in one office was aware of the fact that more intense medical review could increase DI incidence. She stated that in her experience, rejection errors do occur and are sometimes

structure of the DI offices remains unchanged—the final decision to grant benefits still lies with the DI caseworker.

There is also some empirical support for assumption (10). If differential changes in stringency where to occur due to medical review during the pilot, 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, but I collected data on the amount of legal claims for each canton from yearly reports of the cantonal courts. Both the number of total and rejected lawsuits in treated and control regions evolve very similarly over time (Appendix Figure A6). This suggests that it is unlikely that differential changes in stringency occur during the pilot, supporting the lower bound interpretation. Even if the upper bound interpretation applies, given the institutional and empirical evidence it is likely that a non-negligible fraction of the effect can be related to reductions in false positives.

In this section, I have outlined additional assumptions that are required to bound the effect of medical review on award errors and have provided some empirical and institutional support in favor of these assumptions. The reduced-form estimate in sections 4.2 does not permit inference about targeting efficiency of the DI program. Given assumptions (7) and (10) in this section, targeting efficiency improves if medical review reduces inflow. However, this interpretation should be made with utmost care as it is highly dependent on the assumptions that are made. Assuming no change in self-screening as in (7) is neutral with regard to targeting, but directional self-screening would not be. Importantly, constraining the direction of the effect for eligible types as in (10) has direct implications for targeting efficiency. In case this assumption is violated, and rejections are a mix of eligible and ineligible types (the lower left quadrant in Appendix Figure A5), the interpretation ceases to be valid. In this case, rejections of the ineligible have to exceed rejections of the eligible for net targeting to improve, assuming that the same weight is placed on false positive and false negative decision errors, which might not be desirable.

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,

encountered during revisions, but are much less frequent in relation to the amount of award errors uncovered ex post.

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.

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 (corresponding to a 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

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.

health impediments for new awards. I analyse whether medical review changes the relative incidence of partial and full benefit awards. 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 benefit 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 the full benefit amount previously are now downgraded, resulting in a zero net effect for partial DI benefits. This finding is also reflected by a moderate decrease in the aggregate share of full benefit awards—in 2005, 58% of new beneficiaries are awarded full benefits compared to 68% in 2002.

5.2 Incidence of difficult-to-diagnose conditions

The main analysis indicates that DI awards declined substantially due to 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 treating

Table 3: Disability types

	All	Illness	Illness: Psych.	Illness: Nerve	Illness: MSC	Accident	Congenital/ Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
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.

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 conditions (MSC). 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 intensified medical review.

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. 2016). 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 for the full sample show that the share of individuals in registered employment increases. Similarly, the share of individuals with positive (non-benefit) earnings increases as well. In addition, the share of individuals registered with the employment office as job seekers also decreases (columns 1–3). 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 exhausting unemployment the maximum duration for unemployment benefits. Similarly, I find no effect on the receipt of social assistance, 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. The results for the local sample are comparable in sign and magnitude to the estimates for the full sample. However, they are insignificant, most likely due to a lack of power ($p = 0.17$ for the main employment estimate in the local sample).

An explanation for these results is that DI applications are partly made by people capable of gainful employment and driven by moral hazard. One possible mechanism behind this result is the canonical substitution effect interpretation—applicants seek benefits due to a distortion in the relative price of leisure. This distortion is caused by an implicit tax on work due to DI ('cash cliffs'). An alternative explanation is that applications are (partly) due to income effects, i.e., even if work is not implicitly taxed by the DI program, given the transfer payments, beneficiaries may prefer leisure to labor (e.g. Autor and Duggan 2007, Eugster and Deuchert 2017, Gelber et al. 2017). These effects have different welfare implications. If DI reduces labor supply through the substitution effect this implies a deadweight loss, which would be reduced by medical review. Alternatively, medical review would not be welfare improving if all of the labor supply increase is due to a reduced income effect. Since DI is provided (partially) contingent on work, I am unable to separate these effects. Taken together, the evidence from the analysis suggests that distorted incentives are likely to matter in this context.

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

Table 4: Labor market responses to medical review

(a) Full sample

	Working	Positive income	Employment office registration	Employment office dismissal	Social assistance
	(1)	(2)	(3)	(4)	(5)
Treat x pilot	0.009***	0.008**	−0.007***	−0.002	0.000
	(0.004)	(0.003)	(0.002)	(0.001)	(0.001)
Individual covariates	✓	✓	✓	✓	✓
Canton FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
N	556,540	557,270	411,461	411,461	411,461

(b) Local sample (within 20 km)

	Working	Positive income	Employment office registration	Employment office dismissal	Social assistance
	(1)	(2)	(3)	(4)	(5)
Treat x pilot	0.007	0.006	−0.003	0.000	0.001
	(0.005)	(0.005)	(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 clustered at the municipality level given in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.

