Medical Certificates and Sickness Absence: Who Stays Away From Work if Monitoring Is Relaxed?

Yaroslav Yakymovych[*]

Institute for Housing and Urban Research, Uppsala University

February 2025

Abstract

Paid sick leave can cause absenteeism, which is controlled through costly monitoring by medical professionals. I study whether a targeted relaxation of monitoring can reduce costs without affecting absenteeism by leveraging a large-scale experiment which randomised Swedish workers into different monitoring regimes. Using machine learning, I identify systematic heterogeneity in the causal effects of monitoring on the absence of different worker groups. Individuals strongly affected by monitoring have high previous sick leave, low socioeconomic status, and are mostly men. A targeted relaxation of monitoring would be cost-efficient and would halve the increase in sickness absence compared to an untargeted relaxation.

Keywords: Sickness Absence; Monitoring; Heterogeneous Effects; GRF

JEL codes: I18; J38; C45

[*] I am grateful to Per Johansson and IFAU for data access, Cecilia Udin and the Swedish Social Insurance Agency for providing documentation regarding the experiment, and to Adrian Adermon, Susan Athey, Stefan Eriksson, Georg Graetz, Caroline Hall, Michael Lechner, Erica Lindahl, Oskar Nordström Skans, Stefan Pitschner, Lisa Simon, David Strömberg, Erik Sverdrup, Johan Vikström, seminar participants at Uppsala University, IFAU, IFN, Aarhus University, Swedish Social Insurance Inspectorate, Stanford GSB, SEW-HSG and conference participants at SUDSWEC, SNS, EALE, Swedish Conference in Economics and SHEA for their helpful suggestions and comments. Financial support from Vetenskapsrådet, grant number 2018-04581, is gratefully acknowledged. E-mail address: yaroslav.yakymovych@ibf.uu.se

1. Introduction

Paid sick leave is a right enjoyed by workers in most developed countries today, making it possible to stay at home when too sick to work. It shields workers from the economic effects of health shocks and allows them to smooth their consumption over time in a way that would otherwise be unfeasible. However, insuring workers against bad health carries a risk of moral hazard. Workers might stay at home for longer than their health warrants, leading to increased sickness absence costs and lost productivity. Ensuring that the system is fair is paramount for policymakers, especially in light of the substantial public spending on sickness and disability insurance, amounting on average to two percent of GDP in OECD countries in 2017 (OECD, 2021).

Monitoring, typically in the form of medical certificates, is used to make sure that recipients of paid sick leave are too sick to work. However, it is costly because of the high opportunity cost of medical professionals' time. If there is heterogeneity in terms of how monitoring affects the absence of different groups of workers, monitoring requirements can be relaxed for unaffected workers without incurring large increases in sickness absence. This can improve the efficiency of monitoring and reduce costs. Identifying how the sickness absence of different groups of workers responds to monitoring is thus of key relevance for policymakers.

In this paper, I investigate the sensitivity of different worker groups to monitoring using a large-scale randomised controlled experiment conducted in two Swedish regions. In the experiment, individuals were randomised into treatment and control groups based on whether they had odd or even dates of birth. Those with odd dates of birth were required to provide medical certificates if their sick leave spell was longer than seven days (as normal), while those born on even dates were only required to provide certificates if their spell exceeded 14 days. Register data make it possible to comprehensively characterise workers with respect to sickness absence-related factors such as sick leave history, socioeconomic status, family, career and place of residence.

I use a machine learning approach, the generalised random forest (GRF, Athey et al., 2019), to identify heterogeneous effects of the change in monitoring on the behaviour of different groups of workers based on this extensive set of characteristics. GRF has strong advantages compared to traditional sample splitting for studying treatment effect heterogeneity. The choice of characteristics and thresholds when splitting the sample is completely data-driven, minimising the potential for improper data mining. GRF also avoids overfitting predictions to statistical noise, only focusing on systematic predictors of treatment effect heterogeneity. This makes it possible for the analysis to include all characteristics which might ex ante be considered important drivers of sickness absence behaviour. GRF is also nonparametric, and is thus able to capture complex nonlinear functional forms and interactions.

I identify substantial heterogeneity in workers' sensitivity to monitoring. The least sensitive quartile of workers is estimated to increase the duration of their sick leave spells by 0.42 days on average, compared to 2.04 days for workers in the most sensitive quartile. The most important predictors of strong sensitivity to monitoring are a history of high sickness absence, low socioeconomic status (in terms of education, income and reliance on social payments), male gender, high sick leave among colleagues and partners, low socioeconomic status of the neighbourhood of residence, and weak attachment to the main job. For policymakers, sick leave history is particularly important, as it is observed by the insurer and might be seen as a fairer reason for differences in monitoring than other characteristics. The results regarding peer effects suggest that behaviour of colleagues is important and that interventions by firms to improve the work environment or morale can be useful.

Back-of-the-envelope calculations indicate that if monitoring is relaxed for workers who are estimated to be non-sensitive, rather than for a random subset of workers, losses in terms of increased sickness absence can be limited. Relaxing the monitoring regime for the half of workers with the smallest predicted treatment effects (rather than for a randomly selected half of workers as in the experiment) would result in absence rising by 51 percent less than was observed. If monitoring is relaxed based only on workers' sick leave history, absence would still increase 24 percent less compared to a random relaxation of monitoring. Based on comparisons of costs for healthcare provision and savings in terms of reduced absence, more stringent monitoring of all workers is estimated to be inefficient from a social point of view. However, monitoring the most sensitive workers, based on either the full GRF model or on high past sick leave, results in considerable social benefits. The break-even point is when the 81 percent of workers who are most sensitive according to the full model or the 57 percent of workers with the highest past sick leave are monitored more stringently.

There are relatively few earlier studies of the effects of monitoring on sickness absence, but a substantial literature has focused on identifying worker characteristics which are correlates of sickness absence (Paringer, 1983; Winkelmann, 1999; Barmby et al., 2002; Bratberg et al., 2002; Frick and Malo, 2008; Markussen et al., 2011; Treble and Barmby, 2011; Lindbeck et al., 2016). It is well-established that sickness absence is higher among women, public sector employees, low-paid workers, high-tenured workers, and employees at large workplaces. While I find that some factors associated with high sickness absence, such as having low earnings and living in a neighbourhood with high average sick leave, are correlated with high sensitivity to monitoring, this is not the case for a number of other factors. In particular, women and public sector workers are less sensitive to monitoring than men and private sector workers. Some other correlates of sick leave uptake, such as age, marital status and workplace tenure, do not have strong relationships with monitoring sensitivity. On the other hand, some factors which have not received much attention in the literature, such as having a low education and working in the manufacturing industry, are strongly associated with large effects of monitoring.

Ferman et al. (2021) and Boeri et al. (2021) study sensitivity to monitoring directly. Both focus on public sector workers, using a policy change in a Norwegian municipality and an experiment in Italy respectively. The studies arrive at opposing conclusions, with Ferman et al. (2021) finding no increase in sickness absence when medical certificate requirements are relaxed, and Boeri et al. (2021) finding that random visits to the homes of absent employees do have an absence-reducing effect.

Earlier work on the Swedish monitoring experiment has found that the relaxed rules caused a substantial increase in the duration of sickness absence spells. If applied nationally, the less stringent rules have been estimated to increase costs by about one billion SEK (200 million 2021 EUR), representing three percent of the costs of the sickness insurance system (Riksförsäkringsverket, 1989; Hartman et al., 2013). Some previous studies have investigated how the experiment’s effects differed based on worker characteristics such as age, gender and income (Hesselius et al., 2009; Hartman et al., 2013; Hesselius et al., 2013; Johansson et al., 2019). In this paper, the question is approached in a systematic way, by considering heterogeneity across a total of 56 individual characteristics. This is possible thanks to using the GRF instead of traditional approaches to studying treatment effect heterogeneity, as outlined above. The findings of earlier studies regarding higher sensitivity among men and individuals with low incomes are confirmed by my analysis. However, the key importance of sick leave history, social payment share in income, partner behaviour, and neighbourhood characteristics has not been identified previously.

There are large and persistent cross-country differences in levels of sickness absence (see, e.g., Lusinyan and Bonato, 2007), the reasons for which have not been conclusively identified. Explanations put forward include differences in monitoring regimes, replacement rates, workforce health, as well as cultural factors (Barmby et al., 2002). Differences in income tax rates across countries might also contribute to differences in absenteeism (Dale-Olsen, 2013). Evidence from Sweden (Johansson and Palme, 2002; Henrekson and Persson, 2004), Germany (Ziebarth and Karlsson, 2010) and Finland (Böckerman et al., 2018) suggests that absence rates are reduced if replacement rates are lower and the first day of sickness absence is excluded from insurance coverage. Although differences exist, Sweden's public corporatist sickness insurance system is broadly similar to those found in the rest of Scandinavia and continental Europe. Provisions in place at the time of the experiment, such as high replacement rates, lack of an unpaid initial waiting period and lack of monitoring for short absence spells were quite generous, but they are not unlike those prevailing in other European countries today (Palme and Persson, 2020). The results thus provide an insight into how monitoring affects worker behaviour in an institutional setting characteristic of many developed countries.

The issue of monitoring sickness insurance recipients received significant attention during the Covid-19 pandemic. Many countries relaxed rules for obtaining sick leave (OECD 2020). For example, in Sweden, the maximum period a worker could spend on sick leave before having to provide a medical certificate was increased from seven to 21 days during many phases of the pandemic (Försäkringskassan, 2021). This relaxation of monitoring intensity was very similar in spirit to the changes during the experiment I study.

The remainder of the paper is structured as follows. Section 2 provides background about the Swedish sickness insurance system and the context in which the monitoring experiment took place. The sickness absence outcome, as well as the worker characteristics considered as possible drivers of treatment effect heterogeneity, are covered in Section 3. An overview of the machine learning approach used to identify conditional treatment effects is given in Section 4. Section 5 provides evidence that the experimental randomisation was successful and Section 6 presents and discusses the results. Section 7 provides simple policy-relevant calculations and Section 8 concludes.

2. Background

a. The Swedish Sickness Insurance System

Sweden has a comprehensive sickness insurance system, where practically all employees are entitled to paid sick leave. At the time of the experiment in 1988, recipients were entitled to payments from government-run social insurance funds starting from the first day of absence. Normally, workers were required to provide the insurance funds with medical certificates proving that they were sick if the duration of absence was eight days or longer. Workers were reimbursed 90 percent of their wages while they were on sick leave. However, benefits were capped for workers whose annual earnings were greater than 193,500 SEK in 1988 (equivalent to about 69,000 2020 US$). About 7.8 percent of the workers involved in the monitoring experiment had earnings in excess of this cap.[1] There was no time limit on benefit duration. Leave for taking care of sick children was, and has remained, separate from sickness absence in the Swedish system. The rules for such leave were unaffected by the 1988 experiment (Riksförsäkringsverket, 1989).

Since 1988, a number of changes have been enacted to the system, mostly with the aim of reducing moral hazard and overuse. Replacement rates have been reduced to 80 percent of wages, limits on the maximum duration of sickness absence benefits have been introduced, recipients are no longer reimbursed for the first day of sick leave (the "qualification day") and the first two weeks of sick pay are now paid by employers rather than the public insurance system. However, the Covid-19 pandemic brought about a loosening of some rules. Most notably, the qualification day rule was not applied and recipients were reimbursed for all their days of sick leave. Also, monitoring in the form of medical certificate requirements, which had been required from the eighth day of absence, only took place from the 21st day of absence (Försäkringskassan, 2021).

b. The 1988 Experiment

Starting from January 1987, Jämtland county in northern Sweden implemented a policy which allowed workers to be on sickness leave for 14 days without providing a medical certificate, instead of the usual seven days. This affected all workers, without the date of birth differences later introduced by the experiment. Motivations for the new policy included reducing the burden on the medical system, potential positive effects on worker's long-term health, and the large distances to medical establishments in rural Jämtland (Riksförsäkringsverket, 1989). The new policy was evaluated in an experiment, which ran between July and December 1988, and involved Jämtland (70,000 sickness insurance recipients) and Sweden's second largest city Gothenburg (240,000 recipients). During the experiment, individuals born on odd dates were required to provide medical certificates starting on the eighth day of their absence spell, while those born on even dates were required to provide certificates starting on the fifteenth day.

[1] For some workers, such as municipally employed workers and some privately employed white-collar workers, an additional amount was paid by their unions. For these individuals, the replacement rate of sickness insurance could amount to 100 percent for short to intermediate spells. This also mitigated losses for those with earnings in excess of the reimbursement cap. The rules for providing medical certificates for these additional reimbursements were also changed in line with the experiment (Riksförsäkringsverket, 1989), meaning that affected individuals faced no asymmetric incentives.

Thus, the experiment represented a loosening of the rules for some individuals in Gothenburg and a tightening of the rules for some individuals in Jämtland. However, in line with norms prevailing nationally, I define the **control group** as those in both regions who provided certificates on day eight, and the **treated group** as those in both regions who provided certificates on day fifteen. There was a substantial campaign to inform workers about the experiment, and evaluators subsequently assessed workers' understanding of the experimental rules as very good (Riksförsäkringsverket, 1989).

The average duration of sickness absence spells among the treated group was substantially higher than among the control group during the experiment, and the standard rules for medical certificates were reinstated for everyone in Jämtland and Gothenburg for spells which began on January 1989 or later. Evaluators later estimated that the less stringent rules would have increased sickness insurance costs by 3 percent if applied nationally (Riksförsäkringsverket, 1989). The findings of Hartman et al. (2013) confirm this, showing substantially longer average spell duration for the treated group, as well as behaviour indicative of moral hazard.

