Journal of Public Economics 96 (2012) 959–972

Contents lists available at SciVerse ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

# The case for presenteeism — Evidence from Norway's sickness insurance program[☆]

Simen Markussen [a], Arnstein Mykletun [b], Knut Røed [a,*]

[a] The Ragnar Frisch Centre for Economic Research, Norway
[b] The Norwegian Institute of Public Health, Division of Mental Health, Norway

ARTICLE INFO

*Article history:*
Received 3 September 2011
Received in revised form 7 June 2012
Accepted 13 August 2012
Available online 24 August 2012

*JEL classification:*
C26
I18
I38
J48

*Keywords:*
Absenteeism
Disability
Activation
Workfare

ABSTRACT

Can a work-first strategy control moral hazard problems in temporary disability insurance, and accelerate recovery? Based on empirical analysis of Norwegian data, we show that it can. Activation requirements not only bring down benefit claims, they also reduce the likelihood that long-term sickness absence leads to inactivity. Our findings show that absentees who are assigned graded (partial) absence certificates by their physician have shorter absences and higher subsequent employment rates than they would have had on regular sick leave. We conclude that the activation strategies that in recent years have permeated European and US welfare policy may fruitfully be carried over to sick leave insurance for temporary disabled workers.


## 1. Introduction

Rising disability insurance rolls have become a major policy challenge in many industrialized countries; see, e.g., OECD (2003), Duggan and Imberman (2006), and Burkhauser and Daly (2011). On average, OECD countries spend two percent of GDP on temporary and permanent disability benefits — almost three times as much as what was spent on unemployment benefits prior to the recent jobs crisis (OECD, 2010, p. 57). In the present paper, we examine one of the most important public disability programs in European countries, namely insurance against income losses due to sickness absence from work. On a typical working day, around three percent of European employees are absent due to sickness (Bonato and Lusinyan, 2007). And empirical labor market research shows that frequent and/or longer term absence spells significantly reduce subsequent employment and earnings prospects (Hansen, 2000; Ichino and Moretti, 2009; Markussen, 2012), and thus potentially also raise the probability that a worker becomes inactive and dependent on permanent disability insurance. The question we ask in this paper is whether an activation strategy aimed at exploiting the remaining (partial) work capacity of sick pay claimants can reduce absenteeism and subsequent social insurance dependency, and promote self-sufficiency.

Public insurance against income losses during absences caused by temporary disability involves a well-known moral hazard problem: Heavily insured workers tend to be absent too often; see, e.g., Henrekson and Persson (2004) and Johansson and Palme (2005). And firms that can pass their insurance costs on to the public purse exert too little effort to prevent it; see OECD (2010) and Fevang et al. (2011). This is obviously costly for those who pay the insurance premium (typically the taxpayers). In addition, it potentially involves large costs for the absent workers in the form of slower recovery and ensuing earnings losses. Recent medical research indicates that for the illnesses responsible for the vast majority of sick leave days in advanced economies – such as musculoskeletal pain and common mental disorders – regular activity through work helps promote recovery and rehabilitation; see, e.g., Waddell (2004), Waddell and Burton (2006), and OECD (2008).

To the extent that the level of absenteeism is considered to exceed its socially optimal level, possible remedies would be to tighten eligibility regulation, to cut the level of sickness benefits and/or to hold employers accountable for a larger share of insurance costs (e.g. through

☆ This research has received support from the Norwegian Research Council (grant nos. 185201 and 187924). Thanks to Bernt Bratsberg, Stein Atle Lie, Oddbjørn Raaum, two anonymous referees and the Editor for valuable comments, and to Simen Gaure for computational assistance. Thanks also to seminar participants at the University of Utrecht, the Norwegian school of economics and business administration (NHH), the University of Bergen, the University of Oslo, the University of Wurzburg, the University of Erlangen-Nuremberg, the University of Mannheim, and the ASSA 2011 annual meeting in Denver.
* Corresponding author.
*E-mail address:* knut.roed@frisch.uio.no (K. Røed).

0047-2727/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jpubeco.2012.08.008

some form of experience rating). But these options may be blocked, either by competing political priorities or by legally protected contractual obligations. In this paper we show that there is yet another way toward both lower absenteeism and less labor market exclusion, namely to impose *activity requirements* in the temporary disability (sickness) insurance system. Most sick leave days are caused by non-communicable illnesses for which it is far from obvious that 100% absence from work is the appropriate treatment. In particular, we argue that long-term sickness episodes rarely justify complete inactivity. Sickness normally reduces an individual's work-capacity, but it rarely eliminates it. This has motivated some countries to introduce *graded (partial)* sickness insurance arrangements, i.e. insurance that covers the loss arising from reduced productivity or work-hours due to a temporary disability, while requiring the worker to exploit his/her remaining work-capacity. In practice, this is achieved by instructing the physicians, whose certification is required to be eligible for sick-pay, to assess and report the *fraction* of the work capacity that is lost due to the illness. Graded sick leave then implies, for example, that if a worker's work-capacity is considered to be temporarily reduced by 50% due to an illness, he/she is obliged to work at 50% capacity and entitled to the normal wage for this part. Sick-pay applies for the remaining 50%.

Graded sickness insurance thus promotes *presenteeism*, i.e., that workers are present at their workplace even when they are sick, but of course only when the illness is non-infectious and otherwise compatible with work. This idea represents a significant extension of the *activation strategy* that has dominated both European and US welfare and unemployment insurance policies in recent years. With some variations, the use of graded sickness insurance has recently been strongly promoted in the Nordic countries (Kausto et al., 2008), and it has also been introduced in the UK in the form of a so-called “fit note”.[1]

We provide empirical evidence from Norway suggesting that for temporary disabled workers on long-term sick leave (more than 8 weeks) activity requirements not only reduce absenteeism and social security benefit claims, but also significantly improve subsequent employment prospects. The evidence is based on an instrumental variable approach where we use the observed variation in primary care physicians' tendency to grade other patients' absence certificates – controlled for the composition of their clients – as the instrument. Intuitively, we aim to compare employment and other outcomes for sickness insurance recipients paired with physicians who frequently grade their absence certificates with their observably similar counterparts paired with physicians who rarely grade their certificates. The validity of the instrument is examined in detail. We evaluate the consequences of incorporating different groups of control variables, including patient characteristics, medical diagnosis, neighborhood-fixed-effects, and physician characteristics. The latter includes the physicians' workload, their market situation, and proxies for their overall propensity to issue absence certificates (leniency) and the quality of their medical advices. To assess the model's reliability, we also estimate it on outcomes for which we should not expect any causal effects, i.e., *past* absence, employment, and social security dependency. In addition, we assess the potential distortions arising from endogenous family physician selection by looking at the pattern of physician changes just before and during absence spells and by estimating the models based on previous patient–physician-linkages.

The key finding of our paper is that the use of graded rather than non-graded sickness absence certificates reduces the length and volume of long-term absence spells, and significantly improves the likelihood that the absentees are employed in subsequent years. The effects are large, both from an economic and a clinical perspective. Our most conservative instrumental variable estimates indicate that switching from a non-graded to a graded absence certificate before the 8th week of absence reduces the length of the absence spell by more than 90 fulltime-equivalent days and also reduces social insurance claims the next two years – in terms of, e.g., new sickness or disability benefits – by around 80–90 days. Even more importantly, it raises employment propensity two years after by around 16 percentage points. Our findings thus indicate that the introduction of activation requirements in temporary disability insurance schemes is a promising strategy toward reducing sick pay costs and combating labor market exclusion.

## 2. Existing literature

Our paper relates to a large literature on the impacts of workplace-based return-to-work (RTW) interventions essentially showing that workplace interventions tend to improve sick-listed workers' chances for returning to work; see e.g., Franche et al. (2005) and van Oostrom et al. (2009) for recent reviews. However, these interventions typically entail treatment far beyond the “prescription” of return to (some) work, e.g., in the form of physiotherapy, cognitive–behavioral interventions, organizational changes, etc. They are also typically targeted at workers with particular diagnoses, such as musculoskeletal disorders. Hence it is difficult to draw general conclusions regarding the isolated impacts of graded versus non-graded absence certification. There is some corroborating evidence from clinical trials showing that the recovery prospects of back pain patients and individuals with light mental disorders may be enhanced by continuation of “normal activities” even without additional treatments. Malmivaara et al. (1995), for example, conducted a controlled trial among Finnish employees with acute nonspecific low back pain. The patients were randomly assigned to one of three treatments: bed rest for two days, back-mobilizing exercises, or continuation of ordinary activities “as tolerated”. It turned out that the latter ordinary-activity-group had a significantly faster recovery than the other two. A Norwegian randomized controlled trail of long-term absentees with low back pain also found that advice to stay active was associated with better prognosis for return to work during three years of follow-up than “treatment as usual” in primary health care (Hagen et al., 2003). A recent literature review conducted for the UK Department of Work and Pension concluded that work for sick and disabled people is therapeutic and leads to better health outcomes (Waddell and Burton, 2006).

Our paper also relates to a literature examining remaining work capacity among disability insurance applicants by looking at the subsequent employment behavior among those whose applications were rejected. While a typical finding in this literature has been that the employment rate of rejected older disability insurance applicants is relatively low, see, e.g. Bound (1989) and Bound et al. (2003), recent research focusing on younger rejected applicants indicate that the residual work capacity in this group may be significant, although those who do return to work after rejection experience substantial earnings losses when compared to similar non-applicants; see French and Song (2009), Maestas et al. (2011), and Von Wachter et al. (2011). Evidence based on evaluations of the radical Dutch disability policy reforms – with tighter gate keeping and stronger financial incentives aimed at ensuring the use of residual work capacity – also indicate that the work capacity among potential disability program participants is substantial; see Burkhauser and Daly (2011).