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.

regularly, revisions seldom resulted in actual disability degree or benefit cuts and typically involved DI caseworkers 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 benefit increases. With the new regime in place, files that are scheduled for review are now also passed to the DI physicians in charge of medical review.

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.[11] Upward revisions are far more common, downward changes only account for 2.3 percentage points. Still, introducing mandatory medical review appears to cause 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 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. Hazard ratio estimates across all specifications are close to one, precisely estimated and insignificant at conventional levels, supporting the validity of the identification strategy (Appendix Table A4). Placebo results for employment also do not indicate any violation of the common trend assumption (Table B3, Online Appendix).

Another potential concern is that the results are sensitive to the choice of distance window. Figure A7 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 at least a 20% reduction in incidence in the treatment group during the pilot program. More detailed estimates over selected distances are provided in Appendix Table A5.

As discussed in section 4.3, potential outflow effects of the reform might confound the main result. Since previous benefit receipt is unobserved, outflow effects would lead to inflow being measured with error in the sample. I use the stock data to test for outflow effects. 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 Appendix Table A6, 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

[11] 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 and are initiated only in extreme cases.

younger and possibly healthier group more likely to exit. All estimates are consistently indistinguishable from zero and precisely estimated.

Another 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. Replacing the main model using treatment group and pilot period interactions with a model including canton and time fixed effects yields very similar treatment effect estimates (Table B2, Online Appendix). Similarly, the results are not dependent on the application of weights, stratification or the stratification level. Moreover, even though the identifying assumptions are different and the analysis is underpowered, relative effects obtained from a linear probability model for inflow in the full sample are similar in magnitude to the relative average treatment effect estimates in Table 1 (Table B1, Online Appendix).

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 stems 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. Appendix Table A7 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 Conclusion

This paper provides a comprehensive evaluation of the introduction of medical review for DI applications in a setting in which treating physician testimony is decisive. The results indicate that external medical review can reduce insurance inflow substantially. The main estimate suggests that medical review reduces DI uptake by 23%. Reductions 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.

Under additional assumptions, the results suggest that medical review is likely to reduce the amount of false positive award errors and that these errors occur frequently in the absence of medical review.

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 A7 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.

Given the substantial present-discounted value of DI benefits, 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. The calculations are based on the observed increase in the number of physicians, a conservative effect estimate and the average benefit amount and remaining spell duration until retirement, assuming rejections are permanent. Based on these parameters, the yearly savings only in the treated regions during the pilot are likely to be above 650 million Swiss Francs (approximately 650 million US$ in 2018). Extending medical review nationwide in 2005 may have saved in excess of 1.2 billion Swiss Francs in that year alone. Even if all rejected applicants never reenter the labor market and immediately receive social assistance, estimated yearly savings for 2005 are upwards of 500 million Swiss Francs. 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 DI benefits 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.

Taken together, the results cast doubt on the practice to assign a large weight to the treating physician's opinion in DI insurance decisions. 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, treating physicians 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).

Since external medical review by DI physicians appears to be effective in the Swiss setting, it might provide a viable policy option for other countries which are burdened by high disability insurance costs and rely on treating physician assessments for DI. However, it is important to bear in mind that prior to the reform, medical review was conducted almost exclusively by treating physicians and DI physicians could not examine patients. Both the policy impact and the size of award errors are likely to depend on the initial level of screening intensity. Still, treating physician testimony is influential for DI determinations in many OECD countries. The results suggest that subjecting treating physcicians' opinions to medical review by a third party is a cost-effective policy to regulate inflow and award errors. Since the policy also lifted bans on personal medical examinations, the changes in Switzerland can potentially also provide some insight about extending medical review in systems which exclusively rely on file-based review.

It is important to note that screening during the pilot does not necessarily 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. This situation might differ if medical review is announced publicly. Since medical review extracts information, it may also discourage ineligible applicants from applying for benefits, as they have higher chances to be ultimately denied. This deterrence effect is found to be pronounced by Low and Pistaferri (2015).