Workers employed by the central government (11 percent of all workers in Jämtland and Gothenburg) were excluded from the experiment for administrative reasons (Riksförsäkringsverket, 1989). This category included teachers, postal workers, railway employees, police, Church of Sweden clergy, and others. The experiment however applied to local government workers (34 percent of all workers).

3. Outcomes and Characteristics

Thanks to unusually rich microdata collected by Statistics Sweden and the Swedish Social Insurance Agency, I am able to include a broad set of worker characteristics in the analysis. These contain information on sickness absence (from the Register of Sickness Absence Cases), demographic characteristics, place of residence, family, employment, and earnings (all from the Louise/Sys registers).

a. Outcomes

Earlier work by Hartman et al. (2013) has found sizeable effects of the experiment on the duration of treated workers' sickness absence spells, but no evidence of an effect on sickness spell incidence. Because of this, I focus on the intensive rather than the extensive margin.

The main outcome studied is the duration of sickness absence spells in days. This is a natural margin to consider, as costs to the employer and insurer scale with absence duration. Workers in Jämtland and Gothenburg had a total of 261 127 sickness spells during the experiment. In the main analysis, I exclude spells whose duration makes it unlikely that they were affected by differences in monitoring between days 7 and 14. As explained in Section 5.a, there are no noticeable differences in behaviour between treated and controls for spells shorter than four and longer than 21 days. These two categories comprise 111 020 and 21 305 spells respectively. Including long spells is problematic, as they might have outlier effects on estimates. The very large number of short spells would also conceal patterns of behaviour during the period when monitoring intensity varied. For this reason, the main analysis only uses spells between four and 21 days in duration.

To ensure that this choice does not materially affect the results, I perform sensitivity analysis using spells of all durations, but with the duration of spells longer than 30 days set to 30 to avoid outlier effects. Results are similar as shown in Table B2 in Appendix B and explained in Section 6.b. Another alternative outcome definition is the probability of a sickness absence spell ending during its second week. For this outcome, I retain spells shorter than four and longer than 21 days, as outlier effects are not a concern. As explained in Section 6, the findings are also similar to those in the main analysis.

b. Worker and Spell Characteristics

The selection of worker characteristics into the analysis is based on factors which have been identified as important correlates of sickness absence by previous literature. A total of 56 variables are included. Most of them have not been directly linked to monitoring sensitivity. Nevertheless, a simple hypothesis would be that groups of workers with high sickness absence also react more strongly to being monitored.

Health-Related

An individual's health status is, in the absence of moral hazard and reporting costs, the only determinant of sick leave duration. Unfortunately, the data do not allow me to fully characterise individuals' health.[2] However, I am able to use two indirect measures of individual health. The first of these, *the total number of previous days of sickness absence*, is based on the individual's sickness absence in the two and a half years before the experiment.[3] This measure contains information not only on the individual's health, but also on any overuse of sickness insurance. Another variable connected to both health and sickness absence behaviour is the *number of short sickness absence spells in earlier periods.* Short spells are defined as those 1-21 days in

[2] Information about inpatient care spells and diagnoses received during such spells is only available from January 1987. Data on outpatient care contacts and diagnoses, which constitute the vast majority of cases, are unavailable for the period studied.

[3] The length of the pre-period is dictated by data on sickness absence becoming available from January 1986.

length. This measure puts less weight on long spells, instead focusing on whether the individual has been on many short absences in the past, which might be indicative of misuse. The number of short spells is likewise measured in the two and a half years before the experiment. A metric much more directly related to serious health issues is *the total number of days spent in inpatient care* in the one and a half years before the experiment.

Demographic

There are well-established differences between demographic groups in terms of their sick leave uptake. For example, a *female* dummy is included because women tend to have a higher uptake than men (Paringer, 1983). Also, health deteriorates with *age*, which has been shown to affect sick leave (Barmby et al., 2002). Finally, in the Scandinavian setting, immigrants tend to have higher rates of sickness absence more than natives (Markussen et al., 2011; Helgesson et al., 2015). This factor is captured using an *immigrant* variable which takes the value 0 for individuals born in Sweden, 1 for those born in other Scandinavian countries, 2 for those born in the rest of Europe and 3 for those born in the rest of the world. The GRF is able to handle ordinal variables such as this, unlike regression-based methods.

Family

Family factors have been found to play a role in workers' sickness absence. The presence of partners may affect behaviour through an additional source of income, and married individuals tend to have higher sick leave uptake than unmarried ones (Barmby et al., 2002; Angelov et al., 2011). To analyse the importance of such effects, I include dummies for being *married*,[4] *divorced*, *single*, and *widowed*. Furthermore, the *individual's share of household income* directly measures the importance of being insured by a partner's income.

Having children is connected to higher sickness absence, especially for women (Bratberg et al., 2002; Angelov et al., 2013). The presence of children is captured by the variables *number of children younger than 18* and the *age of the youngest child*. As data on children start only in 1990, I impute information for 1988 by subtracting two years from children's ages in 1990. While leave for taking care of sick children is separate from sick leave in Sweden, parents might nevertheless register such spells as own sickness absence. For this reason, I include the *number of days spent taking care of sick children* in the two and a half years before the experiment.[5] The *share of the family's sick child days* variable captures the intra-household division of

[4] Including those cohabiting with a partner with common children.

[5] Data on leave for taking care of sick children begin in January 1986.

childcare responsibilities. For those without children, age of the youngest child and share of family's sick child days are set to missing. The GRF deals smoothly with missing values by grouping them in turn with those with high values of the covariate, those with low values of the covariate and as a separate group when evaluating the splitting criterion. Methodological details in this regard are provided in Section 4.

Education

Education is a strong correlate of factors identified as important for sick leave uptake, such as earnings and occupation, and may also have an independent effect on sickness absence (Piha et al., 2010). To flexibly capture education, I include *years of education*, as well as dummies for broad education fields. The fields are *general education* (found at the low levels of educational attainment), *teacher training*, *administration/law/social science*, *science/engineering*, *health* and *services*.

Neighbourhood

There is evidence that individuals can be affected by the sickness absence attitudes and behaviour of their neighbours (Lindbeck et al., 2016). For this reason, I include several leave-one-out characteristics of the neighbourhood where the sickness insurance recipient lives. The neighbourhoods (defined by Statistics Sweden) are small, corresponding to several urban blocks or small portions of the countryside. Neighbourhood characteristics included are *average annual earnings*, *average share of social payments in income, share of inhabitants with a post-secondary education* and the *immigrant share*. These four measures are constructed based on the population aged 30-64, not taking into account those past working age, or those who are likely to not have completed their education.

The costs of obtaining a medical certificate increase with the *distance to the nearest doctor*. I measure this as the Euclidian distance between the worker's neighbourhood and the neighbourhood of the nearest establishment in the medical industry.[6]

Career-Related

Individuals with high earnings have lower sickness absence compared to those with lower earnings. This could be due to better health, stronger intrinsic motivation, or lower income replacement rates (Barmby et al., 2002; Böckerman et al., 2018).[7] These effects are captured by

[6] I am not able to differentiate between primary clinics, which can provide certificates in cases of mild illness, and other establishments. Thus, there is some measurement error in this variable.

[7] 7.8 percent of individuals with high annual earnings had replacement rates below the standard 90 percent. At least some of these workers were however likely to be (partly) reimbursed by additional union-negotiated sickness insurance. For evidence on sickness absence being affected by replacement rates, see Johansson and Palme (2005).

an *annual labour income* variable. A related concept is the worker's *income rank at his or her workplace*. The rank is measured in relative terms, with 0 representing the worker who earns least and 1 the worker who earns most regardless of workplace size. This measure also captures key worker effects, which imply that workers who are more important for workplace functioning are less likely to be absent. This might be because they continue working even when they are sick, or because individuals with better health tend to select into such roles (Hensvik and Rosenqvist, 2019). Workplace *tenure* has been identified as a correlate of sickness absence, with high-tenured workers tending to take more sick leave than lower-tenured ones (Barmby et al., 2002). Tenure is also correlated with job security, which has been suggested to increase sickness absence (Bratberg and Monstad, 2015). The tenure measure goes from 0 to 3 years and is censored at the top because matched employer-employee data are only available from 1985. The *share of income from the main job* provides a measure of the worker's commitment to his or her main place of work.

Self-employed workers are eligible for paid sick leave in Sweden; in the absence of penalties from employers or colleagues, they have incentives to be absent for longer than other groups. However, a number of studies have found that absenteeism is lower among the self-employed than among other workers (Lechmann and Schnabel, 2014) and that moral hazard might be less of an issue for the self-employed (Spierdijk et al., 2009; Baert et al., 2018). Differences between self-employed and other workers are captured by the *share of income from self-employment*. Many Swedes, even among those who work, receive some social payments, such as child benefits. The importance of these as a source of income relative to earnings is captured by the *social payment share in income*.

Workplace-Related

Different sectors of the economy have different sick leave rates (Barmby et al., 2002). This might be due to intrinsic differences in workforce characteristics, such as gender and age composition, differences in the working environment which translate to employee health, or because some sectors are more permissive of sickness absence overuse. The public sector has higher sickness absence rates than the private sector in many countries (Frick and Malo, 2008); for this reason, I include a *local government sector* dummy. In 1988, the Swedish local government sector included healthcare, elderly care, municipal services and administrative staff. Central government employees were excluded from the experiment and are not a part of this study.

Differences between sectors are further captured by nine broad industry dummies: *primary*, *manufacturing*, *construction*, *utilities*, *wholesale and retail*, *business services*, *health*, *education* and *public administration*.

The *number of workers at an establishment* has been suggested to affect sick leave uptake. This could be both because large workplaces are worse for employees' health and because the importance of a single individual decreases with workplace size, meaning that costs of unnecessary absence spells are lower (Winkelmann, 1999, Lindgren 2012).

Finally, I include the *distance to work*, measured as the Euclidian distance between the worker's neighbourhood and the neighbourhood of the workplace. This captures costs of getting to work, which might induce individuals to stay at home (van Ommeren and Gutiérrez-i-Puigarnau, 2011).

Peer Effects

I consider three kinds of peer effects. The first is colleagues' behaviour, measured as *colleagues' average number of sickness absence days* and *colleagues' average number of short absence spells* in the two and a half years before the experiment. The second set of peer effects relates to neighbours' behaviour, consisting of *employed neighbours' average number of sickness absence days* and *employed neighbours' average number of short spells* before the experiment. This measure is based on those aged 30-64 like the other neighbourhood variables. Finally, peer effects within families are captured by the *partner's number of sickness absence days* and *partner's number of short absence spells* before the experiment.[8]

To capture behavioural effects of the experiment, the *share of colleagues treated* as well as a dummy for the *partner being treated* are included. The share of treated colleagues has been found to be important by Johansson et al. (2019).[9]

Aggregate Characteristics

The *population density of the municipality where the individual lives* is intended to capture any differences between areas with different levels of urbanisation. Relative sick leave uptake between urban and rural areas in Sweden has varied over time and even reversed (Haugen et al., 2008). Given the GRF's nonparametric nature, this variable also nests regional differences

[8] For those at single-worker workplaces, colleague peer effects are missing, and for those without a partner, partner peer effects are missing. The GRF is able to handle such missing values well, as discussed in Section 4.

[9] If colleagues' or partners' treatment status affects behaviour, SUTVA is violated, and this would have to be taken into account when designing a targeted monitoring policy. However, these two variables have little impact on GRF estimates of monitoring sensitivity, suggesting that the role of such spillovers is limited. For this reason, spillovers are not taken into account in the policy analysis in Section 7.

between Gothenburg and Jämtland, as Gothenburg had a much higher population density (963 people/km^2) than any municipality in Jämtland (at most 26 people/km^2). The variable thus also captures any effects of the experiment's different directions in Gothenburg (reduction in monitoring intensity for some individuals) and Jämtland (increase in monitoring intensity for some individuals).

Spell Characteristics

The variables above are identical for all spells taken by the same worker. However, there is seasonal variation in sickness absence (documented since at least Watson, 1927), which is captured by the *spell's starting day* relative to July 1st, 1988. Seasonal sickness absence fluctuations are at least in part driven by higher respiratory infection rates during the cold months of the year. In addition, shirking patterns may vary over the course of the year in response to changes in the opportunity cost of working (in line with Skogman Thoursie, 2004).

It is also possible that workers become more aware of the new monitoring rules as they take out sickness absence spells. This is captured by the worker's *number of previous spells during the experimental period*.

4. Empirical Approach

This section outlines and motivates the GRF approach used for estimating the heterogeneous effects of monitoring, and describes its implementation. A more detailed description of GRF can be found in Appendix A.

The goal is to identify the effects of monitoring on the sickness absence behaviour of different groups of workers, and whether it varies across their absence spells. Given the experimental setting, differences between treated and control spells in a group defined by a set of attributes should be considered causal effects of monitoring within this group. However, which particular characteristic drives the size of the treatment effect is ambiguous, as relevant characteristics can be correlated and interact with each other. Therefore, the focus is on identifying *predictors* of workers' sensitivity to monitoring.

Unlike traditional heterogeneity analysis, the GRF (Athey et al., 2019) used in this paper is fully data-driven. It splits the sample based on the variables and threshold values that maximise treatment effect heterogeneity without input from the researcher. This avoids the subjective choices and data mining concerns that otherwise arise when testing for heterogeneity based on a large number of covariates. Furthermore, GRF identifies the variables and thresholds which are the strongest predictors of treatment effect heterogeneity, and which the researcher might otherwise miss. Only systematic relationships which hold across different subsamples of the data are taken into account, avoiding overfitting by ignoring irrelevant characteristics. This means that all available characteristics which are *ex ante* expected to be related to sickness absence behaviour can be included in the analysis.