We are aware of two published studies that directly examine the impact of graded sick leave certificates on sick leave duration. The first is Høgelund et al. (2010), who based on Danish administrative register data, estimate a proportional hazard rate model and use the timing-of-events approach (Abbring and Van den Berg, 2003) to identify the effect of interest. Their findings suggest that a graded instead of a non-graded absence certificate raises the weekly probability

[1] The UK *fit note* (Statement of Fitness for Work) was proposed in 2008 and implemented in April 2010. In the fit note, physicians are requested to certify whether a sick worker is unfit or (potentially) fit for work. In the latter case doctors may recommend reduced hours or duties, and provide recommendation to employers on how they can help the worker back to ordinary work.

of returning to regular hours by as much as 50%, ceteris paribus. The second is Viikari-Juntura et al. (2012), who perform a randomized controlled trial to investigate the impact of graded versus full-time absence for absences caused by musculoskeletal disorder. The experiment was conducted in six Finnish enterprises and encompassed 63 workers who were unable to perform their regular duties and who were randomly allocated to graded or non-graded sick leave. The findings indicate that grading caused a 60% rise in the hazard rate to regular work activities, and also a 20% reduction in subsequent absenteeism during a one-year follow-up period. Given the small scale of the experiment, however, these estimates are subject to large statistical uncertainty. There is also a number of empirical studies on attitudes toward graded sick leave and of self-reported patient experiences. Most of these have been released in the form of non-peer-reviewed reports and working papers. According to a recent review (Kausto et al., 2008) the results mainly indicate positive attitudes toward graded sick leave, among employees, employers, physicians, and social security administrators. One (published) study examining the subjective views of sick-listed workers in Sweden (Sieurin et al., 2009), for example, reports that 92% of the workers on graded absence, and 63% of the workers on non-graded absence consider graded absence to be (potentially) "good for me".

To our knowledge, no empirical evidence exists regarding the wider impacts of substituting graded for non-graded sick leave certificates, e.g., in terms of subsequent employment and social security dependency. And the few studies that have examined the impacts on the narrower outcome of absence duration are either based on a non-justified proportional hazards assumption (Høgelund et al., 2010) or on a very low number of observations (Viikari-Juntura et al., 2012). The present paper adds to the existing literature by offering the first comprehensive evaluation of implementing an activation strategy in the sick leave insurance system and by introducing a more powerful and reliable strategy for identification of causal effects, based on the variation in "grading-practice" across primary care physicians. Similar identification strategies have previously been used to examine other research questions. Duggan (2005) examines the impact of new/expensive antipsychotic drugs on subsequent health care spending, and instruments each patient's drugs use by his/her psychiatrist's propensity to prescribe expensive drugs. Doyle (2008) investigates the impact of placing a child in foster care on later criminal behavior, and instruments the placement decisions by the observed placement frequencies of child protection investigators. And French and Song (2009) and Maestas et al. (2011) investigate the impact of disability insurance receipt on labor supply, and instruments the disability allowance decision by the examiners' observed allowance rates.

## 3. Institutional setting

Norwegian workers are entitled to a 100% replacement ratio from the first day of sick leave and up to one year. The first 16 days are paid for by the employer, the remaining days are paid for by the social security administration. The only limitation is that a general practitioner (GP) must certify the existence of a genuine disability for absence spells exceeding 3 days (8 days in some firms). During periods of sickness absence, Norwegian workers enjoy a special protection against dismissals, implying that they cannot be dismissed on grounds that are related to their sickness.[2] The moral hazard problems are fairly obvious in this case. Workers have incentives to be absent more than necessary. Firms have incentives to make efforts to prevent short-term absence (since they cover the full costs during the first 16 days), but not necessarily to reduce long-term absence. Indeed, it is typically more beneficial for the firm that a long-term absentee continues to be absent than that he/she returns to work with a high risk of again becoming sick; see Fevang et al. (2011). Moreover, given the level of employment protection in Norway, firms may sometimes find it convenient to pass the costs of temporary redundant labor on to the social security administration. And after one year of sick leave, continued absence becomes a legitimate cause for dismissal.

Norway also has a high level of absenteeism. On a typical working day, around 7% of all workers are absent due to sickness. Long-term sick leave entails a high risk of labor market exit and continued social security dependency. Among those who exhaust their sick-pay entitlements, around 65% move on to other disability programs (medical or vocational rehabilitation), typically with a replacement ratio around 66%; and 3 years after exhaustion, 30% have become permanently disabled.[3] More than 20% of the working age population in Norway is now dependent on a health-related social security transfer (Bratsberg et al., 2010). And in the National Budget for 2011, public insurance payments for sickness absence and disability were projected to account for 5.1% of GDP.

Despite the high level of absenteeism, *all* political parties in Norway, as well as the associations of employers and employees, agree that the existing replacement ratio should be maintained, and that no additional costs should be passed on to firms with absent workers. Instead, policy makers have chosen to focus on "softer" measures, such as more intensive use of graded absence certificates, public information campaigns, and support for improvements in workplace environments. Given that almost 90% of all absence days in Norway are certified by a physician, certification practices have received considerable attention. In 2001, a tripartite "inclusive workplace agreement" (IWA) was made between the state and the associations of employers and employees, in which a target of 20% reduction in absenteeism within four years was set. The agreed strategy included the encouragement of substituting graded for non-graded sick leave certificates, particularly for long-term absences.[4] During the sickness period, the employer is obliged to facilitate modified work within reasonable limits, while the employee is – if necessary – obliged to accept changes in regular duties/tasks.[5]

Graded sick leave normally implies part-time absence (reduced work hours), but in principle it can be implemented in the form of less productive work (with unchanged hours) also. Given the costs associated with implementing workplace adaptations, it is not intended for very short sickness absence spells; hence short-term absence certificates are typically not graded. If the physician expects an absence spell to be long-lasting, it can nevertheless be graded from the start. And for non-graded absence certificates covering sick leaves beyond 8 weeks, the physician is obliged to explain to the social security administration why grading cannot be used. Typical explanations are that it is difficult to implement the required workplace adaptations or that the sick leave spell is expected to end soon anyway. Absence certificates must in any case be renewed at this point, based on a new (and supposedly profound) consultation. It is the grading decisions taken before or at these 8-week consultations that we focus on in this paper.

There are three ways in which grading is intended to affect remaining absence duration and subsequent outcomes (Mykletun et al., 2010): First, it is aimed at pushing unmotivated workers back to work, even when health problems prevent them for working at full capacity, thereby containing moral hazard problems among employees.

[2] The burden of proof lies with the firm. In practice, this implies that absent workers can only be laid off as part of a mass displacement. After the one year absence period, the firm is allowed to lay off the worker if the sickness implies that he/she is no longer able to perform his/her duties.

[3] Source: Own calculations based on all sick leave spells starting in 2002.

[4] Note that the maximum duration of a sick leave spell is 1 year regardless of its grade.

[5] If the firm does not comply with these rules, sick pay refunds may be rejected by the Social Security Administration. If the worker does not comply, he/she may lose the entitlement to sick pay.

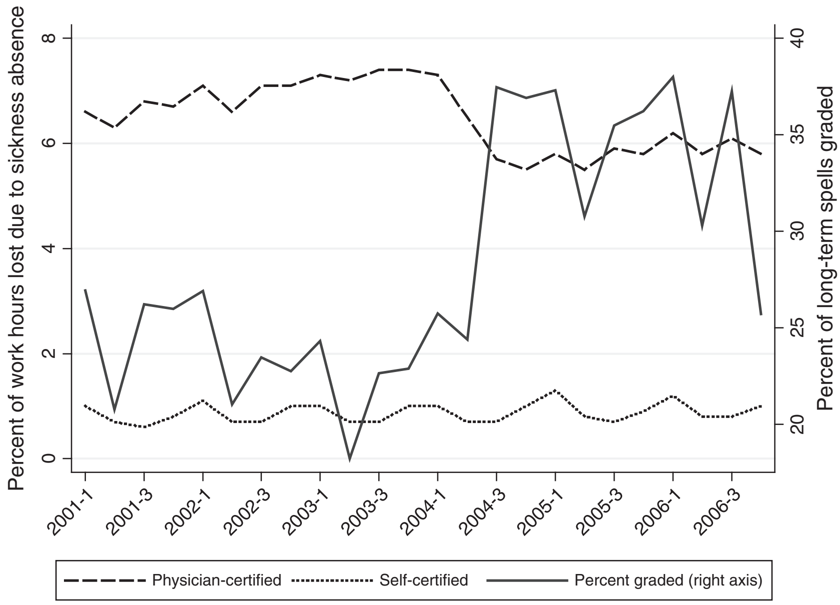


**Fig. 1.** Percent of agreed work-hours lost due to self-certified and physician-certified sickness absence and percent of long-term absence spells (more than 8 weeks) that are graded. Quarterly data (2001.1–2006.4). Note: Percent of agreed work-hours lost due to sickness is calculated as the total number of absence hours divided by the total number of contracted hours in each quarter *(source: Statistics Norway)*. The percent of spells that are graded is for each quarter calculated as the percent of spells that started in the quarter and lasted for at least 8 weeks that became graded before the 12th week *(source: own calculations based on the data used in this paper)*.

Second, it is expected to coerce employers to make appropriate efforts to facilitate adapted work, thereby limiting moral hazard problems among employers. And third, it is intended to improve the employees' health and to speed up their recovery.

In response to relatively modest use of graded sickness absence certificates during 2002 and 2003, a reform in the absence-certification-regulations was implemented in July 2004, which, inter alia, explicitly instructed physicians to use graded absence certificates for all long-term absence spells (exceeding 8 weeks) unless the spell is expected to end shortly or work-related activity is directly harmful to the health of the employee or his/her colleagues. Fig. 1 shows that while the total level of physician-certified absence trended upwards in 2002 and 2003, it declined sharply around the time of the 2004 reform.[6] At the same time, the grading frequency for long term spells rose, from around 25% before the reform to 35% afterwards. Given the significant shift in grading frequency apparently caused by the reform, it is tempting to exploit the reform directly to identify the causal effects of grading. This is problematic for at least three reasons. First, the reform also contained a number of other elements that contributed to a drop in absence incidence; see Markussen (2009).[7] Second, precisely because the reform reduced absence incidence, it is also likely to have raised average severity. And third, significant cyclical fluctuations around the time of the reform (the timing of the reform coincided with a cyclical trough in 2003–2004) make it difficult to isolate the effects of the reform from other changes in the economic environment, particularly for long-term outcomes. It is possible, though, that the reform affected grading propensities in a way that disturbs our identification strategy (because it affected different physicians differently) and/or that it changed the causal effects of grading. As part of our robustness analysis, we therefore estimate effects separately before and after the reform.