The mechanisms behind the results in this paper merit further investigation. One possible channel behind the incidence reductions are inaccurate diagnoses by treating physicians, the first gatekeeper to the DI system. However, whether and how much application behavior suffers from moral hazard remains ultimately unclear. Applicants could be largely myopic or actively engage in malingering. Still, the overall reduction in inflow provides a tentative suggestion that award errors exceed rejection errors in award decisions. This result diverges from previous analyses for the US. However, given that benefits are substantially more generous in Switzerland, this finding is in line with Low and Pistaferri's (2015) result that false applications are strongly increasing with benefit generosity. Hence, the result is also a first indication that the relative prevalence of errors may be different in European DI systems which offer higher replacement rates. 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

Adam, S., Bozio, A. and Emmerson, C. (2010). Reforming disability insurance in the uk: Evaluation of the pathways to work programme. *Working paper*, Insitute for Fiscal Studies, London.

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. and Rust, J. (2004). How Large are the Classification Errors in the Social Security Disability Award Process? *NBER Working Papers 10219*, National Bureau of Economic Research, Inc.

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.

Borghans, L., Gielen, A. C. and Luttmer, E. F. P. (2014). Social Support Substitution and the Earnings Rebound: Evidence from a Regression Discontinuity in Disability Insurance Reform. *American Economic Journal: Economic Policy* 6(4), 34–70.

Bound, J. (1989). The health and earnings of rejected disability insurance applicants. *Working Paper 2816*, National Bureau of Economic Research.

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.

Chen, S. and van der Klaauw, W. (2008). The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics* 142(2), 757–784.

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

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 Deuchert, E. (2017). Income and substitution effects of a disability insurance reform. *Economics Working Paper Series 1709*, University of St. Gallen, School of Economics and Political Science.

Eugster, B. and Parchet, R. (2018). Culture and taxes. *Journal of Political Economy* (forthcoming).

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.

French, E. and Song, J. (2014). The effect of disability insurance receipt on labor supply. *American Economic Journal: Economic Policy* 6(2), 291–337.

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.

Garcia Mandico, S., Garcia-Gomez, P., Gielen, A. and O'Donnell, O. (2018). Earnings responses to disability benefit cuts. *Working paper*, Tinbergen Institute.

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.

Inderbitzin, L., Staubli, S. and Zweimüller, J. (2016). Extended unemployment benefits and early retirement: Program complementarity and program substitution. *American Economic Journal: Economic Policy* 8(1), 253–288.

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.

Karlström, A., Palme, M. and Svensson, I. (2008). The employment effect of stricter rules for eligibility for DI: Evidence from a natural experiment in Sweden. *Journal of Public Economics* 92(10–11), 2071–2082.

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.

Kornfeld, R. and Rupp, K. (2000). The net effects of the project network return-to-work case management experiment on participant earnings, benefit receipt, and other outcomes. *Social Security Bulletin* 63(1), 12–33.

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.

Low, H. and Pistaferri, L. (2015). Disability insurance and the dynamics of the incentive insurance trade-off. *American Economic Review* 105(10), 2986–3029.

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.

Moore, T. J. (2015). The employment effects of terminating disability benefits. *Journal of Public Economics* 124, 30–43.

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 (2003). *Transforming Disability into Ability.* Paris, OECD Publishing.

OECD (2006). *Sickness, Disability and Work: Breaking the Barriers—Norway, Poland and Switzerland, Vol. 1.* Paris, OECD Publishing.

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—A Synthesis of Findings across OECD countries.* 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.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C. and McCulloch, C. E. (2011). *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer Science & Business Media.

von Wachter, T., Song, J. and Manchester, J. (2011). Trends in employment and earnings of allowed and rejected applicants to the social security disability insurance program. *American Economic Review* 101(7), 3308–29.

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 A: Additional tables and figures

Table A1: DI recipients before and after filing for benefits

	-2	*-1*	*DI filing year*	*+1*	*+2*
Worked last week	0.761	0.622	0.368	0.320	0.296
	(197)	(410)	(810)	(1268)	(1457)
Looking for work last month	0.484	0.333	0.120	0.096	0.079
	(31)	(84)	(357)	(748)	(953)
Work contract but absent at work last week	0.326	0.455	0.294	0.120	0.057
	(46)	(154)	(506)	(845)	(1006)
Yearly income (1k CHF)	53.113	47.785	33.262	17.284	11.574
	(197)	(410)	(810)	(1268)	(1457)
Dismissed from unemployment office	0.025	0.053	0.061	0.056	0.066
	(79)	(206)	(445)	(784)	(948)
Social assistance	0.051	0.058	0.038	0.051	0.044
	(79)	(206)	(445)	(784)	(948)
Age	47.365	48.480	50.022	50.445	50.648
	(197)	(410)	(810)	(1268)	(1457)
Mental or physical problem	0.234	0.393	0.688	0.844	0.836
	(124)	(262)	(523)	(841)	(980)
Accident within the last 12 months	0.208	0.176	0.118	0.057	0.082
	(48)	(85)	(136)	(174)	(184)

Note: This table shows the mean values of selected variables for DI recipients from two years prior to filing the application until two years afterwards. The table utilizes the limited longitudinal information that is available in the SESAM data. The number of observations in a cell is given in parentheses. Note that sample sizes vary because not all recipients have the same historic coverage and not all survey modules are administered every year.