GRF is fully nonparametric, splitting the sample at the covariate and threshold value which maximise treatment effect heterogeneity. The resulting subsamples are split according to the same principle (based on the covariate and threshold value which maximise treatment effect heterogeneity in that particular subsample) until the observations have been grouped together with others who have similar treatment effects. This procedure is repeated multiple times, based on a random sample of observations each time. This ensures that relationships which hold across different subsamples are given the most weight. The final GRF prediction combines the output of models ("trees") estimated on each of the random samples.

Because it is nonparametric, GRF is able to capture complex functional forms and interactions between variables. It is also able to handle ordinal variables and missing values.[10] Because of this procedure, it is possible to include features of workplaces and partners in the analysis without dropping individuals who work at one-worker workplaces or are single.

Prior to estimating treatment effects, I use GRF to predict observations' outcome and treatment propensity in order to control for selection into treatment. Prediction of the outcome is based on the spell and worker characteristics listed in Section 3.b. and is done for sickness absence spells between 4 and 21 days in length in the main analysis. Estimation of treatment propensity is based on the characteristics of workers, not spells, and uses all workers resident in Gothenburg and Jämtland, including those who did not have any sickness absence spells during the experiment.[11] The model reveals no differences in treatment propensity based on worker characteristics, which confirms the quality of the experimental randomisation, as discussed in Section 5.a.

To make certain that the GRF model is well-calibrated, I split the data into a training set containing 80 percent of observations and a held-out test set containing 20 percent of observations. The model is constructed using the training set, and the quality of its predictions

[10] The exact numerical value of a variable does not matter, but only whether a split was made, which makes it possible to use ordinal variables. When evaluating splits, observations with missing values are grouped in turn with high values and low values of the variable, as well as placed in a group of their own.

[11] Only workers who fulfil the inclusion criteria listed in Section 5.a. are included in the treatment propensity model. As randomization takes place on the worker level, the spell-specific characteristics (spell start date and spell order) are not included in the treatment propensity model.

assessed using the test set, which is not used in model training. Splitting into the training and test sets, as well as selecting samples for different "trees", is based on family clusters. Estimation for the training set is done out-of-bag (based only on "trees" into which the observation's family cluster was not sampled). This means that predictions are based on the absence spells of individuals from other families, and do not fit idiosyncrasies in the behaviour of particular individuals or their partners.

I estimate my model using the `grf` package in R. Several model parameters can be tuned by cross-validation, and this is implemented in the baseline model. However, using the tuned parameters gives very similar results to using the default parameters, as shown in Table B2 in Appendix B.

GRF has been shown to perform competitively compared to other machine learning methods by Knaus et al. (2021). Key reasons for selecting GRF for my analysis are its flexibility with regard to functional form and ability to handle missing values. A parametric machine learning method such as LASSO entails making at least some functional form assumptions and excluding characteristics whose values are missing for some workers (or dropping the associated workers from the analysis). Nevertheless, I estimate a LASSO model based on a subset of workers and covariates (as well as their higher-order terms and interactions), as explained in Appendix A, and compare the results in Tables B2 and B3 and Figure B14 in Appendix B. The GRF and LASSO models' predictions have a correlation of 0.73. Although both perform well in predicting monitoring sensitivity out-of-sample, the results indicate that GRF is able to classify workers and spells better. An unpenalised OLS model performs clearly worse in the test set than either GRF or LASSO.

5. Randomisation and Balancing

a. Experimental Population and Validity of Randomisation

To be eligible for the experiment, workers had to live in Gothenburg or Jämtland and not be employed by the central government. Location of residence is observed on an annual basis. To make sure that all workers included in the analysis were exposed to the experiment for its entire duration, I drop those who lived in another region in 1987 or 1989. Those employed by the central government during 1988 are also excluded. I also exclude workers under the age of 18 and those with very low annual labour earnings.[12] This leaves 125 541 workers who took 261

[12] The reason for excluding those marginally attached to the labour market is that I want to focus on the behaviour of those whose main source of income is labour earnings. While unemployed and marginally employed individuals are eligible for paid sick leave in Sweden, their behaviour is likely to be influenced by other factors than that of

127 sickness absence spells of any length.[13] After spells shorter than four and longer than 21 days are excluded, 128 802 spells taken by 82 011 individuals remain.

While date of birth considered over the entire year is correlated with many important characteristics and outcomes (see e.g. Bedard and Dhuey, 2006), having an odd or even date of birth is random, as parents are unlikely to be able to determine the exact timing of birth.[14] Swedish social insurance numbers, used for reporting sick leave, include the birth date, meaning that manipulations in response to the experiment would have been prohibitively costly. The lack of differences between the behaviour of workers born on odd and even dates before the experiment is confirmed by the graphs in Figure 1.[15] The left panel plots the survival curve, identifying the share of sickness spells still ongoing a given number of days after their start date. There are no visually discernible differences between workers born on odd and even dates. A fairly sharp drop in the survival rate is evident after seven days of absence, when workers in Gothenburg (77 percent of the sample) were required to provide medical certificates. There is a smaller drop at 14 days of absence, when workers in Jämtland were required to provide certificates. These drops are confirmed by the hazard graph in the right panel of Figure 1, which shows the probability of a spell which has been going on for a given number of days ending on the next day. Like the survival curve, the hazard rates also show very similar patterns of behaviour for individuals with odd and even dates of birth.

A more formal test of the randomisation is provided by comparing treatment propensities (estimated by GRF based on all included worker characteristics) for treated and control workers. Histograms of these are shown in Figure 2. The average treatment propensity is 0.49, reflecting the fact that there are slightly fewer even than odd dates. All sickness insurance recipients' propensity scores lie within 0.08 of the mean, and the distributions are very similar for both treated and controls. GRF is thus unable to predict selection into treatment, in spite of being

employed individuals. From the point of view of public finances, paid sick leave for these groups involves shifting costs between different support mechanisms, which makes cost-benefit estimations difficult. If marginally attached workers are included, their sickness absence duration is estimated to be highly sensitive to medical certificate requirements, comparable to that of the most sensitive workers in the final sample.

[13] 86 757 eligible workers took no sickness absence during the experiment.

[14] There are no laws or other policies in Sweden that have different effects based on date of birth, so there are no incentives to manipulate the date outside of the setting of the 1988 experiment.

[15] Figure 1 is based on the second half of 1987, that is the part of the year that corresponds to the 1988 experiment. Results for the first half of 1987 and for the first half of 1988 closely align with those for the second half of 1987 and likewise reveal no differences in the behaviour of treated and control workers.

able to capture complex interactions between different variables. A balancing table for all covariates included in the analysis is shown in Table B1 in Appendix B.[16]

b. Main Effect of the Experiment

Earlier evaluation of the experiment (Riksförsäkringsverket, 1989; Hartman et al., 2013) has shown a sizeable effect on the duration of sickness absence spells. Hartman et al. (2013) find that the average duration of sickness absence spells among treated workers increased by 0.6 days, but no evidence that the incidence of absence spells per worker responded to the experiment. In Figure 3, survival and hazard rates for sickness spells which began in the second half of 1988 are shown. There are striking differences in the behaviour of the treated and controls, which were absent in the pre-period (Figure 1). The survival curve for the treated is consistently above the one for the controls between days 6 and 14. The gap exists while monitoring intensity differs, suggesting that it is indeed caused by the discrepancy in monitoring rules. The experiment's impact is confirmed by the hazard graph, which shows large spikes in the probability of exiting sick leave at 7 days for the treated and 14 days for the controls. Interestingly, there is no evidence that the experiment continued to affect worker behaviour after it went out of effect. Figure B1 in Appendix B shows survival and hazard rates among workers in Gothenburg and Jämtland in the first and second halves of 1989. Those born on odd and even dates behave identically in both post-experiment periods.

6. Results

a. Extent of Heterogeneity

The distribution of estimated heterogeneous effects of the relaxed monitoring rules on spell duration is shown in Figure 4. Virtually all spells are predicted to become longer if monitoring is reduced; this highlights the nonparametric nature of the GRF, as a regression-based model would have been likely to provide theoretically unreasonable negative treatment effect values for some observations. The median effect is 1.01 days,[17] but there is substantial variation in estimated sensitivity to monitoring across workers and spells. For the least sensitive decile, treatment effects are estimated to be at most 0.47 days, while for the most sensitive decile they

[16] The treated and control groups are very similar with regard to all the characteristics considered, but a few numerically small differences are statistically significant because of the large sample size. This is because non-European immigrants (2.6 percent of the control group and 2.3 percent of the treated group) are slightly more likely to be registered as born on odd dates, particularly January 1st, due to uncertainty about the true birth date. If non-European immigrants are removed from the sample, the means of two variables are statistically different at the five percent level, in line with what is expected when testing for differences in means of 56 variables following a successful randomisation. No difference in means is likely to be economically significant.

[17] This figure is larger than the 0.5-0.7 days found by Hartman et al. (2013) because I drop spells shorter than four and longer than 21 days, whereas they include spells of all durations (censoring spells longer than 28 days).

are estimated to be 1.67 days or more. A corresponding histogram of treatment effects where the outcome is the probability of returning to work on days 8-14 is shown in Figure B2 in Appendix B. The median spell is estimated to be 12 percentage points more likely to end in the second week of absence if the worker is treated. This is a very sizeable effect, as the baseline probability for spells of control group workers is eight percent. There are also large heterogeneities across groups of spells; the effects are estimated to be smaller than six percentage points for the least sensitive decile and larger than 22 percentage points for the most sensitive decile.

The model's performance can be assessed using an omnibus best linear predictor test (Chernozhukov et al., 2020). The best linear predictor test assesses whether both the average treatment effect and variations around this average effect are predicted correctly. The test uses GRF's estimates of treatment effects $\hat{\tau}_x$, predictions of spell duration $\widehat{y_x}$ and estimates of treatment propensity $\widehat{e_x}$ to estimate the following regression (where y_i is the spell's observed duration and W_w is the worker's treatment status):

$$y_i - \widehat{y_x} = \alpha\bar{\hat{\tau}}_x(W_w - \widehat{e_x}) + \beta\big(\hat{\tau}_x - \bar{\hat{\tau}}_x\big)(W_w - \widehat{e_x}) + \varepsilon_i, \qquad \bar{\hat{\tau}}_x = \frac{\sum_i \hat{\tau}_x}{N}$$

The parameter α estimates how well the average predicted treatment effect fits the data and the parameter β measures whether heterogeneity in treatment effects is correctly captured. For the baseline model with spell duration as the outcome, $\alpha = 1.01$ $(SE = 0.02)$ and $\beta = 1.38$ $(SE = 0.05)$. In the model where the outcome is the probability of returning to work in the second week, $\alpha = 0.99$ $(SE = 0.01)$ and $\beta = 1.34$ $(SE = 0.03)$. Both α and β are close to one, indicating that the estimates adequately capture the average effect of reduced monitoring, as well as deviations from this average in different worker and spell groups. The null hypothesis of no heterogeneity (i.e. $\beta = 0$) is strongly rejected at conventional levels of significance. However, the models somewhat underfit the true heterogeneity in treatment effects, compressing estimates of $\hat{\tau}_x$ to the mean.

Because of this, I focus on using the estimated $\hat{\tau}_x$ for ranking absence spells into quantiles of predicted sensitivity to monitoring, and estimating the treatment effect within each quantile using other methods. Due to experimental randomisation, a valid approach is to simply estimate the difference in the duration of treated and control spells within each quantile of $\hat{\tau}_x$. A more sophisticated approach, which makes use of the GRF estimates of $\widehat{y_x}$, $\widehat{e_x}$ and $\hat{\tau}_x$, involves estimating augmented inverse propensity weighted (AIPW) scores (Robins and Rotnitzky, 1995). The average of the AIPW scores in a quantile of $\hat{\tau}_x$ is a doubly robust estimate of the average treatment effect in that quantile, being consistent as long as either the outcome prediction model or the treatment propensity model is correct (but not necessarily both).

I subdivide the spells into four quartiles, with Quartile 1 containing those with the smallest $\hat{\tau}_x$ and Quartile 4 containing those with the largest $\hat{\tau}_x$. Figure 5 shows the average $\hat{\tau}_x$ within each quartile, the mean treated-control difference in duration along with the 95 percent confidence interval, and the average AIPW score. As expected, treatment effects according to each of these measures increase when moving up the quartiles. This confirms that the $\hat{\tau}_x$ estimates correctly rank the spells in terms of sensitivity to monitoring, in spite of being somewhat compressed towards the mean in the upper tail. The simple treated-control difference and the average AIPW score are almost identical in all cases, suggesting that correction for selection into treatment is not necessary, as expected in a randomised experiment. The average treatment effect is quite small for spells in Quartile 1, 0.42 days, compared to a sizeable 2.04 days for spells in Quartile 4. Estimated treatment effects in each quartile are distinct at the 95 percent confidence level from those in other quartiles.

To confirm that the GRF model identifies persistent relationships between worker and spell characteristics and sensitivity to monitoring, I use it to predict treatment effects for absence spells of workers from the held-out test set. I then rank the test set spells by their $\hat{\tau}_x$ and divide them into quartiles analogously to the training set. The survival and hazard rates of spells in each of these test set quartiles are shown in Figure 6. The graphs confirm that the model correctly identifies spells with different responsiveness to monitoring. The survival curves for treated and control spells in Quartile 1 align quite closely. A gap between the two groups appears around day 7, but, crucially, it closes completely already before day 14. The maximum difference in the shares of treated and control spells which are still ongoing is 0.11, on day 8. The gap between treated and controls is wider in the higher quartiles of predicted treatment effects. For those in Quartile 4, a large gap opens up already on day 6, and does not close until day 15. The maximum difference between shares of treated and control spells which are still ongoing is 0.38, on day 8.