Norway has, since 2001, practiced a family (panel) doctor system, whereby each citizen is assigned a single physician who receives a capitation fee from the social security authorities. Sickness absence certificates can in principle be issued by any authorized physician, but in cases of long-term sickness, it will normally be issued by the family doctor (except when the patient is hospitalized or subject to intensive specialist treatment). Norwegian workers are free to choose their family doctor insofar as the physician in question has vacant patient slots. As we return to below, this implies that we must deal with a potential endogeneity problem when we use family doctor characteristics to instrument absence certificates.

## 4. Data

The data we use are collected from Norwegian administrative registers and include encrypted information about all citizens and their primary care physicians from 2001 and onwards. The data include detailed longitudinal information on employment and social security spells, and annual information on earnings. They also include longitudinal information on all certified absence spells from 2001 through 2005, including starting and stopping dates, diagnosis and grade (graded or not graded).[8] By merging different administrative registers (employer–employee, education, demography, social security, income/taxes) we are able to obtain ample information about each patient.

On average, around 40% of all employees experience a physician-certified absence spell during the course of a year. Fig. 2, panel a)

[6] Norwegian attempts at promoting the use of graded absence certificates were inspired by similar earlier efforts in Sweden. In Sweden, the fraction of absence certificates that are graded rose from around 25% in the 1990s to well over 30% after the turn of the century. It peaked in 2007 at a rate of 37.2%. And interestingly, also in Sweden, the rise in the grading fraction turns out to be mirrored in a decline in the overall absence rate; see www.forsakringskassan.se/press/statistik_och_analys.

[7] The most important of these were tightening of the regulations regarding self-reported absenteeism (relevant for those with multiple spells only), stricter requirements for work ability assessment in absence certificates, the introduction of sanctions against physicians who repeatedly violate absence certification regulations, and cutbacks on the use of so-called "active sick-leave" certificates, which give some employees the opportunity to work while being on *full-time* sick-leave and receiving 100% compensation from the social security administration.

[8] Note that we only use grading information as a dichotomous variable (graded or not graded). We do not exploit information on the actual grade. The reason is that there is not enough variation across physicians in grading percentages to identify the effects of actual grade on the basis of an instrumental variable approach. The most commonly used grade is 50%, which is used in around 60% of the graded absence certificates.

presents survival curves for all these absences for the year before and the year after the 2004 reform, while panel b) reports the fraction of spells that were graded at each duration. Most spells are very short (around half of them end within two weeks). As shown in Markussen et al. (2011), the group of short-term spells is dominated by respiratory infections, virus diseases, and gastrointestinal diseases, implying that grading is not relevant for these spells. It is evident that the grading-fraction rises sharply with spell duration, particularly until the 8th week, at which point grading is supposed to be the default option. Although the fraction rose significantly in response to the 2004 reform, grading is still only used for the minority of long-term absences. In the analysis of the effects of grading, we exploit data on all absence spells lasting at least 8 weeks, provided that they are handled by the family doctor. As shown in Fig. 2, around 20% of all spells last 8 weeks or more, and this fraction changed very little in response to the 2004 reform (it declined by one percentage point). Approximately 63% of these spells are handled by the family doctor, and can thus potentially be used in our statistical analysis. We drop from the analysis population all observations for pregnant women, since pregnancies typically imply a period of nonparticipation in the labor market, thus rendering some of our outcome variables difficult to interpret. In total, this leaves us with 339,251 spells. A spell is interpreted as graded if a partial absence certificate is issued before the end of week 8. We ignore grading decisions taken after week 8, since including them would introduce a source of reverse causation when we study the impact of grading on absence duration. The strategy of disregarding subsequent grading decisions introduces an element of contamination bias, however, since some of the spells in the presumed "control group" in reality become "treated" later on. To evaluate the robustness of our findings with respect to the duration-threshold used to construct our dataset, we also perform analyses with the threshold set at 4 weeks and 12 weeks, respectively. We return to these analyses in Section 5.3 after having presented the main results.

Table 1 summarizes our patient data when we use spell duration of minimum 8 weeks as the inclusion criterion Graded absence certificates were issued in around 23% of the cases, and it was used much more frequently for women than for men. Patients with graded certificates were slightly older, had slightly higher education, and had higher earnings prior to the absence spell than patients with non-graded certificates. Table 1 also presents the key patient *outcomes* that we intend to focus on in the empirical analysis:

i. The total number of days from the start to the stop of the absence spell (including holidays and days off).
ii. The number of lost fulltime equivalent working days during the absence spell (i.e., the total number of days adjusted for expected days off (two per week), regular work-hours, and absence grade).
iii. The number of additional fulltime equivalent days on social security during the 24 months following the end of the long-term absence spell (caused by new absence spells, medical and vocational rehabilitation, unemployment benefits, social assistance, or permanent disability benefits).
iv. Employment in the second year after the start of the absence spell (e.g., if the spell started in 2001, employment is evaluated in 2003).[9]

The descriptive statistics in Table 1 show that patients with graded absence certificates on average had much more favorable outcomes than patients with non-graded absence certificates. They had shorter absence durations, lower degree of social security dependency afterwards, and higher subsequent employment rates. Differences in outcomes between patients with graded and non-graded absence certificates represent a combination of sorting and causality. A person who is too sick to work at all will obtain a non-graded absence certificate and at the same time probably have a low likelihood of a quick recovery. On the other hand, a graded absence certificate normally entails non-trivial costs/efforts for the employer and the employee who need to agree on the required workplace adaptations, e.g., in the form of changes in work-hours and job contents. For this reason, an absence spell is typically not graded at the 8-week consultation if it is expected to end soon. This mechanism induces a negative correlation between grading status and future prospects.

We intend to disentangle causality from sorting by exploiting the variation in grading-propensities across physicians. There are 4044 family doctors included in our analysis. On average, each of them issued 100 long-term absence certificates during our data window (23 graded, 77 non-graded). The variation in actual use of graded absence certificates was substantial. While some physicians almost never issued graded absence certificates, others used it in more than 50% of the cases. As we return to below, this does *not* prove that physicians have different grading-propensities; the observed variation in grading frequencies results from a combination of variation in physician-practices, systematic patient sorting, and randomness.[10] The remainder of this paper seeks to identify and isolate the variation stemming from physician-practices in order to estimate the causal effects of grading.

## 5. Empirical analysis

Consider a situation where a physician faces the choice between certifying a *graded or a non-graded* absence spell for a particular worker, where the former decision implies that the worker continues working to the extent deemed tolerable.[11] The physician's decision may affect the length and the ultimate outcome of the worker's absence spell, and hence also his/her future employment prospects. The purpose of our analysis is to quantify these effects empirically within the population of workers for whom the physician's discretion actually influences the choice of treatment (the type of absence certificate).

Let $y_i$ be one of the outcomes for worker $i$, such as absence duration or subsequent employment. The regression equations of interest can then be written

$$y_i = x'_i\beta + \alpha P_{ij} + \varepsilon_i, \tag{1}$$

where $x_i$ is a vector of observed conditioning variables (including a constant term), $P_{ij}$ is an indicator variable taking the value *1* if the absence certificate issued by patient *i*'s physician *j* prescribes graded (partial) absence (and *0* if it prescribes non-graded absence), and $\varepsilon_i$ is an unobserved residual. The parameter of interest is $\alpha$. Now if we had experimental data – with controlled randomization of $P_{ij}$ – we could estimate Eq. (1) directly with an appropriate statistical model (or simply compare the outcomes for patients with full and graded absence certificates). Since we do not have experimental data, we must take into account that $P_{ij}$ is not likely to be independent of $\varepsilon_i$. A physician obviously takes the employee's recovery prospects into account when making decisions about the absence certificate, and it is unlikely that we can find observed control variables ($x_i$) that fully captures the information available to the physician. To eliminate

[9] A person is interpreted as employed in a year if earnings from regular or self employment exceeded approximately 120,000 NOK ($ 20,000) in 2005-value.

[10] Existing evidence suggests, however, that variation in clinical practice between physicians in Norway can be substantial. In particular, Grytten and Sørensen (2003) show that physician-specific effects explain more than half of the variation in expenditure for laboratory tests and almost two thirds of the variation in expenditure for consultations lasting over 20 min.

[11] The physician can obviously also decide not to certify a continuation of the absence spell at all, and in some cases, this may be the most realistic alternative to a graded absence certificate. We abstract from this complication at the present stage, but return to it below.

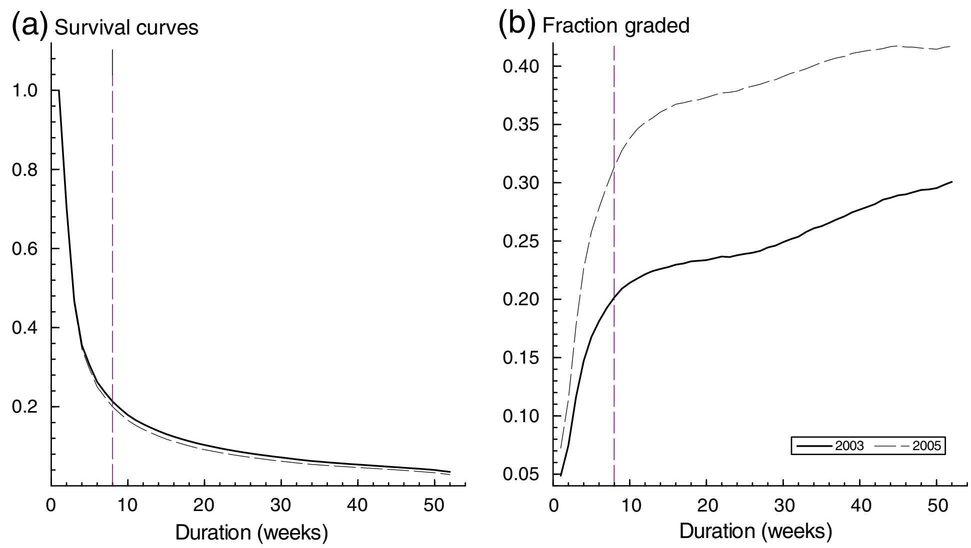


**Fig. 2.** Survival curve for all physician-certified absence spells and the fraction of spells graded at each duration. Before (2003) and after (2005) the reform. Note: The curves are computed from all absence spells starting in the years 2003 and 2005 (1,419,706 spells in total). Source: own calculations based on the register data used in this paper.

this source of endogeneity bias, we pursue an instrumental variable strategy. Our instrument is going to be a variable that we can think of as the physician's *grading practice*. This instrument will be estimated separately for each individual *i* on the basis of the observed grading frequency for all physician *j*'s patients other than *i*. While this can be shown to constitute a powerful instrument – in the sense that it strongly affects an employee's likelihood of obtaining a graded sickness absence certificate – it is more challenging to ascertain that it is completely independent of the residual ($\varepsilon_i$) in the outcome equations. There are two mechanisms by which dependency may arise. The first is that patients are non-randomly sorted to physicians, both as a direct result of the geographical location of physicians and patients and through endogenous choice of family doctor. The second is that physicians who are different with respect to their usage of graded absence certificates are different along other dimensions as well, e.g., in their overall indulgency toward issuing (unjustified) absence certificates or in their general skills. The next sub-sections explain how we have dealt with these problems and how we have ascertained empirically – by means of robustness exercises – that the problems have actually been appropriately dealt with.