Table A2: Descriptive statistics

(a) Full sample

	Mean	SD	Min	Max	N
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 A3: 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.

Table A4: 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 A5: 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.

Table A6: Stock outflow

	(a) All individuals			(b) Age ≤ 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

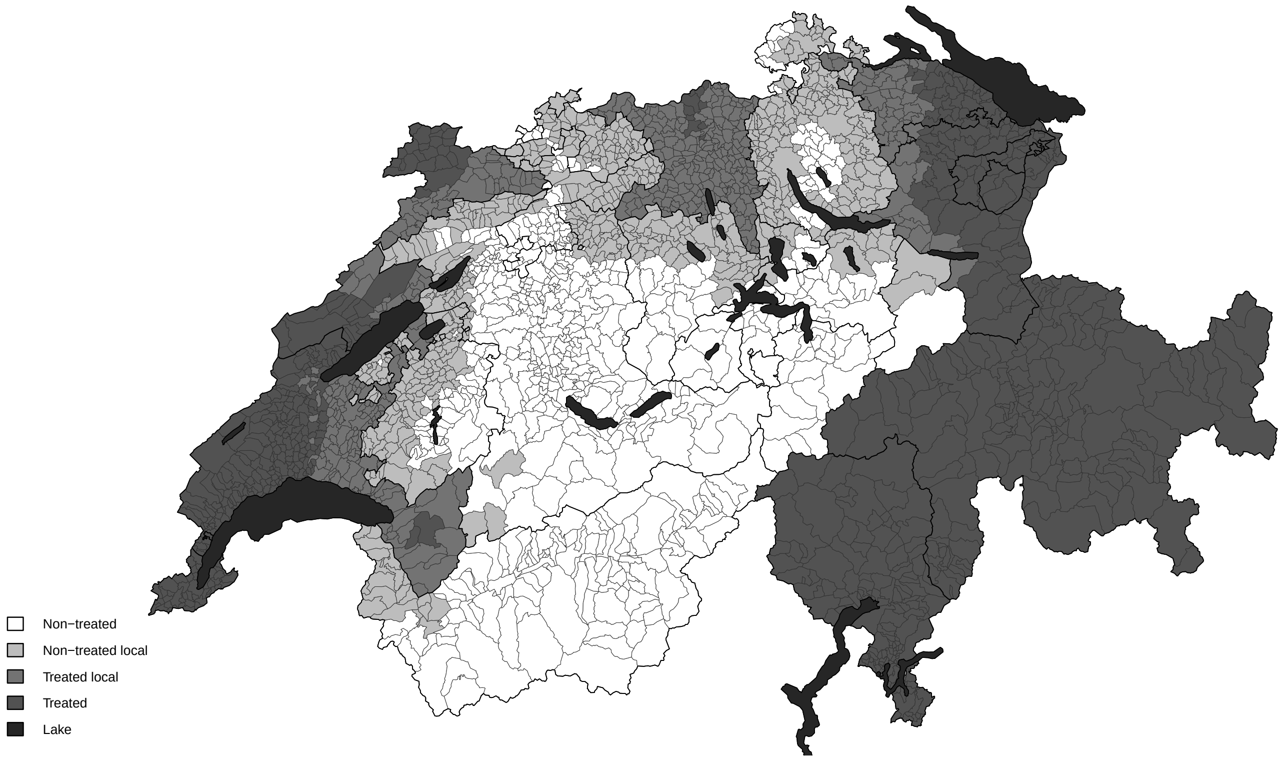
Note: Cox Proportional Hazard estimates for individuals in treated and control regions based on SESAM individual-level survey and administrative data sampled during 1999–2011. Baseline hazard for all regressions stratified by 5-year birth cohorts. Survey weights applied for the full sample. Observations 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 A7: 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