The hazard rate spikes on days 7 and 14 for all groups of spells, but the size of this spike is larger in the higher quartiles. The hazard rate for control spells in Quartile 1 on day 7 is 48 percent, while the hazard rate for treated spells on day 14 is 50 percent. For spells in Quartile 4, the corresponding rates are 54 percent on day 7 for control spells and 72 percent on day 14 for treated spells. This is further evidence that differences in behaviour between the quartiles are driven by different responsiveness to monitoring.

Overall, Figure 6 confirms that my GRF model identifies relationships between worker characteristics and monitoring which hold across different groups of workers in the data. To test external validity further, I implement a robustness check where I train a model on a training set consisting only of the spells of workers from Gothenburg. The correlation between the predictions of this model and of the baseline model is 0.94 for the Gothenburg sample. Then, I use this model to predict the sensitivity of absence spells of workers in Jämtland. For them estimates from the Gothenburg-based model are also very similar to those of the baseline model, with a correlation coefficient of 0.92. I test the validity of the Gothenburg-based predictions for workers in Jämtland through exercises analogous to those in Figures 5 and 6, with results presented in figures B3 and B4 in Appendix B. They confirm that patterns of behaviour observed in Gothenburg can be used to infer the monitoring sensitivity of workers and absence spells in Jämtland. The relationships between individual and spell characteristics and monitoring sensitivity are thus stable across two very different Swedish regions. Hence, the model's predictions are at least somewhat externally valid.

b. Predictors of Monitoring Sensitivity

To form hypotheses about why workers behave differently, it is necessary to know which worker and spell characteristics predict monitoring sensitivity. Characteristics of sensitive and non-sensitive individuals can provide evidence in favour of or against mechanisms such as health, replacement rates, career ambitions, irreplaceability in the workplace, family responsibilities, social conscientiousness, peer influence and learning about the new rules over time. Furthermore, policymakers are unlikely to have information about all the characteristics included in the GRF model's training, and some characteristics are unfeasible to use for moral or legal reasons. If a given characteristic or set of characteristics is a strong predictor of monitoring sensitivity, it can be used as an approximation of the full model.

In this part of the analysis, I continue dividing the sample into four quartiles, with Quartile 1 containing absence spells with the smallest $\hat{\tau}_x$ and Quartile 4 containing absence spells with the largest $\hat{\tau}_x$. Differences between the characteristics of workers and spells in Quartile 4 and Quartile 1 are presented in Figure 7. The left-hand panel shows differences in terms of continuous characteristics, which have been normalised by their mean and standard deviation. The panel on the right shows dummy characteristics, with the bars representing differences between the two quartiles in terms of the share who have the given characteristic. Both panels are ordered from the smallest difference to the largest; variables with higher values among non-sensitive workers and spells are at the top, and those with higher values among sensitive workers and spells are at the bottom.

Several factors stand out as strong predictors of monitoring sensitivity. Being sensitive to monitoring is associated with high past sick leave uptake, low socioeconomic status, male gender, weak workplace attachment, and having peers with high sick leave uptake.

Both the number of days of sick leave and the number of short sickness absence spells in the past predict monitoring sensitivity. On average, workers in the most sensitive quartile took 125 days of sick leave and 8.8 short sickness absence spells in the two and a half years before the experiment. The corresponding figures for workers in the least sensitive quartile are 37 days of sick leave and 6.1 short absence spells. The strong association between past sickness absence and sensitivity to monitoring could be due to workers in Quartile 4 having poorer health, or due to relaxed attitudes to shirking among this group.

Responsive workers have lower socioeconomic status than non-responsive workers. This is most apparent in the higher share of social payments in their income (23 percent versus six percent), lower education (9.4 years versus 11.6 years) and lower earnings (84 thousand SEK versus 118 thousand SEK).[18] Immigrants are also overrepresented among sensitive workers. Lower sensitivity among workers with high earnings is in line with the results of Hartman et al. (2013), but the share of social payments in income, which they do not study, turns out to be a stronger predictor of sensitivity. These patterns might be driven by causal effects of socioeconomic status, or by its correlation with factors such as health.

A somewhat surprising finding is that women’s sickness absence is much less sensitive to monitoring than men’s. It is well-established that women on average take out more sick leave than men, and that sickness absence is more prevalent in the public sector, where women are overrepresented (see, e.g., Paringer, 1983, and Frick and Malo, 2008).[19] Furthermore, uneven division of family responsibilities and childcare might give women stronger incentives to use sickness absence to gain time for work in the home. The results show no evidence in favour of such hypotheses. On the contrary, women constitute 69 percent of the least sensitive quartile and only 27 percent of the most sensitive quartile. This is in spite of past sick leave, which is higher among women, being strongly associated with monitoring sensitivity. As women are concentrated in the public sector, public sector employees are on average not sensitive to

[18] Having education in a “general field” is also much more common among sensitive workers (56 percent compared to 27 percent). This is closely associated with lower levels of education.

[19] However, Hartman et al. (2013) have found that the experiment had smaller effects on women than men.

monitoring. The same holds for the female-dominated health industry (which also includes elderly care). On the other hand, workers in the male-dominated manufacturing sector react much more strongly to monitoring. The results are consistent with a higher conscientiousness among women, or with an aversion to obtaining medical certificates among men even when they are sick. Another interesting finding is that there is no strong connection between monitoring responsiveness and days spent taking care of sick children in the pre-period or with the person's share of total family sick child days. This suggests that family responsibilities are not a major factor in determining sickness absence behaviour.

Those who are less important for or less attached to their workplaces react more to monitoring. Sensitive workers have a lower earnings rank at their workplace, suggesting a lower position in the workplace hierarchy, and are more likely to have income from an additional job or from self-employment.

The results show that there are peer effects or sorting across neighbourhoods, workplaces and families. A range of neighbourhood variables is strongly correlated with sensitivity to monitoring. Sensitive workers tend to live in neighbourhoods with low socioeconomic status as reflected in low average earnings, small shares of highly educated inhabitants, high reliance on social payments, and high shares of immigrants. In particular, their neighbours take more days of sick leave and more short sickness absence spells. This is consistent with neighbourhood effects driving differences in local benefit cultures, as identified by Lindbeck et al. (2016), but might also be due to residential sorting. There are similar, but weaker, patterns when it comes to colleagues' days of sickness absence. Finally, sensitive individuals have partners who have taken more sick leave and more short spells in the past. This is consistent with, for example, correlations in health or attitudes to sickness absence among partners, staying at home to help partners with weak health, or shirking from work to spend time with a partner.

The spell-level variables are not particularly strong predictors of monitoring sensitivity. However, more sensitive spells on average start earlier in the experimental period and at the same time are taken out by individuals with more previous spells during the experiment. This apparent contradiction is an effect of sensitive workers taking more absence spells (1.65 on average for workers in Quartile 4, compared to 1.24 in Quartile 1). If an individual takes more absence spells, he or she is likely to both take some spells early in the experimental period, and to have had more spells before a given spell. This is consistent with workers who have high sick leave uptake being more sensitive to monitoring. Importantly for policymakers, there is no evidence of workers adjusting their behaviour over time in response to the new monitoring rules. When conditioning on individual fixed effects, the estimated sensitivity of spell duration to monitoring does not increase with the number of previous spells during the experiment.

The high predictive power of the individual's sick leave history is encouraging, as this characteristic is readily available to policymakers and its use for targeting monitoring is not counter to legal restrictions. Another characteristic that can be of interest for policymakers is the average sickness absence at a workplace.

A concern when implementing targeted monitoring policies is whether individuals are affected by their peers' treatment status, that is whether the SUTVA assumption holds. The results do not suggest that this is a major issue, as sensitive and non-sensitive workers do not differ in terms of partners' and colleagues' treatment status. Another piece of evidence against disruptive behavioural effects is the even representation of individuals from Gothenburg and Jämtland among the sensitive and non-sensitive groups, in spite of the experiment's different directions in the two regions (a loosening of the rules for the treated group in Gothenburg and a tightening of the rules for the control group in Jämtland). There is thus no evidence of e.g. control workers in Jämtland staying on sick leave for longer because they perceive themselves as unfairly discriminated or of treated workers in Jämtland taking out more sick leave because they have gotten more used to the relaxed rules over time. Overall, this suggests that adaptation to monitoring policies over time or in response to others' treatment status is limited.[20] Because of this, I do not take these aspects into account in the policy discussion in Section 7.

Classifying individuals and spells as sensitive or not based on the probability of returning to work on days 8-14 has little effect on who is identified as sensitive to monitoring, as the correlation between the two measures is 0.86. Differences between the most and least sensitive quartiles based on the probability measure, corresponding to Figure 7, are shown in Figure B5 in Appendix B. The same patterns are apparent as when using the duration-based measure of sensitivity to monitoring.

c. Drivers of the Model's Predictions

To measure which variables drive the GRF model's predictions of sensitivity, I estimate partial dependence functions for each variable. I set the variable to a given value for all workers and spells, while keeping the other variables at their empirically observed values, and predict $\hat{\tau}_x$

[20] Further evidence against strong effects driven by the direction of the experiment is provided in Figures B3 and B4 in Appendix B where I use the spells of workers from Gothenburg to estimate the sensitivity of spells of workers from Jämtland. The model that only uses spells of workers from Gothenburg yields estimates which are very highly correlated with those of the baseline model, and predicts the monitoring sensitivity of spells of workers in Jämtland well.

using the model. If there are big changes in the $\hat{\tau}_x$, the manipulated variable is an important driver of the model's estimates.

I evaluate the mean $\hat{\tau}_x$ when each continuous variable is set to its first – ninth decile values and when each binary variable is set to zero and one. For variables that are concentrated at a few mass points, I evaluate the mean $\hat{\tau}_x$ when the variable is set to all values that contain five percent or more of observations. In the case of variables where most individuals have a value of zero, I evaluate the mean $\hat{\tau}_x$ at zero and at the average value among those with nonzero values. If many observations have missing values for a variable (e.g., partner's previous sick leave among those who do not have a partner), I also evaluate the mean $\hat{\tau}_x$ if the variable is set to missing. Plots of partial dependence functions for each variable are provided in Appendix B. Demographic and health-related characteristics are shown in Figure B6, family-related characteristics in Figure B7, education in Figure B8, work-related characteristics in Figure B9, sector of work in Figure B10 and neighbourhood characteristics in Figure B11.

The results in Figures B6-B11 suggest that the model's predictions are driven by variables which are important predictors of monitoring sensitivity according to Figure 7. If all workers' previous sick leave duration is set to five days (the 10th percentile value), the average estimated increase in spell duration is 0.89 days; if it is set to 180 days (the 90th percentile value), the estimated average increase is 1.15 days. Similar patterns hold for the social payment share of income, with average increases in spell duration of 0.87 days and 1.25 days when all workers' shares are set to the 10th and 90th percentile values (two percent and 30 percent) respectively. The model predicts that workers would respond more strongly to monitoring if they had low earnings or less than high school education, but the relationship becomes flat at higher values of earnings and education. Consistent with Figure 7, workers are estimated to be less sensitive to monitoring if they would all behave as if they were women or worked in the public sector or health industry.

An interesting difference compared to Figure 7 is the pattern regarding the spell start date. Spells which start at the beginning of the period (late July) are highly sensitive to monitoring, whereas spells which start in the fall months are less sensitive. However, spells which start in December, and especially around Christmas, are again sensitive to monitoring. This pattern seems difficult to explain by medical factors and indicates higher rates of shirking at times of the year when the opportunity cost of working is higher.

It is often unrealistic to manipulate variables separately, as different characteristics can be highly correlated.[21] Figure 8 presents partial dependence functions when groups of related variables (past sick leave, socioeconomic status, gender and gender-typical industry and sector, attachment to main job, colleague and partner behaviour, neighbourhood socioeconomic status) are manipulated at the same time. It is apparent that socioeconomic status, as measured by education, earnings and social income share, is an important driver of the model's monitoring sensitivity predictions. The average increase in spell duration would be 0.75 days if all workers had high socioeconomic status and 1.47 days if all workers had low socioeconomic status. Sick leave history, as well as gender and gender-related work sector choices are also important drivers of predictions, followed by neighbourhood socioeconomic status. Variables which capture the behaviour of peers and importance at or attachment to the workplace are not as important. These findings confirm that much of the variation in sensitivity to monitoring can be captured by socioeconomic status, sick leave history and gender.[22] It is thus possible to use these groups of variables when designing targeted monitoring policies.

7. Targeted Monitoring Policy

The model's predictions of $\hat{\tau}_x$ can be used for selective monitoring of workers. Relaxing monitoring for workers with low $\hat{\tau}_x$ can reduce the resources spent on monitoring while avoiding large increases in sickness absence. The targeted monitoring policy I consider involves requiring medical certificates from more sensitive workers after seven days, as is done currently, while requiring medical certificates from less sensitive workers after 14 days.

A targeted monitoring policy implies tension between efficiency in the use of the healthcare system's limited resources on the one hand and equal treatment of all recipients of paid sick leave on the other. Limitations on how policymakers can target monitoring in practice arise because of ethical considerations and anti-discriminatory laws. Discrimination based on characteristics such as ethnicity, gender and age is legislatively forbidden in most countries. While basing a targeted policy on socioeconomic status using variables such as earnings and

[21] For instance, the correlation between average neighbourhood earnings and the share of highly educated individuals is 0.73; spells taken out by women comprise 76 percent of public sector spells and 88 percent of health industry spells.