Note that even though some of our outcome variables call for non-linear models (e.g., a duration model for absence duration and a logit model for subsequent employment), we have chosen to base our analyses on a linear regression framework to facilitate our instrumental variable strategy. Some of the reduced forms are re-estimated with alternative functional form assumptions in the robustness exercises in Section 5.3.

**Table 1**
Descriptive statistics.

| | Graded absence certificate | Non-graded absence certificate |
|---|---|---|
| # Observations (long-term spells) | 77,655 (22.9%) | 261,596 (77.1%) |
| *I. Patient characteristics(year t)* | | |
| Age | 44.1 | 42.4 |
| Percent female | 67.8 | 53.0 |
| Years of schooling | 13.2 | 12.4 |
| Initial annual earnings (NOK) | 330,583 | 312,660 |
| *II. Diagnosis (%)* | | |
| Mental disorders | 23.6 | 22.3 |
| Musculoskeletal diseases, back pain | 46.4 | 47.2 |
| Other | 30.0 | 30.4 |
| *III. Outcomes* | | |
| Mean absence duration (days) | 138.3 | 197.6 |
| Mean number of fulltime-equivalent sick leave days | 56.4 | 119.8 |
| Mean number of additional fulltime equivalent days with social security dependency next two years | 81.2 | 114.8 |
| Employed year $t+2$ | 89.5 | 74.6 |

Note: Year *t* is the calendar year in which the long-term absence spell starts. Initial annual earnings are measured in year *t*. These earnings are unaffected by the absence spell since sickness benefits are included in our earnings measure and the replacement ratio is 100% during the first year.

### 5.1. The instrumental variable approach

Let $j=1,\ldots,J$ be the set of family doctors in Norway. Assume that a patient's probability of obtaining a graded rather than a non-graded absence certificate at the 8-week consultation can be written as a function of the observed covariates that also affect the outcomes ($x_i$), and of his/her physician's grading propensity. We write

$$P_{ij} = x'_i\theta + u_{ij}, \tag{2}$$

where $u_{ij}$ is a residual capturing the influence of the physician, of (remaining) unobserved individual characteristics, and of genuine randomness. It seems likely that $Cov(u_{ij},\varepsilon_i)\neq 0$,since all factors affecting individual outcomes potentially also affect the grading decision. Let $\hat{\theta}$ be the OLS-estimator for $\theta$, and let $\hat{u}_{ij}$be the corresponding estimated residual from Eq. (2). The residual $\hat{u}_{ij}$ can then be interpreted as the estimated covariate-adjusted grading propensity at the patient level; hence, a natural indicator for the physicians' grading propensity is the average covariate-adjusted grading propensity among his/her patients; i.e.

$$\bar{u}_j = \left(N_j\right)^{-1}\sum_{i\in N_j}\hat{u}_{ij}, \tag{3}$$

where $N_j$ is the number of observations belonging to physician *j*. The idea behind using the physicians' grading propensities to instrument patients' grading outcomes is that some sickness insurance recipients would obtain a graded certificate with a high-grading physician, but not with a low-grading physician. This implies that our effect estimates will be based on a limited group of patients, namely those whose grading outcomes are manipulated by the physician; see Angrist and Krueger (2001, p. 77). Given the suspected correlation between $\varepsilon_i$ and $u_{ij}$, however, it is clear that $\bar{u}_j$ is not a valid instrument for $P_{ij}$ in Eq. (1), even if the assignment of patients to physicians is as good as random. The reason is that it suffers from a non-ignorable "reflection problem" (Manski, 1993) caused by the fact that each worker's own grading outcome contributes to its computation. We deal with this problem by excluding patient *i* from the computation of his/her own physician's grading-propensity, using the strategy proposed by Duggan (2005); i.e., by exploiting the identity:

$$\bar{u}_{j,-i} \equiv \frac{N_j \bar{u}_j - \hat{u}_{ij}}{N_j - 1}, \tag{4}$$

where $\bar{u}_{j,-i}$ is the covariate-adjusted grading frequency for physician *j*'s patients *excluding patient i*.[12] The first step equation in our instrumental variable approach thus becomes

$$P_i = x'_i \delta + \lambda \bar{u}_{j,-i} + \xi_i. \tag{5}$$

The reduced form outcome equation takes the form

$$y_i = x'_i \beta + \varphi \bar{u}_{j,-i} + \zeta_i, \tag{6}$$

and the instrumental variables (2SLS) estimator for $\alpha$ in Eq. (1) is

$$\alpha_{IV} = \frac{\hat{\varphi}}{\hat{\lambda}}. \tag{7}$$

Now, the assignment of physicians to patients is clearly not random; and, potentially, the characteristics of a physician's clients may be correlated to the physician's grading propensity. It is therefore critical that the covariate vector $x_i'$ removes the influence of patient sorting. We have at our disposal a comprehensive set of control variables. The potential variable sets include sickness diagnosis, patient and job characteristics, time indicators, neighborhood indicators, and physician characteristics. A complete description of these variables is provided in Appendix A. In a baseline version of the model, we control for sickness diagnosis, patient/job characteristics, and calendar time only. Diagnosis is represented by 60 dummy variables based on the International Classification of Primary Care (ICPC 2), accounting for the patient's most important diagnosis or symptoms. To some extent, this system also incorporates the sickness' severity. Patient/job characteristics include age, nationality, marital status, education, industry, work-hours, and earnings. To avoid unjustified functional form restrictions, most of the variables are entered in the form of extensive dummy-sets. For example, we use 40 age dummies, 70 education type/level dummies, 61 industry dummies (two-digit NACE), and 10 earnings dummies (corresponding to deciles in the age-specific earnings distribution). We control for calendar time by means of one dummy variable for each of the 56 possible entry months.

The baseline model is convenient to illustrate how the probability of obtaining a graded absence certificate depends on the characteristics of the workers and their jobs. Based on Eq. (2), Table 2 and Fig. 3 show how various selected patient/job characteristics are estimated to affect the probability of obtaining a graded absence certificates. Table 2 first shows that graded certificates are used much more frequently for women than for men, that it is used more for married

**Table 2**
The grading decision (OLS on Eq. 2).

| | Coefficient (standard error) |
|---|---|
| Female | 0.0949 (0.0020) |
| Age-specific income decile (first decile is the reference) | |
| 2 | 0.0391 (0.0032) |
| 3 | 0.0591 (0.0032) |
| 4 | 0.0762 (0.0033) |
| 5 | 0.0843 (0.0034) |
| 6 | 0.0871 (0.0034) |
| 7 | 0.0924 (0.0034) |
| 8 | 0.0980 (0.0035) |
| 9 | 0.1000 (0.0035) |
| 10 | 0.1150 (0.0037) |
| Employment type (full time position is the reference) | |
| Less than half position | −0.0930 (0.0032) |
| Half to full position | −0.0796 (0.0026) |
| Marital status (married is the reference) | |
| Not married | −0.0216 (0.0019) |
| Widow/Widower | −0.0184 (0.0060) |
| Divorced | −0.0249 (0.0021) |
| Legally separated | −0.0148 (0.0039) |
| Foreign origin (native is the reference) | |
| Western Europe | −0.0099 (0.0032) |
| Eastern Europe | −0.0572 (0.0066) |
| Africa | −0.0317 (0.0085) |
| Asia | −0.0311 (0.0046) |
| North America | 0.0010 (0.0026) |
| South America | −0.0261 (0.0242) |
| Oceania | −0.0232 (0.0114) |
| Number of observations | 339,251 |

than for unmarried persons, and that it is used more for natives than for immigrants. The probability of obtaining a graded absence certificate is higher for full-time than for part-time workers, and it also rises with the patient's position in the (age-specific) income distribution. Fig. 3 summarizes the estimated impacts of calendar time and age. Panel (a) shows that the probability of obtaining a graded certificate rises with age up to around 45 years, after which it remains more or less constant. Panel (b) confirms that grading became more popular after the reform in July 2004. It also reveals that there is a strong seasonal pattern in grading propensities, with particularly low grading propensity for spells starting in the summer holiday season. As explained above, the model also contains indicator variables representing the influences of education, industry, and sickness diagnosis. Given the limited interest of the particular coefficient estimates for these variable groups, we do not present these results in any detail. It is notable, however, that job characteristics – as captured by educational attainment and industry – seem to matter more for the grading decision than sickness diagnosis.[13]

Even though our baseline model includes a rather extensive set of control variables, we may still worry that $\bar{u}_j$, based on Eq. (3), to some extent absorbs uncontrolled-for patient characteristics that vary systematically across physicians. The by far most important patient–physician allocation determinant in Norway is *geography*. Virtually all patients are listed with a doctor whose practice is located close to their home. Since residential areas are socially segregated, this implies an obvious risk of non-random physician–patient sorting which may or may not have been appropriately controlled for by our observed covariates. Hence, in an extended version of the model, we also include dummy variables for 12,906 "neighborhoods" in Norway. These neighborhoods are drawn up by Statistics Norway for statistical purposes, and are designed to be socially and economically homogeneous. On average, there are 350 inhabitants in each neighborhood who together

[12] We have also estimated the model by means of split-sample techniques where we have used the observed grading propensity for female patients as an instrument for men's grading outcome and vice versa. These models produced similar results as those reported here, but with substantially larger standard errors. There were only modest differences in estimated effects for men and women. The results are documented in a previous discussion paper version of this paper; see Markussen et al.(2010).