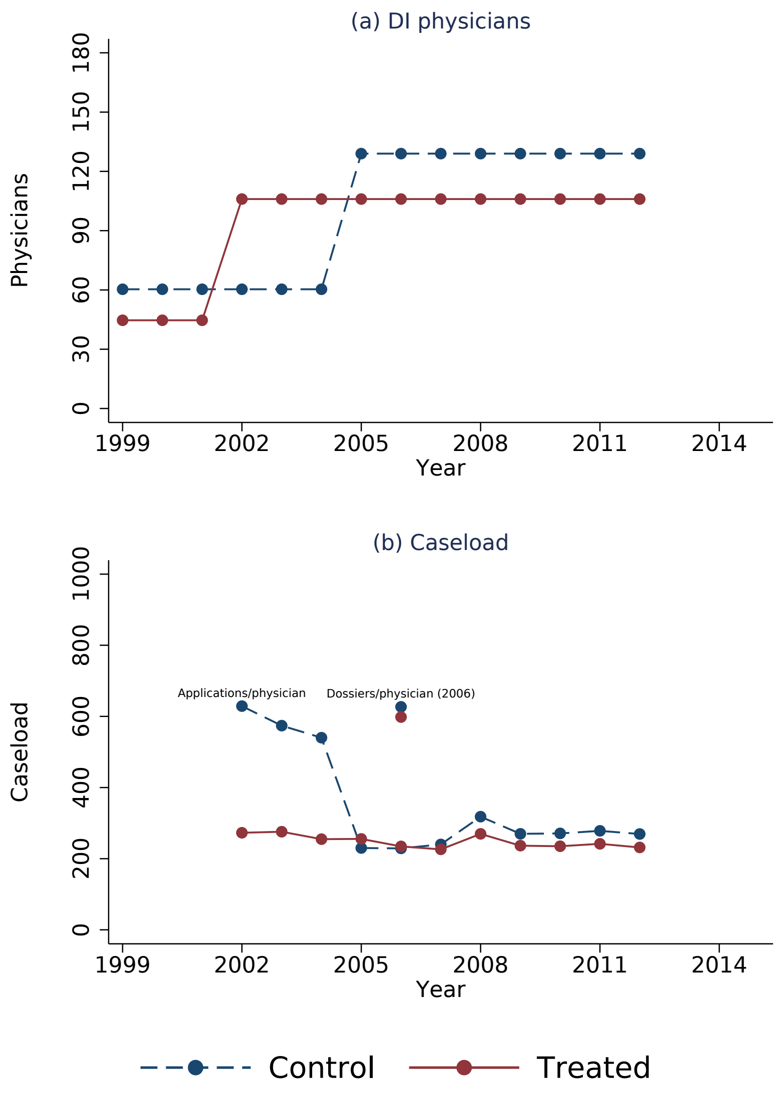
Note: Probit estimates for the probability to be included in the local sample separately for treated and control regions. Marginal effects at the mean reported. Standard errors clustered at the municipality level in parentheses, number of observations given below. *, ** and *** denote significance at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

Figure A1: Sample composition



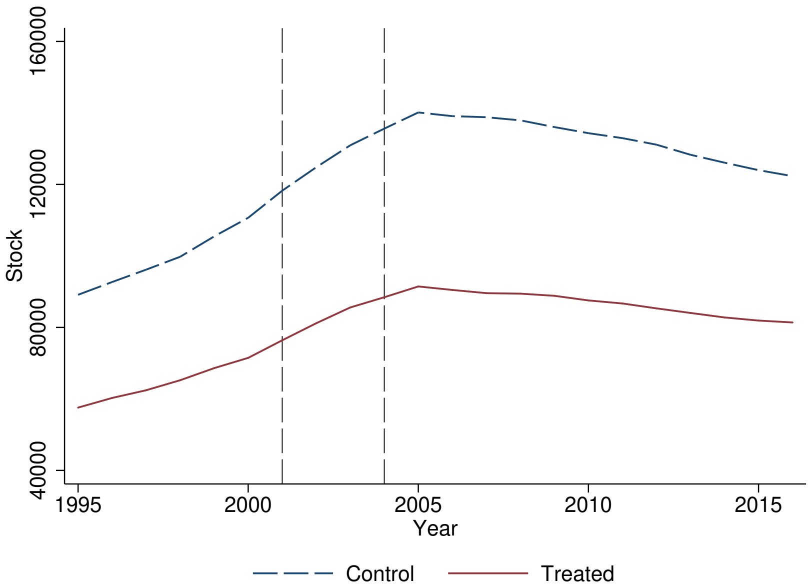
Note: Pilot cantons in shaded in dark and medium grey, control cantons shaded in light grey and white. Intermediate shades indicate the municipalities that are included in the local sample. Lakes shown in black.