[22] The `grf` package also has a simple built-in variable importance measure, intended as a rough diagnostic of the model. The measure is based on the number of times the causal forest's trees split on a characteristic up to depth $d = 4$. Each split is weighted by the depth at which it is made, with a split at depth d having half the weight of a split at $d - 1$. This importance measure is presented in Figure A10 in the Appendix, where the bars represent the weighted share of times each characteristic was used for splits up to $d = 4$. This metric gives results similar to those provided by evaluating partial dependence functions, but overestimates the importance of e.g. plant size and average sick leave among colleagues.

the share of social payments in income might not be explicitly illegal, it would likely be seen as unfair targeting of low-income individuals.

The most promising way of designing a targeted monitoring policy would be to base it on recipients' history of sickness absence. This would entail a more relaxed monitoring regime for those with little past sick leave and a stricter regime for those with higher past sick leave. The advantage of such a system is that it self-regulates against misuse among the less stringently monitored group. If a sick leave recipient increases his or her sick leave in response to reduced monitoring, he or she will eventually end up in the more stringently monitored group.

I estimate the gains of implementing policies based on GRF and on previous sickness absence using workers from the held-out test set. The results, compared to a random relaxation in monitoring, as took place during the experiment, are shown in Panel A of Figure 9.[23] The increase in sickness absence from reduced monitoring is shown as a share of the total increase that would take place if all workers were monitored less stringently, plotted against the share of workers for whom monitoring is relaxed. The blue line corresponding to the 45° line shows the proportional increase in sickness absence if the workers for whom monitoring is relaxed are randomly selected. Using the $\hat{\tau}_x$ from the GRF model allows for greater efficiency. If workers are ranked by their $\hat{\tau}_x$, and monitoring relaxations are targeted toward those who are less sensitive, the increase in absence is shown by the maroon line. It is far below the blue line, showing that it is possible to limit the increase in sickness absence through targeted monitoring, and providing further validation of the model on held-out data. The green line shows what can be achieved if the policymaker only uses information about previous days of sickness absence, relaxing monitoring for those with lower sickness absence.[24] About half of the gains of the full GRF model are retained if only sick leave history information is used. In particular, consider

[23] In Figure B13 in Appendix B I provide estimates corresponding to those in Panel A of Figure 9 for monitoring based on the social payment share of income. The gains of targeting monitoring using the social payment share are similar to those achieved by using past sick leave when it comes to identifying sensitive workers, but slightly smaller when it comes to identifying non-sensitive workers. Furthermore, I estimate optimal policy trees (Athey and Wager, 2021) to identify the best simple rules for selecting workers and spells in the top quartile, top half and top three quartiles of monitoring sensitivity. The rules identified by the trees involve selection based on social payment share in income and on whether the individual works in the health industry. A policy based on both the social payment share and a health industry indicator does slightly better than using sick leave history (Figure B13). However, I focus on the sick leave history policy, as using the social payment share of income is likely to be politically unfeasible. Formal rank-weighted average treatment effect (RATE, Yadlowsky et al., 2021) metrics for different targeting rules are presented in Table B3 in Appendix B. As measured by the QINI coefficient, the GRF model outperforms the simple rules, although the simple rules perform significantly better than randomisation. Using information on social payment share in income and a health industry indicator as suggested by the optimal policy trees achieves a higher RATE than using either past sickness absence or social payment share on its own.

[24] Among those with equal past sickness absence, it is random who is selected into relaxed monitoring first.

the case of relaxing monitoring for the same share of workers as in the experiment, 49 percent. In expectation, randomising who gets relaxed monitoring results in an increase in sickness absence equivalent to 49 percent of the total increase if monitoring is relaxed for all workers. Targeting based on the GRF model can significantly improve performance, and is estimated to only increase sickness absence by an amount equivalent to 24 percent of the total increase; only using sick leave history yields an increase equivalent to 37 percent of the total increase. The full-information policy thus makes it possible to reduce monitoring for the same share of workers as in the experiment at a 51 percent smaller loss in terms of extra sickness absence; the sick leave history policy still gives a 24 percent smaller loss.

It is possible to quantify the costs and benefits of more stringent monitoring under assumptions about the costs of monitoring and the output lost due to absence from work. In the following analysis, I assume that the costs of more stringent monitoring are equal to the individual's pre-tax wage times the increase in spell duration. The benefits are assumed to equal the saved cost for a visit to a primary clinic; the cost of a primary clinic visit is assumed to be a constant across worker types. This suggests that monitoring efforts should be focused on workers who have high wages and whose sickness absence reacts strongly to monitoring.

I assume that the cost of monitoring is equivalent to 1941 SEK (192 USD), the calculated cost of providing a primary care visit in Gothenburg in 2022 (Västra sjukvårdsregionen, 2022). The median daily pre-tax wage in Sweden (including payroll taxes) was 1 478 SEK (146 USD) in 2022. For someone who has a median wage, the medical certificate must thus reduce absence by at least 1.31 days in order to break even from a social point of view (ignoring any costs of worse long-term health or contagion arising from premature returns to work). However, workers with different incomes are not equally sensitive: those who increase their absence more when monitoring is relaxed tend to have lower incomes on average. To take this into account, I rescale workers' wages to 2022 levels, keeping the size of their wage relative to the mean the same as in 1988. The resulting cost-benefit analysis of stringent monitoring of workers according to different policies is shown in Panel B of Figure 9.

On average, monitoring workers more stringently, as under the current system, is estimated to be socially inefficient. The expected benefit of monitoring a random worker is 1682 SEK (166 USD), as shown by the blue line, amounting to only 87 percent of the cost. However, relaxing monitoring for non-sensitive workers, and only leaving the current regime in place for sensitive workers can be socially efficient. The benefit of stringent monitoring of the most sensitive workers according to the full GRF model is over 4000 SEK (395 USD), which is well above the cost. If workers are arranged according to their monitoring sensitivity as estimated by the GRF model, monitoring the 81 percent who are most sensitive breaks even from a social point of view. Monitoring those with a history of high sick leave also has potential for gains relative to monitoring everyone. If medical certificate requirements are retained at seven days for the 57 percent with the highest sick leave in the past, and relaxed to fourteen days for the remainder of workers, the average social benefit would equal the social cost. While simplified, this analysis suggests that the current monitoring regime may be made more efficient through a targeted relaxation of the monitoring rules.

8. Discussion

The findings point to substantial systematic heterogeneity in how recipients of paid sick leave react to reduced monitoring. For the least sensitive decile of individuals, sickness absence spells increase by only 0.47 days, compared to 1.67 days for the most sensitive decile. The key predictors of high sensitivity to monitoring are high previous sick leave, low socioeconomic status and male gender. The degree of attachment to the job, as well as colleagues' and neighbours' behaviour also have predictive power. A key finding is that many predictors of high sick leave uptake, such as female gender and working in the public sector, are not predictors of high sensitivity to monitoring. The existence of workplaces with high sick leave uptake and high monitoring sensitivity suggests that the management at such establishments should take steps to improve working conditions, especially given findings that such measures are effective in reducing absenteeism (Huber et al., 2015).

For policymakers, a selective relaxation of monitoring for some groups of workers can be a way of reducing costs while minimising the effect on sickness absence. Such a policy might be motivated in light of the strained resources of many countries' healthcare systems and the high opportunity cost of medical professionals' time. Back-of-the-envelope calculations suggest that monitoring could be reduced by the same amount as in the experiment, but causing only 49 percent of the increase in sickness absence if efforts are targeted using all the characteristics included in this study, or 77 percent of the increase if only sick leave history is used. Simple cost-benefit calculations favour such selective monitoring policies. I estimate that the current policy of monitoring all workers after seven days of sickness absence is not socially efficient, suggesting that monitoring of workers who are estimated to not be sensitive should be relaxed. For groups who are estimated to be sensitive, on the other hand, the benefits of the current monitoring regime exceed the costs.

While targeted monitoring has high potential when it comes to increasing efficiency, ethical concerns must also be taken into account when designing policy. Monitoring based on many of the worker characteristics included in the full GRF model would likely be seen as discriminatory or unfair. In particular, it would be highly controversial to use variables such as gender, immigrant background or income for monitoring purposes. A policy which varies monitoring intensity based only on sick leave history would thus be preferable for ethical and practical reasons. Another upside of such a policy is that it self-regulates against overuse by individuals who have little past sick leave uptake. If these workers increase their sickness absence by a significant amount in response to the reduction in monitoring, they will eventually end up in the more intensely monitored group.

Another concern to keep in mind is that not all reductions in sick leave are socially beneficial. If there are monetary or time costs of obtaining a medical certificate, workers might forgo days of absence which would have been medically motivated. This might lead to both negative longer-term effects on the worker's own health (see e.g. Marie and Vall Castelló, forthcoming) and to the infection of others at the workplace, an issue which has been prominent during the Covid-19 pandemic. The design of a policy which takes these broader issues into account is left for future research.

References

Angelov, Nikolay, Johansson, Per, and Lindahl, Erica. (2013). *Gender Differences in Sickness Absence and the Gender Division of Family Responsibilities*, IZA Discussion Paper No. 7379

Angelov, Nikolay, Johansson, Per, Lindahl, Erica, and Lindström, Elly-Ann. (2011). *Kvinnors och mäns sjukfrånvaro,* IFAU Report 2011:2

Athey, Susan, and Imbens, Guido. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113.27:7353-7360.

Athey, Susan, Tibshirani, Julie, and Wager, Stefan. (2019). Generalized random forests. *Ann. Statist*. 47.2:1148 – 1178

Athey, Susan and Wager, Stefan. (2021). Policy Learning With Observational Data, *Econometrica*, 89:133–161.

Baert, Stijn, Bas van der Klaauw, and Gijsbert Van Lomwel. (2018). The effectiveness of medical and vocational interventions for reducing sick leave of self-employed workers. *Health economics,* 27.2:139-152.

Barmby, Tim A., Ercolani, Marco G., and Treble, John G.. (2002). Sickness Absence: An International Comparison. *The Economic Journal*, 112:480

Bedard, Kelly, and Dhuey, Elizabeth. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics* 121.4: 1437-1472.

Böckerman, Petri, Kanninen, Otto, Suoniemi, Ilpo. (2018). A kink that makes you sick: The effect of sick pay on absence. *Journal of Applied Econometrics* 33.4: 568-579.

Boeri, Tito, Di Porto, Edoardo, Naticchioni, Paolo, and Scrutinio, Vincenzo. (2021). *Friday Morning Fever. Evidence from a Randomized Experiment on Sick Leave Monitoring in the Public Sector*. CEPR Discussion Paper No. DP16104.

Bratberg Espen, Dahl, Svenn-Åge, and Risa, Alf Erling. (2002). The double burden - Do combinations of career and family obligations increase sickness absence among women?, *European Sociological Review*, 18:233-249.

Bratberg, Espen, and Monstad, Karin. (2015). Worried sick? Worker responses to a financial shock. *Labour Economics,* 33:111-120

Breiman, Leo. (2001). Random forests. *Mach. Learn*. 45:5–32.

Breiman, Leo, Friedman, Jerome H., Olshen, Richard. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA.

Chernozhukov, Victor, Demirer, Mert, Duflo, Esther and Fernàndez-Val, Ivàn. (2020). *Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India*. National Bureau of Economic Research, WP 24678

Dale-Olsen, Harald. (2013). Absenteeism, efficiency wages, and marginal taxes. *The Scandinavian Journal of Economics* 115.4:1158-1185.

Ferman, Bruno, Torsvik, Gaute, and Vaage, Kjell. (2021). Skipping the doctor: evidence from a case with extended self-certification of paid sick leave. *J Popul Econ* 1-37

Försäkringskassan. (2021). *Effekter som covid-19 har på sjukförsäkringen*, FK 2020/000065

Frick, Bernd, and Malo, Miguel Á. (2008). Labor market institutions and individual absenteeism in the European Union: the relative importance of sickness benefit systems and employment protection legislation. *Industrial Relations: A Journal of Economy and Society*, 47.4:505-529.

Hartman, Laura, Hesselius, Patrik, and Johansson, Per. (2013). Effects of eligibility screening in the sickness insurance: Evidence from a field experiment, *Labour Economics*, 20:48-56

Haugen, Katarina, Holm, Einar, Lundevaller, Erling, and Westin, Kerstin. (2008). Localised attitudes matter: a study of sickness absence in Sweden. *Population, Space and Place*, 14.3:189-207.

Helgesson, Magnus, Johansson, Bo, Nordqvist, Tobias, Lundberg, Ingvar, and Vingård, Eva. (2015). Sickness absence at a young age and later sickness absence, disability pension, death, unemployment and income in native Swedes and immigrants, *European Journal of Public Health*, 25. 4:688–692

Henrekson, Magnus, and Persson, Mats. (2004). The effects on sick leave of changes in the sickness insurance system. *Journal of Labor economics* 22.1:87-113.

Hensvik, Lena and Rosenqvist, Olof. (2019). Keeping the Production Line Running: Internal Substitution and Employee Absence, *J. Human Resources* 54:200-224

Hesselius, Patrik, Nilsson, Peter J., and Johansson, Per. (2009) Sick of Your Colleagues‘ Absence?, *Journal of the European Economic Association*, 7.2-3

Hesselius, Patrik, Johansson, Per, and Vikström, Johan. (2013). Social behaviour in work absence. *The Scandinavian Journal of Economics*, 115.4:995-1019.

Huber, Martin, Lechner, Michael, and Wunsch, Conny. (2015). Workplace health promotion and labour market performance of employees. *Journal of Health Economics*, 43:170-189.