[13] The standard deviations in estimated education, industry, and diagnosis impacts are 0.041, 0.039, and 0.033, respectively. By comparison, the standard deviations generated by age and calendar time (Fig. 3) are 0.024 and 0.061, respectively.

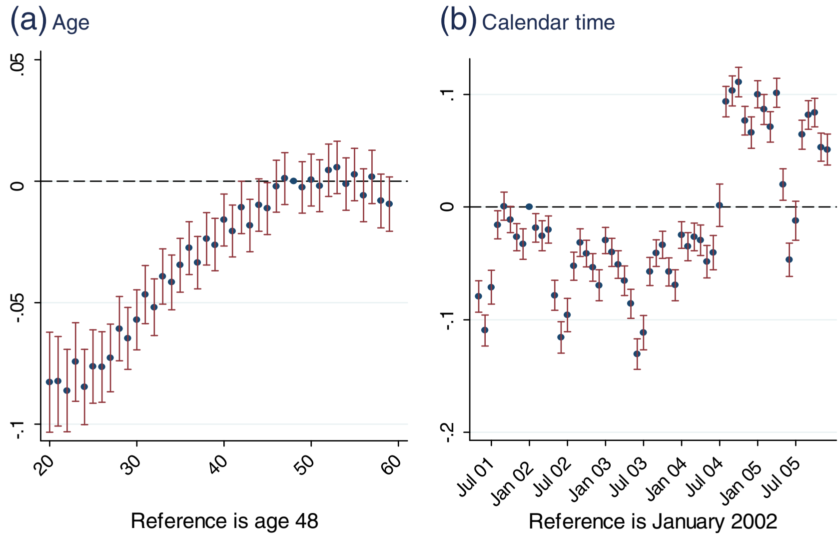


**Fig. 3.** Estimated impacts of age and calendar time on the probability of obtaining a graded absence certificate (with 95% confidence intervals). Note: The graphs report OLS-results based on Eq. (2).

contribute around 30 long-term absence spells to our analyses. In the extended model, we also control for patient history in terms of employment, earnings, and sick-leave during the past three years prior to the spell used in our analysis.[14]

Fig. 4 illustrates the resultant distribution of the physicians' grading propensities. The physicians are here divided into 20 bins (with around 190 physicians in each bin), sorted on the basis of their estimated grading propensities. The graphs show the distributions for three different models, i.e., without any controls (the observed grading variation), with the controls used in the baseline model, and with the controls used in the extended/full model. It is clear from these graphs that parts of the variation in grading frequencies are indeed explained by patient sorting. However, even when we include the most extensive set of controls, the variation in estimated grading propensities remains substantial. It is not much larger, though, than what would be expected from a completely random allocation of grading decisions (i.e., with all patients having the same probability of obtaining a graded absence certificate). Hence, the extent to which the distribution in Fig. 4 constitutes a useful foundation for an instrumental variable strategy remains to be explored.

To check for the presence of systematic patient–physician sorting related to the physicians' grading behavior, we have compared the physicians' estimated grading propensities with some key characteristics of their *complete* patient lists; i.e., all their registered clients (recall that we have only used a small fraction of their patients – namely those with long-term absence – to estimate their grading propensity). This exercise reveals that there are signs of patient–physician sorting. We find that patients' employment propensity and years of schooling rise slightly with their physician's grading propensity, with correlation coefficients equal to 0.09 and 0.11. For annual earnings (conditional on employment) and mortality, on the other hand, the relationship is much weaker (and also non-monotonic), with correlation coefficients equal to 0.04 and −0.04, respectively.

Note that patient–physician-sorting does not invalidate our empirical strategy insofar as we control appropriately for the patient characteristics responsible for inducing a correlation between the patient outcomes and estimated grading propensity. To evaluate this, we examine the correlation between the physicians' grading propensities and their clients' outcomes for the population of clients that did *not* have any physician-certified sickness in the first part of our observation window (2001–2004). The idea behind this exercise is that subsequent outcomes for these workers cannot have been much affected by their physicians' grading decisions; hence they can serve as a test for remaining uncontrolled-for sorting. We use the patient–physician affiliation in 2003 as the basis for this exercise, and investigate whether the physician's estimated grading propensity $\bar{u}_j$ is a predictor for his/her clients' outcomes in 2005, conditional on a full set of explanatory variables. The results are presented in Table 3. The estimated effects on the two absence incidence outcomes are for all practical purposes equal to zero. The effect on employment propensity is positive and statistically significant at conventional levels, but close to zero. To illustrate the size of the coefficient, we note that a one standard deviation increase in the physician's grading propensity (by 5.7 percentage points) raises the subsequent employment rate among his/her non-absent clients by $0.057 \times 0.019 = 0.001$, i.e., one out of thousand. While these results indicate that our control variables do remove most of the systematic patient–physician sorting on outcomes such as future employment and absence propensities, we cannot completely rule out elements of remaining sorting-problems.

A particularly troublesome source of patient–physician sorting arises if workers who are offered a graded absence certificate respond by choosing a new physician, rather than by going back to work. This could generate a pattern of patient–physician sorting such that the motivated workers end up with the physicians with high grading propensity. One way to examine the empirical relevance of this mechanism is to look at the pattern of physician changes in the period around the issuing of absence certificates. Table 4 shows that there was indeed a higher tendency for employees with non-graded absence certificates to have changed physician just before or during the sick leave spell than for employees with graded certificates (3.95 versus 2.79%). There was also a tendency for patients to move toward physicians with low grading propensity in the period around the issuing of a long-term absence certificate. But again, the numbers as well as the differences are small. We evaluate their potential impacts on the results by linking patients to their past – instead of their current – physician in the robustness analysis in Section 5.3 below.

Even if our explanatory variables had fully controlled for patient–physician sorting, we are not ensured a valid instrument for the

[14] The inclusion of 12,906 neighborhood dummy variables in addition to several other large dummy variable sets raises some computational challenges. We have used a novel algorithm based on the method of alternating projections implemented by Simen Gaure in the R-package "lfe"; see http://cran.r-project.org/web/packages/lfe/citation.html.

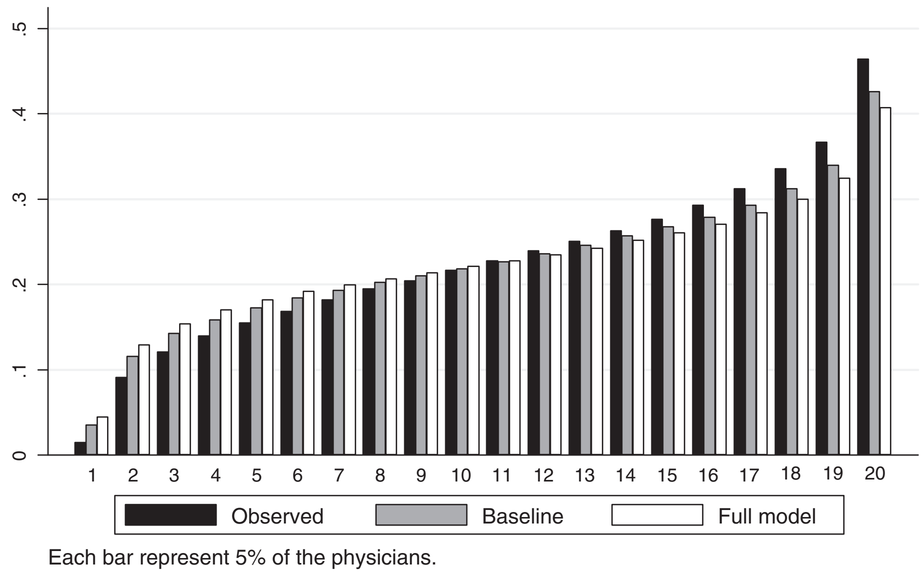


**Fig. 4.** The estimated distribution of physicians' grading propensities. Note: The numbers on the x-axis correspond to the 20 physician-groups ranked according to their position in the grading propensity distribution, so that group 1 consists of the 5% with lowest grading propensities and group 20 consists of those with the highest grading propensities. To make the estimated distributions from the baseline and full models comparable to the observed distribution, we have centered the estimated physician effects on the observed average grading frequency.

patients' grading outcomes in Eq. (1). As indicated above, it is also required that the physicians' grading propensities are uncorrelated to other relevant physician behaviors that directly affect the evaluated outcomes. In particular, we may worry that physicians who frequently grade their absence certificates do so precisely because they pursue a more "lenient" certification practice in the first place, implying that they also tend to have relatively healthy absentees on their list. We assess this potential problem by investigating the observed relationship between the physicians' grading propensities, on the one hand, and the fractions of their patients who obtain *any* absence-certificate and a *long-term* absence-certificate, respectively (the latter implies inclusion in our analysis population). As it turns out there is a weak *negative* correlation between physicians' grading propensities and their overall inclination to issue absence certificates; the correlation coefficients are −0.02 for all absences and −0.03 for long-term absences. If anything, this indicates that physicians with high grading propensities are *less* lenient than other doctors, and hence have absentees with more severe health problems.

To sum up, depending on the set of control variables, we see some patterns of systematic patient–physician sorting and also indications that physicians who are different with respect to grading propensity also differ along other dimensions. In the statistical analysis, we therefore put considerable emphasis on robustness with respect to the inclusion of controls, and also estimate models where we control for a number of physician characteristics, including variables designed to reflect the physicians' overall leniency (with respect to certifying absence) and the quality of their medical treatment. These latter "practice style indicators" are computed in a similar fashion as the grading propensities (i.e., by auxiliary regressions), but based on larger/different patient populations; see Appendix A for details.[15] They have already been shown to have strong predictive power for worker absenteeism in general; see Markussen (2012). The inclusion of physician characteristics among the control variables reduces the risk that our grading-instrument captures other dimensions of the physicians' practice styles that happen to be correlated with their grading strategy, but have distinct impacts on the outcome variables. On the other hand, it raises the risk that we attenuate the true impact of grading by attributing genuine grading effects to physician characteristics that affect outcomes precisely through their correlation with the grading indicator. We therefore report results for the full model both with and without physician characteristics included. After having presented the main results, we also run a battery of robustness/specification checks, with alternative data inclusion strategies, and alternative (non-linear) functional forms. We also report regressions on *past* outcomes, i.e., outcomes for which a valid instrument should not have any effect (placebo regressions).