Figure A2: Trends in DI physicians and caseload



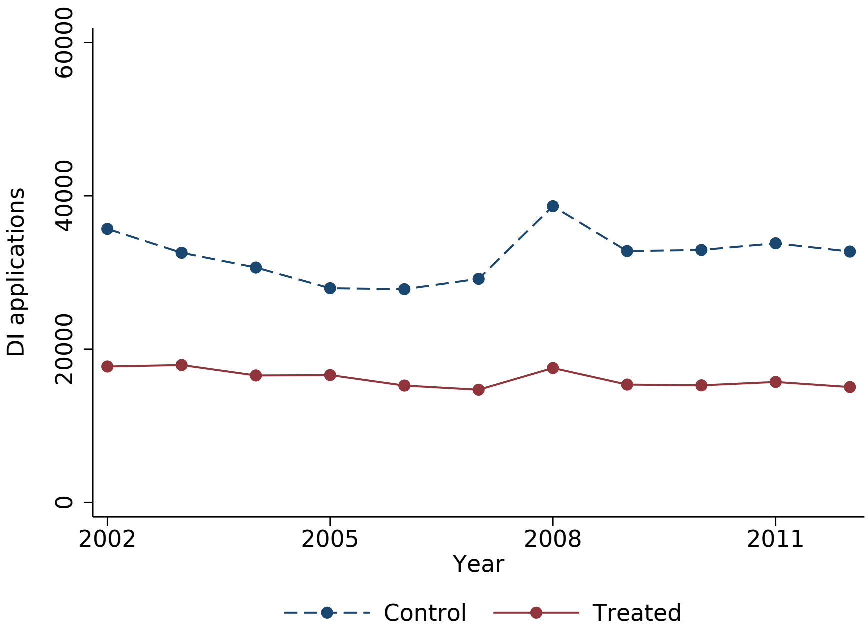
Note: Panel (a) shows the number of full-time equivalent medical staff positions before and after the reform changes, panel (b) approximates the application caseload per physician. Cantons in western Switzerland for which the electronic reporting system is known to have been faulty are omitted from the sample for the statistics in panel (b) (Fribourg, Genève, Jura, Neuchâtel, Vaud). Applications are only available from 2002.

Figure A3: Trends for aggregate disability insurance stock



Note: Disability insurance stock for treated and control regions.

Figure A4: Trends for disability insurance applications

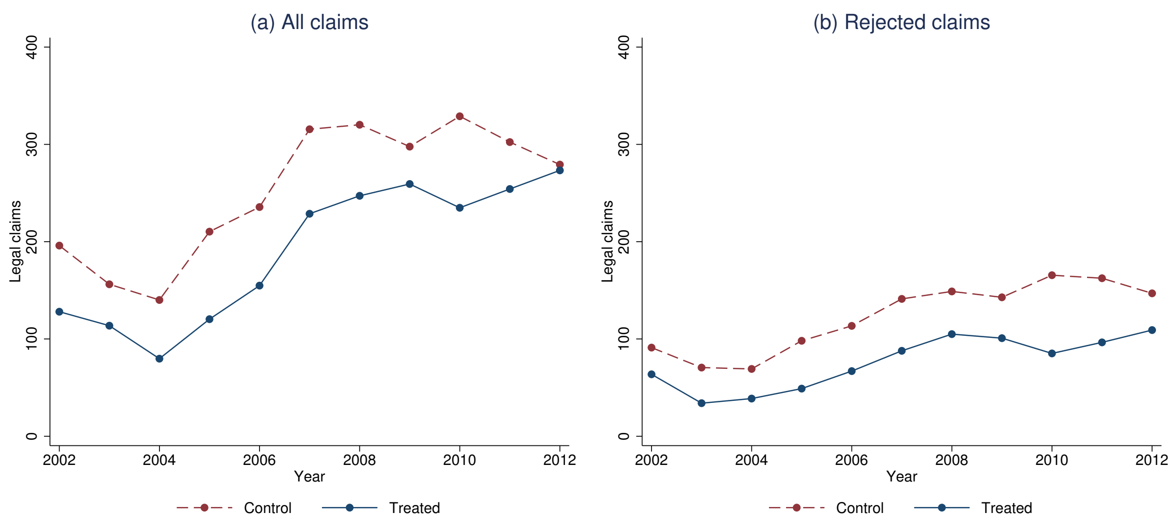


Note: Disability insurance applications for the years 2002–2012. Cantons in western Switzerland for which the electronic reporting system is known to have been faulty are omitted from the sample (Fribourg, Genève, Jura, Neuchâtel, Vaud).