Johansson, Per, Karimi, Arizo, and Nilsson, Peter J. (2019). Worker absenteeism: peer influences, monitoring and job flexibility. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182.2:605-621.

Johansson, Per, and Palme, Mårten. (2002). Assessing the effect of public policy on worker absenteeism. *Journal of Human Resources*, 37.2:381-409.

Johansson, Per and Palme, Mårten. (2005). Moral hazard and sickness insurance, *Journal of Public Economics*, 89.9–10:1879-1890

Knaus, Michael C., Lechner, Michael, and Strittmatter, Anthony. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal* 24.1:134-161.

Lechmann, Daniel SJ, and Schnabel, Claus. (2014). Absence from work of the self-employed: a comparison with paid employees. *Kyklos* 67.3:368-390.

Lindbeck, Assar, Palme, Mårten, and Persson, Mats. (2016). Sickness absence and local benefit cultures, *The Scandinavian Journal of Economics*, 118.1: 49-78

Lindgren, Karl-Oskar, *Workplace size and sickness absence transitions*, IFAU Working paper 2012:26

Lusinyan, Lusine, and Bonato, Leo. (2007). Work absence in Europe. *IMF Staff Papers* 54.3:475-538.

Marie, Olivier and Vall Castelló, Judit. (2022). Sick Leave Cuts and (Unhealthy) Returns to Work. *Journal of Labor Economics*, forthcoming

Markussen, Simen, Røed, Knut, Røgeberg, Ole J., Gaure, Simen. (2011). The anatomy of absenteeism, *Journal of Health Economics*, 30.2: 277-292

Nie, Xinkun, and Wager, Stefan. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108.2:299-319

van Ommeren, Jos N. and Gutiérrez-i-Puigarnau, Eva. (2011). Are workers with a long commute less productive? An empirical analysis of absenteeism, *Regional Science and Urban Economics*, 41.1

OECD. (2020). *Paid sick leave to protect income, health and jobs through the COVID-19 crisis*.

OECD. (2021). *Public spending on incapacity (indicator)*.

Paringer, Lynn. (1983). Women and absenteeism: health or economics?. *The American economic review* 73.2:123-127.

Palme, Mårten, and Persson, Mats. (2020). Sick Pay Insurance and Sickness Absence: Some European Cross-Country Observations and a Review of Previous Research. *Journal of Economic Surveys* 34.1:85-108.

Piha, Kustaa, Laaksonen, Mikko, Martikainen, Pekka, Rahkonen, Ossi, and Lahelma, Eero. (2010). Interrelationships between education, occupational class, income and sickness absence, *European Journal of Public Health*, 20.3:276–280

Riksförsäkringsverket. (1989). *Utvidgad egen sjukskrivning*, AD 1989-954:01

Robins, James M., and Rotnitzky, Andrea. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90.429:122–129.

Skogman Thoursie, Peter. (2004). Reporting sick: are sporting events contagious? *Journal of Applied Econometrics*, 19.6:809-823.

Spierdijk, Laura, van Lomwel, Gijsbert, and Peppelman, Wilko. (2009). The determinants of sick leave durations of Dutch self-employed. *Journal of health economics*, 28.6:1185-1196.

Treble, John and Barmby, Tim. (2011). *Worker absenteeism and sick pay*, Cambridge University Press, Cambridge/New York

Västra sjukvårdsregionen. (2022). *Utomlänspriser 2022*, HS 2021-01109

Watson, Alfred W. (1927). National Health Insurance: A Statistical Review. *Journal of the Royal Statistical Society* 90.3:433-486.

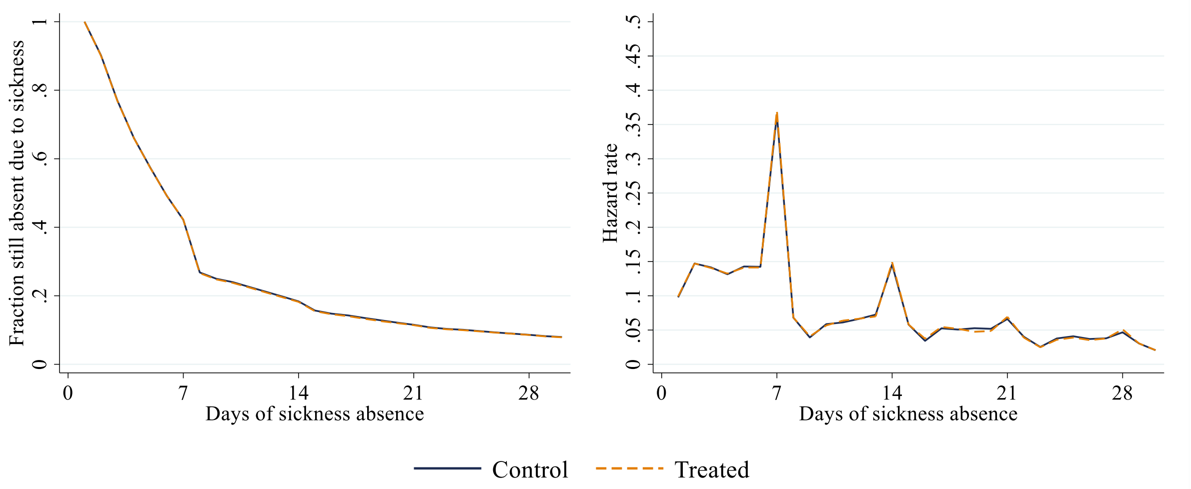
Winkelmann, Rainer. (1999). Wages, firm size and absenteeism, *Applied Economics Letters*, 6:6, 337-341

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*.

Ziebarth, Nicolas R., and Karlsson, Martin. (2010). A natural experiment on sick pay cuts, sickness absence, and labor costs. *Journal of Public Economics*, 94.11-12:1108-1122.

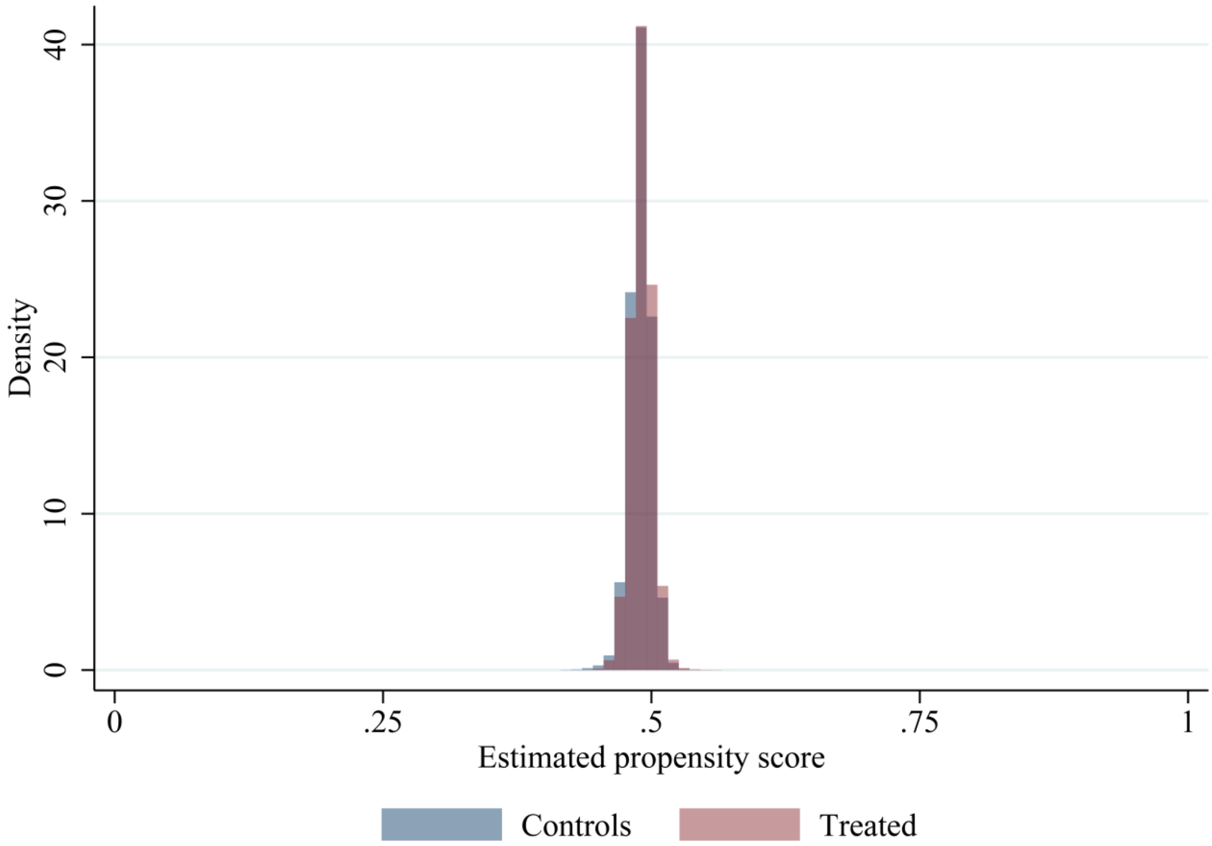
Figures

FIGURE 1. SURVIVAL AND HAZARD RATES FOR SICKNESS ABSENCE SPELLS OF TREATED AND CONTROL WORKERS IN GOTHENBURG AND JÄMTLAND BEFORE THE EXPERIMENT.



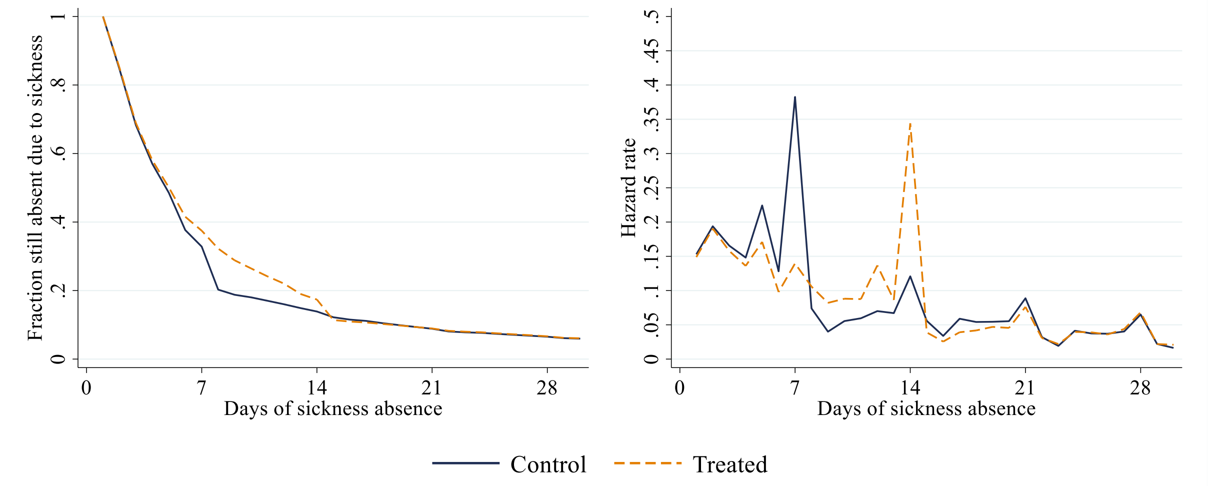
Note: The hazard rate represents the probability that a worker who has been absent for a given number of days returns to work on the next day. Spells which began between July 1st and December 31st 1987 (pre-period). Control workers born on odd dates, treated workers on even dates.

FIGURE 2. PROPENSITY SCORE ESTIMATES FOR TREATED AND CONTROLS FROM GRF.



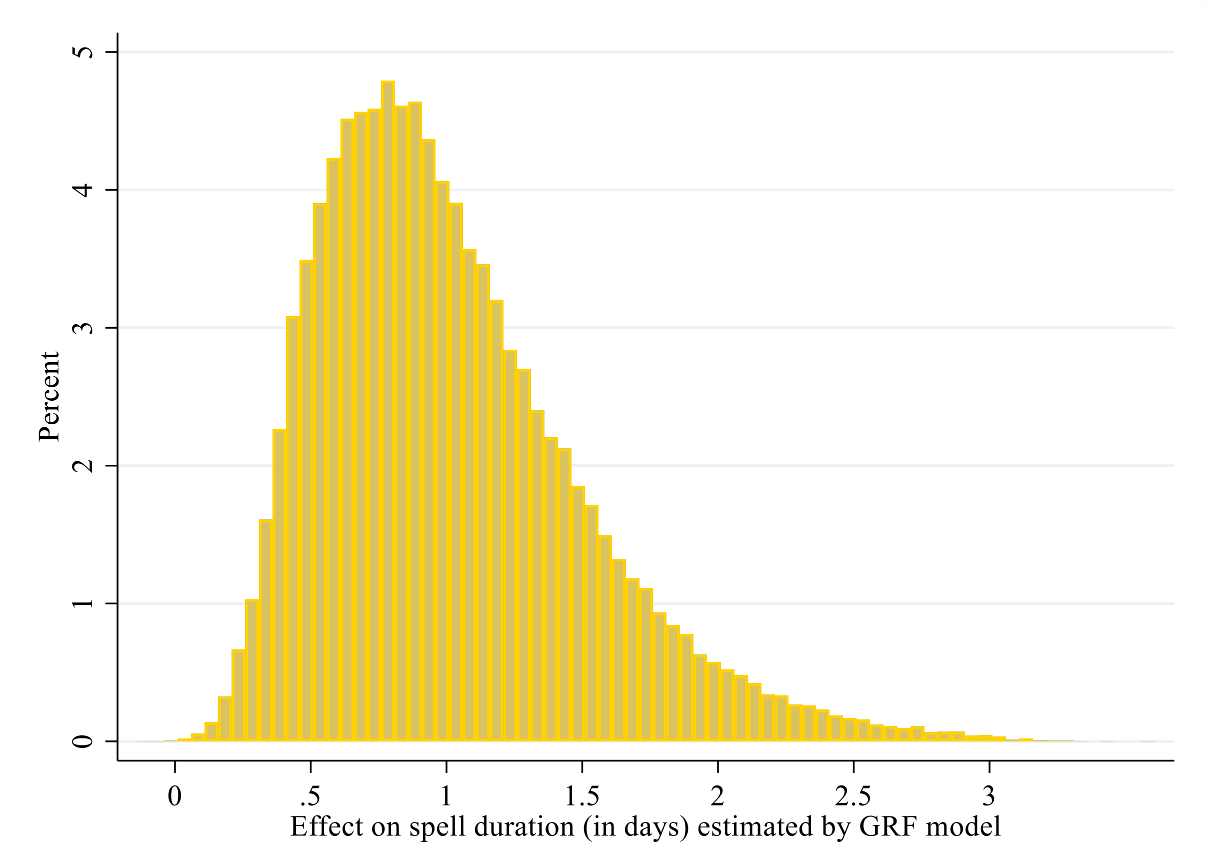
Note: GRF estimates of probability of entering treatment based on the included worker characteristics, plotted separately for treated and control workers (including those without any absence spells during the experiment). Controls born on odd dates, treated on even dates. Bin width = 0.01.