[15] Observed physician characteristics also account for a part of the variation in the physicians' grading propensities. In particular, physicians who have fewer clients than their target number (i.e., patient shortage) use grading *less* than physicians who are "fully booked". It is also notable that we identify a significant negative correlation between grading propensity and the overall propensity to issue long-term absence certificates, but a small positive correlation with the propensity to issue any kind of absence certificates. We identify a *positive* correlation between grading propensity and treatment quality.

### 5.2. Main results

The main estimation results are presented in Table 5; for the baseline model, for the extended/full model, and for the full model with a vector of physician characteristics included among the control variables (see Appendix A for details). While we have argued that instrumental variable (2SLS) is the appropriate estimation technique in this case, we also report OLS estimates for comparison. A first point to note is that the physician's estimated grading propensity has a significant impact on the likelihood of obtaining a graded absence certificate; confer the first-stage coefficients. Hence, our instrument is strong. When a physician's covariate-adjusted grading frequency for other patients rises by 1 percentage point, the probability of obtaining a graded absence certificate rises by 0.4–0.5 percentage points, ceteris paribus. The second stage estimates indicate that the grading decision has large and lasting impacts on all patient outcomes. The 2SLS results are similar for the different models, despite huge differences in the conditioning sets. In particular, the estimated effects change only marginally when we add 12,906 neighborhood dummy variables and a vector of 38 physician variables (including proxies for leniency and treatment quality; see Appendix A). All the models indicate that by grading long-term absence certificates, physicians contribute to shorter absence durations, less subsequent social security dependency,

**Table 3**
Estimated effects of physicians' grading propensities on outcomes for clients who were employed in 2003 and had no absence spells 2001–2004 (robust standard errors in parentheses, clustered on physicians).

| Effect on: | |
|---|---|
| Employment propensity 2005 | 0.019 |
| | (0.005) |
| Incidence of sick leave 2005 | 0.000 |
| | (0.011) |
| Incidence of long-term sick leave (>8 weeks) 2005 | −0.004 |
| | (0.004) |
| # Observations | 402,314 |

Note: Estimated by OLS. The regression includes patient and job characteristics, and neighborhood indicators; see Appendix A for details.

**Table 4**
Observed physician changes before (1 month) and during sick leave spells. By patients' grading status and physicians' grading propensities (standard errors in parentheses).

| | Actual grading | | | |
|---|---|---|---|---|
| | Graded | Non-graded | | |
| Percent changing before spell | 0.93 | 1.14 | | |
| | (0.03) | (0.02) | | |
| Percent changing during first 8 weeks of the spell | 1.86 | 2.81 | | |
| | (0.04) | (0.03) | | |
| | Physician's grading propensity. By quartile | | | |
| | Q1 | Q2 | Q3 | Q4 |
| Percent changing before spell | 1.51 | 0.98 | 0.88 | 0.98 |
| | (0.05) | (0.03) | (0.03) | (0.04) |
| Percent changing during first 8 weeks of the spell | 3.32 | 2.27 | 2.10 | 2.47 |
| | (0.07) | (0.05) | (0.05) | (0.06) |

Number of observations is 335,665.

and higher employment propensities. The effects are large from an economic – as well as from a clinical – viewpoint. Based on the 2SLS estimates, we conclude that a graded absence certificate reduces the expected number of (fulltime-equivalent) work-days lost during the spell by more than 90 days. In addition it reduces subsequent social security claims and raises employment. The number of saved social security days in the two-year period following just after the end of the absence spell is similar to the number of days saved during the spell, i.e., around 80–90. The potentially most important impact of grading, however, is that it significantly raises the probability of remaining employed afterwards. Issuing a graded rather than a 100% absence certificate raises the employment probability two years later by around 16 percentage points (according to the most conservative 2SLS-estimate).

Comparing the 2SLS with the OLS results, we find that there is a tendency for 2SLS to yield similar effects on absence duration, but larger effects on the other outcomes. This suggests that unobserved sorting into graded absence certificates is generally negative — in the sense that patients with poor log-term outcome prospects have higher grading likelihoods than patients with more favorable outcome prospects. A likely explanation for this is that grading is generally not used – even after 8 weeks of absence – when health problems are expected to end in the near future; conf. the discussion in Section 4. However, 2SLS not only differs from OLS in that it removes sorting bias; as pointed out above it also bases the effect estimates on a more limited group of patients, namely those whose grading outcomes are manipulated by the instrument. Our 2SLS estimates should thus be given a local average treatment effect (LATE) interpretation. It is conceivable that the complier-group – those whose grading outcomes are influenced by the physician – are workers of relatively good health, but with an uncooperative employer and/or limited own motivation, implying that grading serves the dual purposes of containing moral hazard problems and promoting healthy work. By contrast, for motivated "always-takers" – i.e., patients who actively ask for a graded rather than a non-graded absence certificate – it is likely that only the latter effect is relevant.

### *5.3. Robustness and reliability*

In this section, we perform a battery of robustness checks and "placebo" regressions to assess the reliability of our findings. The first set of robustness checks are presented in Table 6. For convenience, Column A repeats the estimates from the full model (Table 5, Column F). Columns B and D report estimation results based on alternative duration conditions for sample inclusion. While we have used absence spells exceeding 8 weeks in the main analysis, we use all spells exceeding 4 weeks in Column B and spells exceeding 12 weeks in Column C. Consequently, the results in Column B are based on more than twice as

**Table 5**
Main results (robust standard errors in parentheses, clustered on physicians).

| | Baseline model | | Full model | | Full model w/physician char. | |
|---|---|---|---|---|---|---|
| | A<br>OLS | B<br>2SLS | C<br>OLS | D<br>2SLS | E<br>OLS | F<br>2SLS |
| *I. First stage*: | | | | | | |
| Effect of excluded instrument | – | 0.52 | – | 0.41 | – | 0.38 |
| | | (0.01) | | (0.02) | | (0.02) |
| F excluded instrument | – | 1350.9 | – | 484.6 | – | 373.4 |
| *II. Second stage. Effect of grading on*: | | | | | | |
| Absence duration (days) | −59.6 | −58.8 | −60.2 | −66.1 | −60.1 | −62.6 |
| | (0.5) | (8.0) | (0.5) | (10.8) | (0.5) | (11.3) |
| # Fulltime-equiv. lost working days in abs. spell | −63.9 | −93.8 | −63.9 | −97.8 | −63.9 | −96.7 |
| | (0.3) | (4.7) | (0.3) | (6.5) | (0.3) | (7.0) |
| # Fulltime-equiv. soc. sec. days next two years | −26.8 | −102.3 | −26.7 | −91.0 | −26.5 | −79.7 |
| | (0.5) | (8.2) | (0.5) | (9.7) | (0.5) | (11.8) |
| Employment in year $t+2$ | 0.10 | 0.21 | 0.10 | 0.19 | 0.10 | 0.16 |
| | (0.00) | (0.03) | (0.00) | (0.04) | (0.0) | (0.04) |
| *III. Controls* | | | | | | |
| Patient char., job char., and diagnosis | Yes | Yes | Yes | Yes | Yes | Yes |
| Calendar time controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Patient history (empl., earnings, absences) | No | No | Yes | Yes | Yes | Yes |
| Neighborhood fixed effects | No | No | Yes | Yes | Yes | Yes |
| Physician characteristics | No | No | No | No | Yes | Yes |
| # Observations | 339,251 | 339,251 | 336,134 | 336,134 | 336,134 | 336,134 |

Note: Year $t$ is the calendar year in which the long-term absence spell starts.

**Table 6**
Robustness analysis (robust standard errors in parentheses, clustered on physicians).

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| *I. First stage*: | | | | | | |
| Effect of excluded instrument | 0.38 | 0.52 | 0.30 | 0.34 | 0.35 | 0.43 |
| | (0.01) | (0.02) | (0.02) | (0.02) | (0.02) | (0.03) |
| F excluded instrument | 373.4 | 704.9 | 230.0 | 327.9 | 249.3 | 203.6 |
| *II. Second stage. Effect of grading on*: | | | | | | |
| Absence duration (days) | −62.6 | −47.1 | −74.8 | −68.8 | −77.2 | −40.3 |
| | (11.3) | (8.8) | (10.3) | (12.3) | (14.5) | (16.0) |
| # Fulltime-equiv. lost working days in abs. spell | −96.7 | −70.1 | −117.1 | −103.4 | −108.1 | −78.8 |
| | (7.0) | (5.2) | (8.2) | (7.6) | (9.2) | (9.5) |
| # Fulltime-equiv. soc. sec. days next two years | −79.7 | −67.7 | −99.8 | −98.5 | −95.6 | −52.9 |
| | (11.8) | (8.6) | (15.6) | (13.6) | (15.7) | (17.6) |
| Employment in year $t+2$ | 0.16 | 0.13 | 0.20 | 0.21 | 0.18 | 0.13 |
| | (0.04) | (0.02) | (0.05) | (0.04) | (0.05) | (0.05) |
| *III. Model, data, and controls* | | | | | | |
| Duration condition for inclusion in dataset | 8 weeks | 4 weeks | 12 weeks | 8 weeks | 8 weeks | 8 weeks |
| Patient-physician linkage | Correct | Correct | Correct | 6 m. bef. spell | Correct | Correct |
| Data period (start of abs. spell) | 2001.5–2005.12 | 2001.5–2005.12 | 2001.5–2005.12 | 2001.5–2005.12 | 2001.5–2004.6 | 2004.7–2005.12 |
| Patient char., job char., and diagnosis | Yes | Yes | Yes | Yes | Yes | Yes |
| Calendar time controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Patient history | Yes | Yes | Yes | Yes | Yes | Yes |
| Neighborhood fixed effects | Yes | Yes | Yes | Yes | Yes | Yes |
| Physician characteristics | Yes | Yes | Yes | Yes | Yes | Yes |
| # Observations | 336,134 | 524,782 | 252,649 | 298,084 | 242,449 | 95,764 |

many observations as the results reported in Column C. Spells are treated as graded if a graded certificate is issued before or at the time of entry into the dataset in question (i.e., before the end of weeks 4 and 12, respectively). Otherwise, we have used exactly the same instrumental variable strategy as the one described in Sections 5.1 and 5.2. The results are also very similar. There is a tendency for the estimated impacts to be larger the longer are the included spells, probably reflecting that grading indeed is more effective in cases of long-term sicknesses. The main conclusions coming out of the analyses, however, are remarkably similar across the different datasets.