Figure A5: Possible effects of medical review on latent types

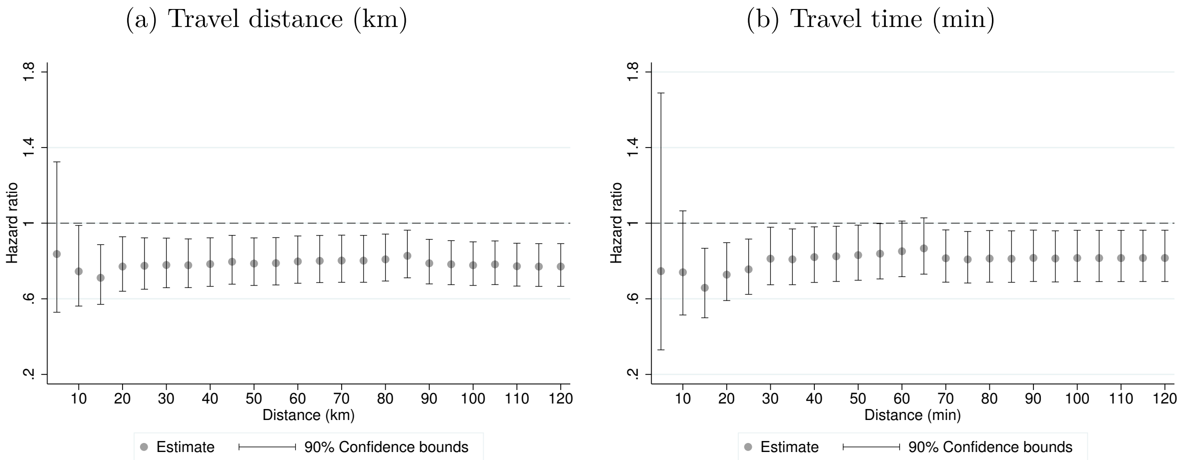
Type		*Eligible* $E = 1$		
	Effect of medical review	*perverse* $h^1(t\|E=1) < h^0(t\|E=1)$	*non-perverse* $h^1(t\|E=1) \geq h^0(t\|E=1)$	
Ineligible $E = 0$	*perverse* $h^1(t\|E=0) > h^0(t\|E=0)$	E=1 rejected more E=0 accepted more Increase in false rejections must exceed increase in false acceptances. Implausible.	E=1 accepted more E=0 accepted more Indiscriminate acceptances. Incompatible with lower inflow.	Irreconcilable with effect.
	non-perverse $h^1(t\|E=0) \leq h^0(t\|E=0)$	E=1 rejected more E=0 rejected more Indiscriminate rejections, reduced inflow is a mix of both types. Upper bound of effect on award errors identified.	E=1 accepted more E=0 rejected more Reduced inflow by ineligible types may be partially offset by higher inflow of eligible types. Lower bound of effect on award errors identified.	
		Potentially ruled out by assuming *MTR* for $E = 1$.		

Figure A6: Disability insurance court cases



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

Figure A7: 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.

Supplementary material for online publication only

Appendix B: Further tables and figures

Title: Does external medical review reduce disability insurance inflow?
Author: Helge Liebert

Table B1: Main disability incidence results, linear probability model

	(a) Full sample		(b) Local sample (within 20 km)	
	(1)	(2)	(3)	(4)
Treat x pilot	−0.000265	−0.000272	−0.000106	−0.000107
	(0.000519)	(0.000518)	(0.001146)	(0.001144)
relative ATT (implied)	-0.1698	-0.1742	-0.0696	-0.0701
$\bar{y}$	0.001559	0.001559	0.001530	0.001530
Other controls	-	✓	-	✓
Canton fixed effects	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓
N	592491	592491	299545	299545