FIGURE 3. SURVIVAL AND HAZARD RATES FOR SICKNESS ABSENCE SPELLS OF TREATED AND CONTROL WORKERS IN GOTHENBURG AND JÄMTLAND DURING THE EXPERIMENT.



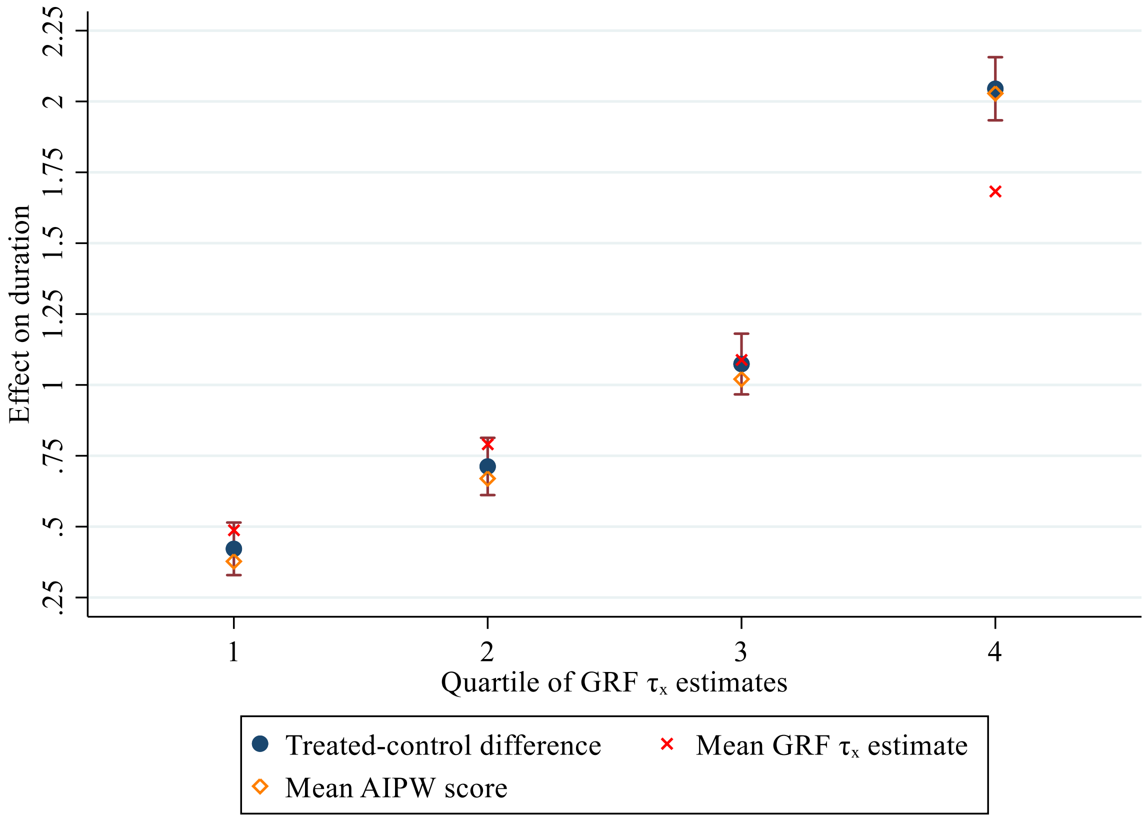
Note: Spells which began between July 1st and December 31st 1988. The hazard rate represents the probability that a spell which has been ongoing for a given number of days ends on the next day. Control workers born on odd dates, treated workers on even dates.

FIGURE 4. DISTRIBUTION OF PREDICTED TREATMENT EFFECTS ON SICKNESS ABSENCE SPELL DURATION.



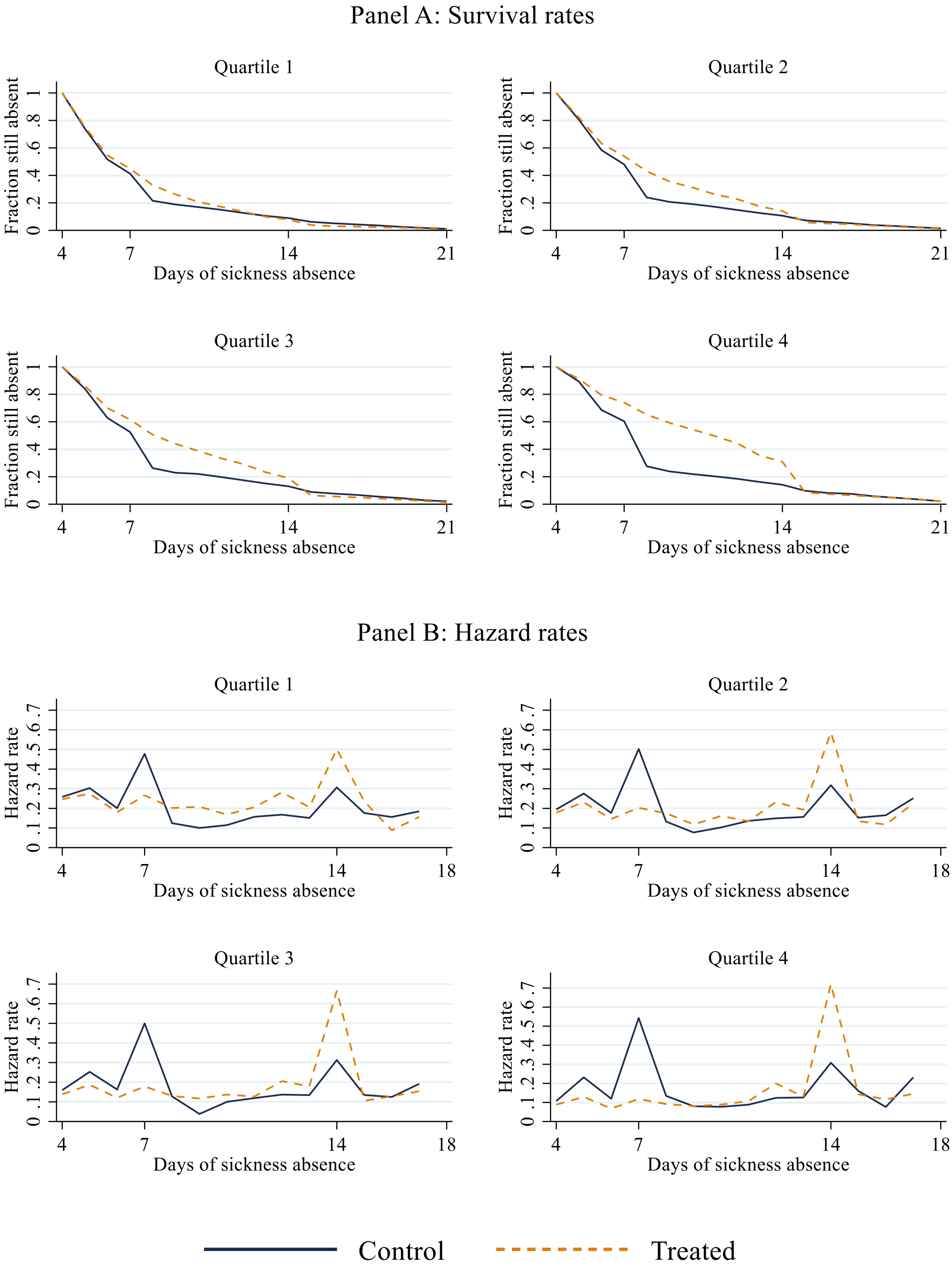
Note: Out-of-bag treatment effect estimates for training set workers. Bin width = 0.05 days.

FIGURE 5. TREATMENT EFFECTS FOR QUARTILES OF SPELLS RANKED BY THEIR ESTIMATED $\hat{\tau}_x$.



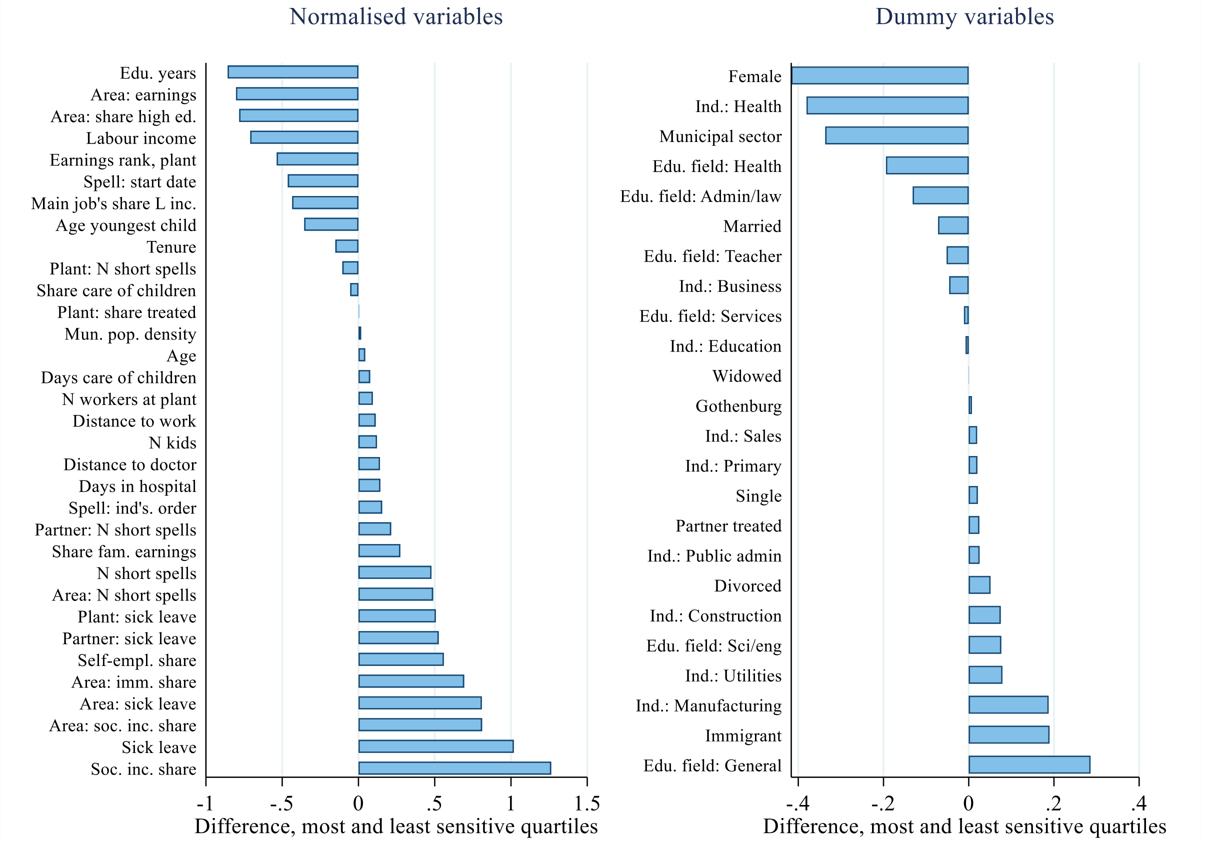
Note: Quartiles defined by ranking spells based on treatment effects estimated by GRF. Q1 contains spells estimated to be least affected and Q4 spells estimated to be most affected. Treated-control differences in duration within each quartile estimated as $\hat{\tau} = \bar{y}_i|(W_w = 1) - \bar{y}_i|(W_w = 0)$. Effects for training set workers; out-of-bag estimates of $\hat{\tau}_x$ and AIPW scores. Confidence intervals at the 95 percent level for the treated-control differences.