To investigate the possible bias arising from endogenous physician-switching, Column D reports estimates based on a model where each patient is linked to the physician he/she had 6 months before the start of the analyzed spell. We then lose around 38,000 (11%) of the observations, since we cannot identify physicians 6 months before for entrants in the first half year of our observation window. The estimation results indicate even larger effects than those obtained using the current physician. The instrument becomes weaker, however, and the standard errors become larger in this model, hence the results are less precise and reliable. We nevertheless conclude that our main results hardly can be explained by physician-shopping in the period just prior to, or during, absences.[16]

As described in Section 2, there was a reform in the absence certification regulations in July 2004, which raised the use of graded absence certificates and also reduced the overall number of certified absence spells. This reform may potentially have affected different physicians' certification and grading practices differently and, hence, reduced the reliability of our instrument. It is also probable that the sharp increase in the use of graded certificates entailed a change in the true local average treatment effect. To assess these concerns, Columns E and F report results from models estimated separately on the data before and after the reform. For both periods, the results are similar to the results based on the full model. However, it does indeed seem to be the case that the favorable effects of grading declined somewhat in response to the sharp increase in its usage after the 2004 reform.

For two of our outcomes – absence duration and employment propensity – it may be argued that the linear modeling framework is inappropriate. The reason why we nevertheless have chosen a linear model specification is that it facilitates the instrumental variable strategy. For the reduced forms, it is possible to assess the potential consequences of the linearity assumption by estimating alternative models based on more appropriate functional forms. We have thus re-estimated reduced form models (i.e., models where we use the physician-based instrument directly as an explanatory variable) for absence duration and for employment propensity (two years after). For absence duration, we have estimated a hazard rate model with duration- and calendar time dummy variables for each absence-week. For employment propensity, we have estimated a logit model. Both models use the explanatory variables from “the baseline model”, i.e., we do not include the neighborhood indicators in this exercise. The results are reported in Table 7. In order to compare them with the

**Table 7**
Estimated effects of physician's grading propensity based on a hazard rate model for absence duration and a logit model for employment propensity (standard errors in parentheses).

| | Effect on log (hazard rate) | Average marginal effect (logit) |
|---|---|---|
| Effect of physician's grading propensity on: | | |
| The return to work hazard rate | 0.21 | |
| | (0.03) | |
| Employment in year $t+2$ | | 0.11 |
| | | (0.01) |
| # Observations | 339,251 | 339,251 |

Note: The models include patient/job characteristics, diagnoses, and calendar time controls. The hazard rate model also includes duration dummy variables (one for each week).

[16] Similar results are obtained when we base the analysis on patients' physician-affiliation 12 months before the absence spell We have also estimated models on the subset of non-switchers only, and the results from these models confirm this conclusion; see Markussen et al. (2010).

2SLS results reported for the baseline model in Table 5, we need to multiply the first- and second stage coefficients to obtain the reduced form coefficients. It is then clear that average marginal employment effect of 0.11 reported for the logit model in Table 7 is equal to the corresponding linear reduced form coefficient underlying Table 5 ($0.21\times0.52=0.11$). For the absence duration measure, it is a bit more difficult to compare the two models. For a typical absence spell, the estimated hazard rate effect of 0.21 corresponds to a drop in absence duration (conditional on at least 8 weeks duration) of around 23% — or 32.4 days. This is strikingly similar to the reduced form estimate underlying the results reported in Table 5 ($-58.8\times0.52=-30.6$). Based on these exercises, we find it unlikely that our results have been significantly biased as a result of the linearity assumption.

To further evaluate the reliability of our preferred 2SLS model, we finally estimate it on a set of *past outcomes* for which the grading decisions under study could not possibly have had any causal effect. Since we have used patient history during the three years prior to the start of the spell in our control variable set, we use observed outcomes four years prior to the start of the absence spell to test the model (this also reduces the likelihood that the physician has affected past outcomes through previous grading decisions). We focus on three outcome measures in this exercise, i) incidence of social insurance dependency, ii) employment and/or regular education, and iii) incidence of long-term sickness absence (conditional on employment). The first of these outcomes captures all kinds of social security transfers, including sickness benefits, unemployment benefits, rehabilitation benefits, disability benefits, and social assistance. A person is defined as being dependent on social security in a given year if benefits were received for at least three months. The reason why we use this comprehensive outcome measure in our back-in-time estimation is that a large fraction of our absentees would not have been eligible for sick-pay or rehabilitation/disability benefits in the past due to lack of employment experience at that time. The combined employment-education outcome is chosen as an alternative to a pure employment-measure for the reason that many of our absentees did not participate in the labor market in *t-4*, implying that employment alone would be an awkward indicator for labor market success in the past.

Estimation results for the back-in-time analysis are presented in Table 8. We find no “effects” whatsoever of the current grading decision on past social security dependency and past employment/education. For past absence (conditioned on employment) we find a small *positive* effect, if anything indicating that there is adverse unobserved selection (in terms of absenteeism) to physicians with high grading propensity. This is consistent with our finding in the previous subsection that physicians with high grading propensity tend to be “stricter” than other physicians, and clearly opposite of what we would worry about if we were searching for a sorting-explanation for the estimated favorable effects on future outcomes.

**Table 8**
Estimated second stage (2SLS) effects on past outcomes (t-4) (robust standard errors in parentheses, clustered on physicians).

| “Effect” of grading on | |
|---|---|
| Social insurance dependency in *t-4* | −0.014 |
| | (0.037) |
| Employment/education in *t-4* | 0.029 |
| | (0.023) |
| Absence (conditional on employment) in *t-4* | 0.075 |
| | (0.032) |
| # Observations | 336,134 |
| | 289,818 (*cond. on empl.*) |

Note: The models include individual/job characteristics, patient history, diagnoses, physician characteristics, and neighborhood indicators.

## 6. Concluding remarks

Our estimation results indicate that requiring temporary disabled workers on long-term sick leave to be active – and work to the extent deemed tolerable by their physicians – is an effective tool to reduce social security dependency and raise subsequent employment. The effects are significant, both from economic and statistical viewpoints. Our most conservative estimates imply that substituting a graded for a non-graded absence certificate cuts the number of lost (fulltime-equivalent) working days during the sick leave by more than half, and raises the employment propensity two years later by 16 percentage points. Although our results are robust with respect to a number of alternative data collection strategies and model specifications, we obviously cannot establish with certainty that our instrumental variable strategy has solved all the sorting problems that come with non-experimental data. In particular, we cannot rule out that elements of unobserved patient–physician sorting remain or that the physicians' estimated grading propensities correlate with other unobserved physician characteristics with potential influence on patient outcomes. We will argue, however, that taken as a whole, the evidence presented in this paper convincingly shows that activation requirements do have the intended favorable effects on social security dependency and employment. We also find no evidence whatsoever that the use of graded absence certificates reduces the threshold for claiming sickness benefits. To the contrary, physicians who frequently use graded absence certificates also seem to issue fewer absence certificates in the first place.

Balancing the objectives of appropriate social insurance and sufficient work-incentives is a difficult task. For unemployment insurance and social assistance programs, policy makers in many countries have to an increasing extent resorted to various *activation strategies*, essentially requiring benefit claimants to participate in temporary employment or training programs. The key idea behind this strategy – with potential appeal to the political right as well as to the left – is that by pairing insurance with activity requirements it becomes possible to partly escape the unpleasant tradeoff between equality and work incentives; i.e., it facilitates a reduction of the moral hazard problem, given the level of insurance, or, alternatively, an improvement in insurance coverage, given the level of moral hazard. The results presented in the present paper suggest that the same strategy may successfully be pursued for publicly provided insurance against income losses due to sickness absence. When workers with health problems are required to exploit remaining work capacity in order to be eligible for insurance payments, it probably becomes less attractive to report sick when it is not strictly necessary. It also becomes more difficult for employers to “get rid of” workers with health problems. Moreover, recent empirical evidence indicates that work is actually a *healthy* activity for workers with the illnesses and symptoms responsible for the vast majority of absence days in industrialized countries (musculoskeletal diseases, back pain, and light mental disorders). Although it could be argued that patients have every opportunity to take the expected adverse long-term consequences of inactivity into account when determining their own absence behavior, it is not difficult to imagine that some of them fail to do so, either because of insufficient information or limited self-control. In any case, the existence of generous sick leave insurance and the absence of experience rating imply that the decisions of employers and employees are distorted to some extent.

What we have shown in this paper is that more intensive use of graded absence certificates has the potential of significantly reducing the volume of social security insurance payments caused by sickness, and also of strengthening the employees' subsequent labor market attachment. The apparent decline in the estimated favorable effects following the huge expansion of the activation strategy in 2004 indicates, however, that there may be some declining returns to the use of graded absence certificates. There are obviously also some potential costs involved that we have not looked into in this paper. In particular, graded absence may involve workplace adaptations that impose costs on employers and/or co-workers. Physicians' assessments of sick workers' remaining work capacity are also likely to be imprecise, implying that the insurer will not typically compensate the true productivity loss caused by health problems. Given the large gains involved, we nevertheless conclude that incorporating activation requirements into temporary disability insurance stands out as a promising avenue for future social security reform.