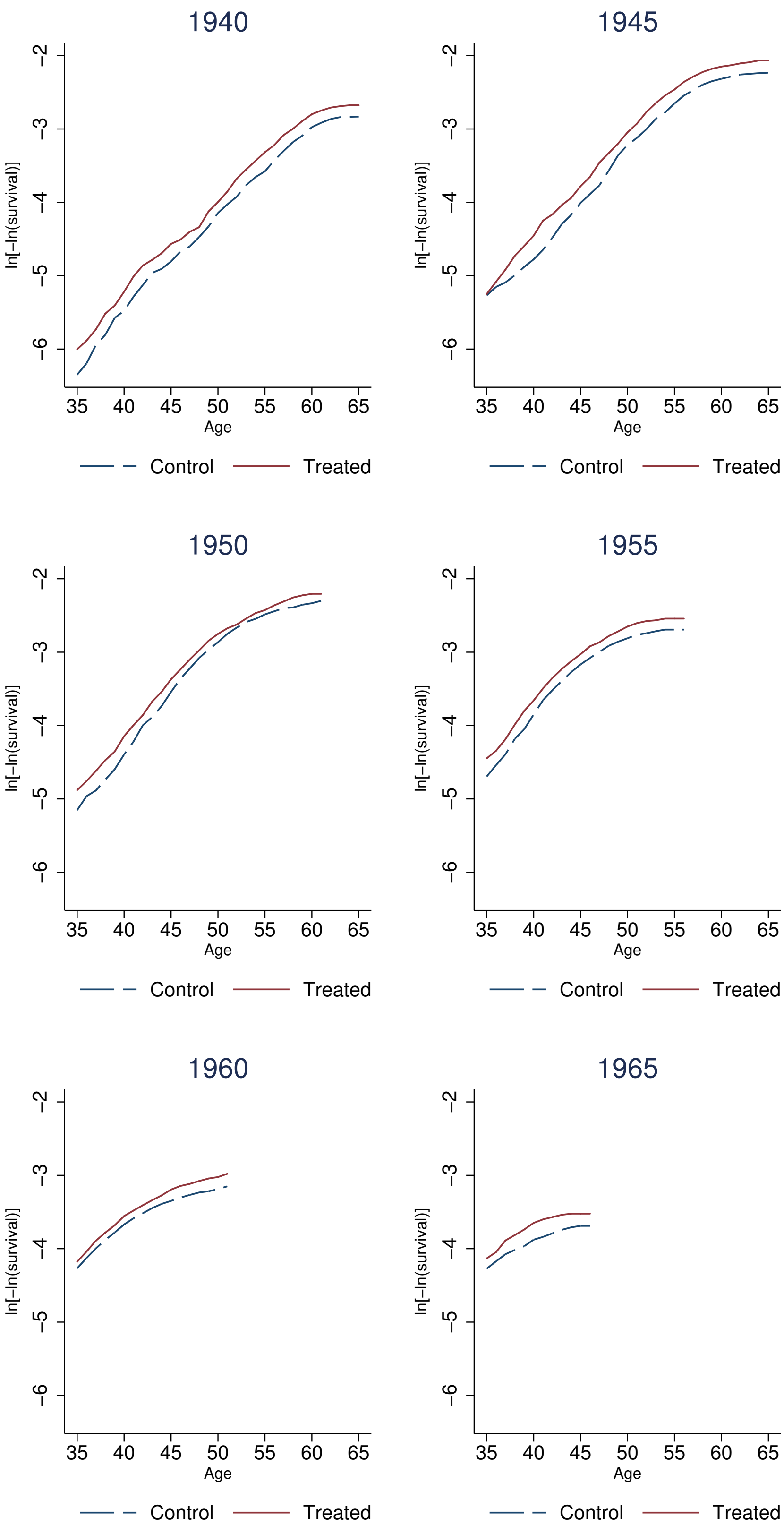
Note: Linear probability model estimates of DI receipt for individuals in treated and control regions based on SESAM individual-level survey and administrative data sampled during 1999–2011. Estimations separately for a complete representative sample of the Swiss population and only for individuals in the vicinity of the border between treated and non-treated regions. Standard errors clustered at the cantonal level in parentheses.

Table B2: Main disability incidence results with canton fixed effects

	(a) Full sample			(b) Local sample (within 20 km)		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat x pilot	0.857**	0.855**	0.859*	0.765**	0.765**	0.760**
	(0.067)	(0.067)	(0.068)	(0.086)	(0.087)	(0.086)
Canton fixed effects	✓	✓	✓	✓	✓	✓
Time fixed effects	✓	✓	✓	✓	✓	✓
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.

Figure B1: Log cumulative hazard by age, treatment region and birth cohort strata



Note: Log-log plot showing log cumulative hazard estimates by age and birthcohort for individuals in treated and control regions, separately for major birth cohort strata.

Table B3: Robustness: Placebo test labor market participation

	(a) Full sample				(b) Local sample (within 20 km)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat x pre	0.0107		0.0004		0.0127		0.0004	
(2001)	(0.0076)		(0.0063)		(0.0109)		(0.0090)	
Treat x pre		0.0045		−0.0055		0.0055		−0.0064
(2000, 2001)		(0.0077)		(0.0046)		(0.0112)		(0.0066)
Treat x pilot			0.0091***	0.0087***			0.0068*	0.0063*
			(0.0026)	(0.0026)			(0.0038)	(0.0038)
Individual covariates	✓	✓	✓	✓	✓	✓	✓	✓
Canton FE	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓
Only years before 2002	✓	✓			✓	✓		
All years			✓	✓			✓	✓
N	52,016	52,016	556,540	556,540	27,887	27,887	282,858	282,858

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. Robust standard errors given in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% level respectively.