## Appendix A. Overview of control variables

**Table A1**
List of control variables used in the regression analysis (numbers include omitted reference group).

| |
|---|
| Diagnosis |
| 60 dummy variables describing type and severity of illness. |
| Patient characteristics |
| Gender: One dummy for each gender (2 dummy variables) |
| Age: One dummy variable for each yearly age (40 dummy variables) |
| Education: One dummy for each type-level education (70 dummy variables) |
| Industry: One dummy for each industry, two-digit NACE (61 dummy variables) |
| Nationality: One dummy for each region of origin (8 dummy variables) |
| Marital status: 9 dummy variables describing marital status |
| Earnings level: One dummy variable for each decile in the age-specific income distribution (10 dummy variables) |
| Work-hours: One indicator for each of three working time arrangements (3 dummy variables) |
| Patient history |
| Employment and earnings: One dummy variable for employment and one scalar with earnings in NOK (deflated to 2005-value) for each of the years *t-1*, *t-2* and *t-3*. |
| Absenteeism: One dummy for each of six possible sick leave histories last three years (6 dummy variables) |
| Neighborhood |
| One dummy variable for each neighborhood in Norway (12,906 dummy variables) |
| Calendar time |
| One dummy variable for each starting month occurring in our observation window (56 dummy variables) |
| Physician characteristics (see description below) |
| Age (5 dummy variables based on 10-year grouping) |
| Specialist education (2 dummy variables) |
| Gender (2 dummy variables) |
| Sharing office with other physicians (2 dummy variables) |
| Fixed or variable salary (2 dummy variables) |
| Taking part in emergency service (2 dummy variables) |
| Number of patients on capitation list |
| Desired number of patients relative to actual number |
| Physician leniency (centiles in the distribution of estimated leniency for issuing short-term and long-term absence certificates (20 dummy variables)); see below. |
| One scalar indicator for treatment quality; see below. |

The list of controls used in the regression analysis includes 369 variables, most of them dummy variables. In addition, in some of the models, the data are centered on 12,906 neighborhood dummies. A list of all control variables used in this paper is provided in Table A1

*Derivation of the physicians' practice style indicators*

The physicians' practice style indicators have been computed by means of auxiliary regressions. Leniency is measured by the physicians' propensity to issue absence certificates to the employees on their capitation list. To compute physician leniency, we exploit data on all Norwegian employees (not only those with long-term absence), and set up auxiliary linear probability models in a similar fashion as we did to compute grading propensity, with a dummy variable for each physician along with a large number of individual controls. We compute two leniency-indicators, one based on the annual incidence of certified sick leave (denoted overall leniency) and one based on the probability of having a long-term sick leave (denoted long-term leniency). The dataset includes more than 7 million worker-year observations. The estimated coefficients attached to the physician dummies are taken to represent physician leniency, and subsequently used as additional control variables in the instrumental variable analysis along with other observed physician characteristics. Given the large number of observations behind these regressions, we have ignored the “reflection problem” in this exercise, i.e., we have not removed person *i* from the computation of his/her own physician's leniency.

Physicians' treatment quality is clearly a latent variable, and since the health status of patients is also unobserved, it is not obvious how we should control for this variation. We may hypothesize that the quality of a physician's medical advice is correlated to the patients' survival, conditioned on all observed patient characteristics, in which case we may use data on mortality to compute a proxy for physician influences beyond their absence certification practices. We realize, of course, that physicians have limited influence on their elderly patients' survival, and, hence, that a practice style measure based on observed mortality is likely to be noisy. To compute our indicator, we have exploited an additional cut of our data consisting of elderly retirees (around 4 million person-year observations). Again, we have used an auxiliary linear probability model, this time with survival as the outcome and again with physician dummies and patient characteristics as explanatory variables.

## References

Abbring, J.H., Van den Berg, G.J., 2003. The non-parametric identification of treatment effects in duration models. Econometrica 71 (5), 1491–1517 (September).
Angrist, J.D., Krueger, A.B., 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. Journal of Economic Perspectives 15 (4), 69–85.
Bonato, L., Lusinyan, L., 2007. Work absence in Europe. IMF Staff Papers 54 (3), 475–538.
Bound, J., 1989. The health and earnings of rejected disability insurance applicants. American Economic Review 79 (3), 482–503.
Bound, J., Burkhauser, R.V., Nichols, A., 2003. Tracking the household income of SSDI and SSI applicants. Research in Labor Economics 22, 113–159.
Bratsberg, B., Fevang, E., Røed, K., 2010. Disability in the welfare state: an unemployment problem in disguise? IZA Discussion Paper No. 4897.
Burkhauser, R.V., Daly, M.C., 2011. The Declining Work and Welfare of People with Disabilities: What Went Wrong and a Strategy for Change. AEI Press, Washington D.C.
Doyle, J., 2008. Child protection and adult crime: using investigator assignment to estimate causal effects of foster care. Journal of Political Economy 116, 746–770.
Duggan, M., 2005. Do new prescription drugs pay for themselves? The case of second-generation antipsychotics. Journal of Health Economics 24, 1–31.
Duggan, M., Imberman, S., 2006. Why are disability rolls skyrocketing? In: Cutler, D., Wise, D. (Eds.), Health in Older Ages: The Causes and Consequences of Declining Disability among the Elderly. University of Chicago Press.
Fevang, E., Markussen, S., Røed, K., 2011. The sick pay trap. IZA Discussion Paper No. 5655.
Franche, R.-L., Cullen, K., Clarke, J., Irvin, E., Sinclair, S., Frank, J., The institute for work & health workplace-based RTW intervention literature review research team, 2005. Workplace-Based Return-to-Work Interventions: A Systematic Review of the Quantitative Literature. Journal of Occupational Rehabilitation 15 (4), 607–631.
French, E., Song, J., 2009. The effect of disability insurance receipt on labor supply. Working Paper, Federal reserve Bank of Chicago.
Grytten, J., Sørensen, R., 2003. Practice variation and physician-specific effects. Journal of Health Economics 22, 403–418.
Hagen, E.M., Grasdal, A., Eriksen, H.R., 2003. Does early intervention with a light mobilization program reduce long-term sick leave for low back pain: a 3-year follow-up study. Spine 28, 2309–2316.
Hansen, J., 2000. The effect of work absence on wages and wage gaps in Sweden. Journal of Population Economics 13, 45–55.
Henrekson, M., Persson, M., 2004. The effects on sick leave of changes in the sickness insurance system. Journal of Labor Economics 22, 87–113.
Høgelund, J., Holm, A., McIntosh, J., 2010. Does graded return-to-work improve sick-listed workers' chance of returning to regular working hours? Journal of Health Economics 29, 158–169.
Ichino, A., Moretti, E., 2009. Biological gender differences, absenteeism and the earnings gap. American Economic Journal: Applied Economics 1 (1), 183–218.
Johansson, P., Palme, M., 2005. Moral hazard and sickness insurance. Journal of Public Economics 89, 1879–1890.
Kausto, J., Miranda, H., Martimo, K.-P., Viikari-Juntura, E., 2008. Partial sick leave — review of its use, effects and feasibility in the Nordic countries. Scandinavian Journal of Work, Environment & Health 34 (4), 239–249.

Maestas, N., Mullen, K., Strand, A., 2011. Does disability insurance receipt discourage work? Using examiner assignments to estimate causal effects of SSDI receipt. Rand Working Paper, WR-853-2.
Malmivaara, A., Häkkinen, U., Aro, T., Heinrichs, M.-L., Kosenniemi, L., Kuosma, E., Lappi, S., Paloheimo, R., Servo, C., Vaaranen, V., Hernberg, S., 1995. The treatment of acute low back pain — bed rest, exercises, or ordinary activity? The New England Journal of Medicine 332 (6), 351–355.
Manski, C.F., 1993. Identification of endogenous social effects: the reflection problem. The Review of Economic Studies 60 (3), 531–542.
Markussen, S., 2009. Closing the gates? Evidence from a natural experiment on physicians' sickness certification. Memorandum No. 19/2009. Department of Economics, University of Oslo.
Markussen, S., 2012. The Effects of Sick-Leaves on Earnings. Journal of Population Economics 25 (4), 1287–1306.
Markussen, S., Mykletun, A., Røed, K., 2010. The case for presenteeism. IZA Discussion Paper No. 5343.
Markussen, S., Røed, K., Røgeberg, O., Gaure, S., 2011. The anatomy of absenteeism. Journal of Health Economics 30 (2), 277–292.
Mykletun, A., Eriksen, H.R., Røed, K., Schmidt, G., Fosse, A., Damberg, G., Christiansen, E.C., Guldvog, B., 2010. Tiltak for reduksjon i sykefravær: Aktiviserings-og nærværsreform. Ekspertgrupperapport til Arbeidsdepartementet. English Summary available at: http://www.regjeringen.no/en/dep/aid/topics/welfare-policy/inclusive-working-life/summary-of-.html?id=592754.
OECD, 2003. Transforming Disability into Ability. OECD, Paris.
OECD, 2008. Are all jobs good for your health? The impact of work status and working conditions on mental health. Chapter 4 in OECD Employment Outlook 2008 . Paris.
OECD, 2010. Sickness, Disability and Work: Breaking the Barriers — Key Trends and Outcomes in Sickness and Disability. OECD, Paris.
Sieurin, L., Josephson, M., Vingård, E., 2009. Positive and negative consequences of sick leave for the individual, with special focus on part-time sick leave. Scandinavian Journal of Public Health 37, 50–56.
Van Oostrom, S.H., Driessen, M.T., de Vet, H.C.W., Franche, R.L., Schonstein, E., Loisel, P., Mechelen, W., Anema, J.R., 2009. Workplace Interventions for Preventing Work Disability (Review). The Cochrane Library.
Viikari-Juntura, E., Kausto, J., Shiri, R., Kaila-Kangas, L., Takala, E.-P., Karppinen, J., Miranda, H., Luukkonen, R., Martimo, K.-P., 2012. Return to work after early part-time sick leave due to musculoskeletal disorders: a randomized controlled trial. Scandinavian Journal of Work, Environment & Health 38 (2), 134–143.
Von Wachter, T., Song, J., 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), 3308–3329.
Waddell, G., 2004. The Back Pain Revolution, Second Edition. Elsevier Health Sciences.
Waddell, G., Burton, A.K., 2006. Is work good for your health and well-being? Report for Department for Work and Pensions. TSO (The Stationery Office), London.