FIGURE 6. SURVIVAL AND HAZARD RATES OF ABSENCE SPELLS OF WORKERS IN THE HELD-OUT TEST SET, RANKED BY QUARTILES OF PREDICTED TREATMENT EFFECTS



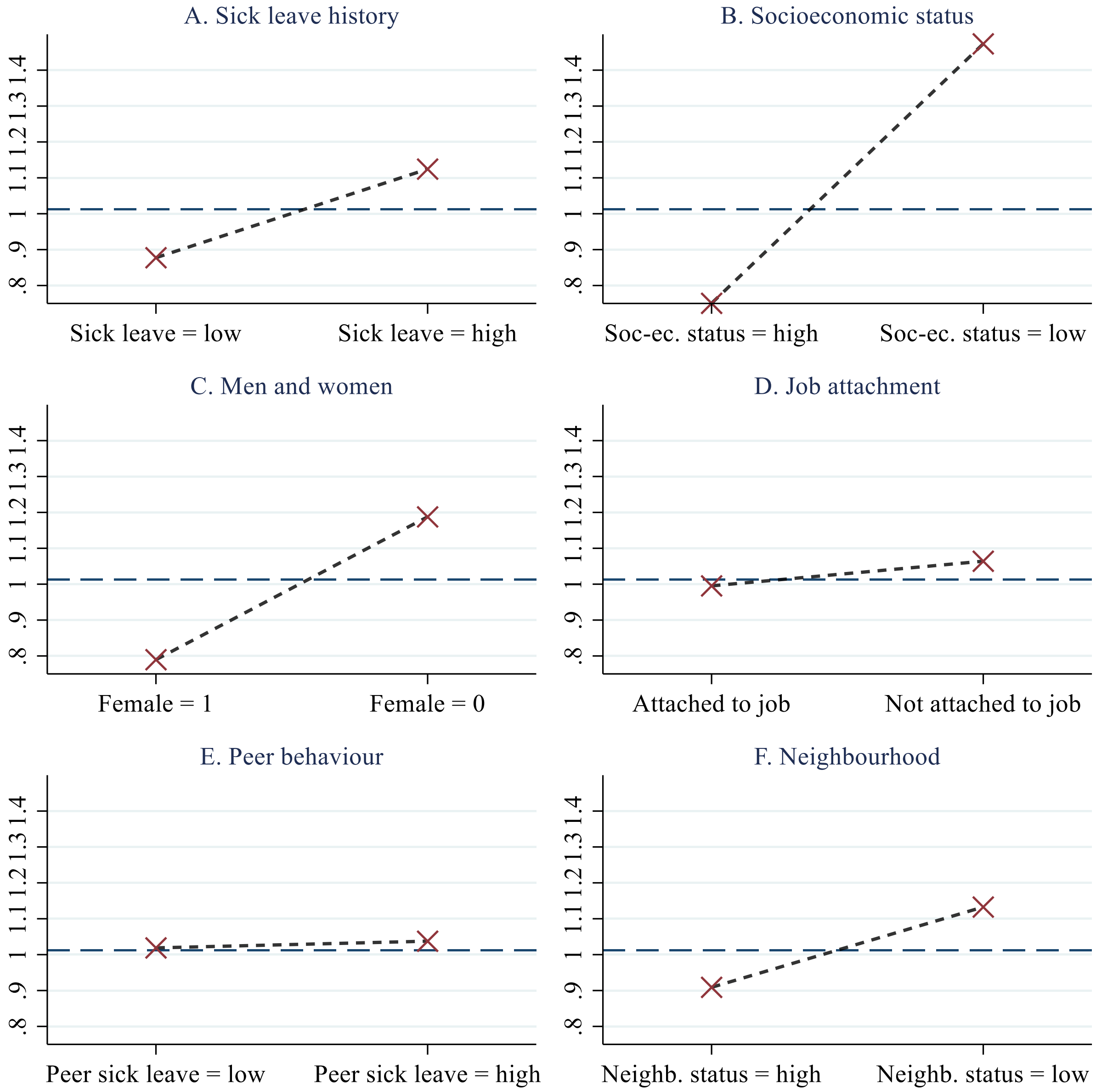
Note: Rates for absence spells of the 20 percent of workers randomised into the held-out test set. The hazard rate represents the probability that a spell that has been ongoing for a given number of days ends on the next day. Workers divided into quartiles based on predicted $\hat{\tau}_x$. Q1 contains spells estimated to be least affected and Q4 spells estimated to be most affected. Controls born on odd dates, treated on even dates.

FIGURE 7. DIFFERENCES IN TERMS OF INDIVIDUAL AND SPELL CHARACTERISTICS BETWEEN THE QUARTILES MOST AND LEAST RESPONSIVE TO MONITORING.



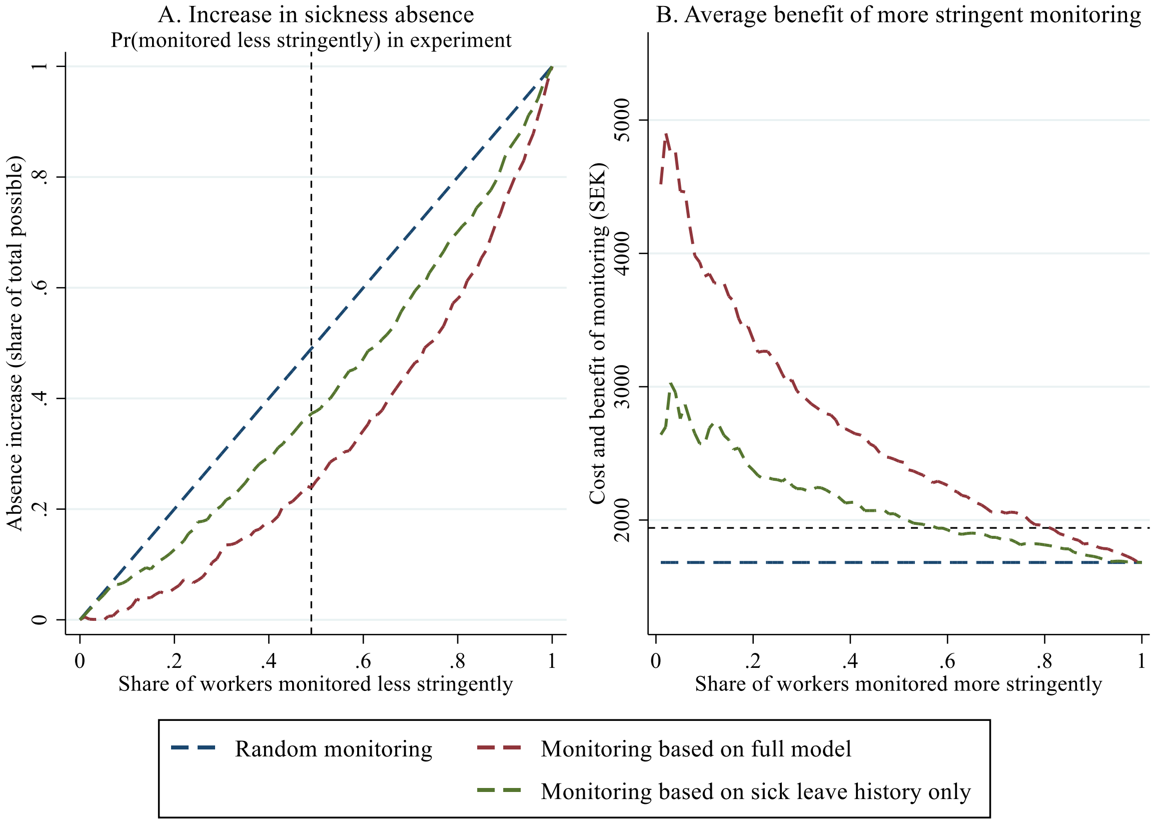
Note: Quartiles defined based on GRF $\hat{\tau}_x$ so as to include equal numbers of spells. Continuous variables normalised to have mean zero and standard deviation one. For dummies, differences are based on shares with the given characteristic. Positive differences mean higher values among the most sensitive (Quartile 4), negative differences mean higher values among the least sensitive (Quartile 1). Spells of training set workers.

FIGURE 8. PARTIAL DEPENDENCE OF GRF ESTIMATES ON GROUPS OF CHARACTERISTICS.



Note: Spells of training set workers. Estimates are means of GRF $\hat{\tau}_x$ evaluated when groups of variables are set to given values for all observations, while remaining variables are held at their empirically observed values. Manipulated variables are set to the 10th and 90th percentile values among all observations to avoid using extreme values. As fewer than 10 percent of observations have any self-employment income, its share is set to zero for weak attachment to the main job and to the empirically observed value for strong attachment to the main job. The variables whose values are manipulated are: *Panel A*: days of previous sick leave (low – 5 days, high – 180 days); number of previous short spells (low – 1, high – 15); *Panel B*: annual earnings (high status – 152 kSEK, low status – 46 kSEK); education years (high status – 14 years, low status – 6 years); social income share (high status – 0.02, low status – 0.30); *Panel C*: female dummy (women – 1, men – 0); public sector (women – 1, men – 0); health industry (women – 1, men – 0); manufacturing industry (women – 0, men – 1); *Panel D*: share main job in income (attached – 1, non-attached – 0.67); share of self-employment in income (attached – 0, non-attached – empirically observed value); workplace income rank (attached – 0.88, non-attached – 0.13); *Panel E*: colleagues' previous sick leave (low – 23, high – 75); colleagues' previous short spells (low – 2.8, high – 6.5); partner's previous sick leave (low – 0, high – 167); partner's previous short spells (low – 0, high – 11); *Panel F*: neighbourhood earnings (low status – 89 kSek, high status – 122 kSEK); neighbourhood highly educated share (high status – 0.39, low status – 0.06); neighbourhood immigrant share (high status – 0.03, low status – 0.37); neighbourhood sick leave (high status – 34, low status – 76); neighbourhood short spells (high status – 2.9, low status – 5.1).

FIGURE 9. COMPARISON OF DIFFERENT MONITORING POLICIES.



Note: Effects of monitoring spells of test set workers less stringently (after 14 days of absence) or more stringently (after seven days of absence) when monitoring is targeted according to different rules. Monitoring based on full model assumes workers are ranked based on their estimated $\hat{\tau}_x$ and monitoring is targeted toward those with higher $\hat{\tau}_x$. Monitoring based on sick leave history assumes workers are ranked based on past sickness absence and monitoring is targeted toward those with higher past sickness absence. Order of monitoring among those with equal past sickness absence is random. Pr(monitored less stringently) in experiment = 0.49. Cost of monitoring = 1941 SEK (192 USD).

Appendix A: Methodological details

The goal is to estimate treatment effects τ_x for groups of spells i defined by a vector of characteristics $\boldsymbol{x}_i$. As treatment W is randomised at the level of workers w, the heterogeneous treatment effects that are the object of interest can be written as:

$$\tau_x = E(y_i|\boldsymbol{x}_i, W_w = 1) - E(y_i|\boldsymbol{x}_i, W_w = 0)$$

Given the large number of characteristics and values which can be used for splitting the sample, there are extremely many possible partitions of the sample for which it is possible to estimate τ_x. It is impossible to study all the possible ways of splitting the sample using traditional methods. For this reason, the variables and threshold levels used for making splits in heterogeneity analysis have traditionally been selected based on theory. However, there often are many different theoretical predictions, and evaluating all of them is prohibitively time-consuming. Furthermore, theoretical predictions are almost always to some degree inexact, and choosing variables and threshold values for sample splitting always involves some degree of arbitrariness.

GRF, on the contrary, identifies the characteristics and threshold values which yield the maximum heterogeneity in τ_x in an entirely data-driven way. GRF involves first estimating a model for selection into treatment (the propensity model)[25] and a model for the value of the outcome (the outcome model) based on the attributes $\boldsymbol{x}_i$, but without using information on treatment status W_w. These models are estimated using regression forests (Breiman, 2001). The resulting propensity scores $\widehat{e_x}$ and predicted outcomes $\widehat{y_x}$ are used to calculate residualised treatment status $\widetilde{W_w}$ and outcome $\tilde{y}_i$. The residualised values are used for estimating heterogeneity in τ_x using a causal forest (Athey et al., 2019). The different components of GRF, as well as how I have implemented it in practice in this paper are explained below.

a. Training and Test Sets

I split the observations into a training set and a held-out test set prior to estimating treatment effects using GRF. The training set contains 80 percent of the families in the sample and their associated sickness absence spells, while the test set contains the remaining 20 percent of families. The training set is used for constructing the GRF, while the test set is used to validate that the model predicts sensitivity to monitoring well out-of-sample. This shows whether there

[25] Strictly speaking, if experimental randomization holds, estimating conditional propensity scores $\hat{e}_x = W_w|x_w$ is unnecessary, as then $\hat{e}_x = e = 0.49 \; \forall \; \boldsymbol{x}$. However, to be more conservative, and to avoid some minor balancing issues as discussed in Section 5, I estimate $\hat{e}_x$ using GRF. This estimation provides little gain compared to the naïve model of perfect randomisation.

are, for instance, problems with overfitting (that is, the model replicating patterns in the data which are not due to stable relationships between covariates and the outcome, but rather due to random noise). If the model is able to predict sensitivity well out-of-sample, it is likely that it has been able to identify persistent relationships between the characteristics $\boldsymbol{x}_i$ and the outcome. The division into sets is based on families rather than spells to ensure that the training data contain no information on the individual's or their partner's behaviour.

b. Regression Trees and Causal Trees

Regression forest and causal forest estimation relies on constructing a large number of recursive "trees". Regression trees (Breiman et al., 1984) divide observations into groups with similar values of an outcome y, while causal trees (Athey and Imbens, 2016) divide observations into groups with similar treatment effects τ. In both cases, the divisions are based on a vector of characteristics $\boldsymbol{x}$. A tree is grown as follows:

1. The full set of absence spells is randomly divided into two groups,[26] which constitute the splitting and estimation subsamples. The trees are grown using only the workers in the splitting subsample; the estimation subsample is used to populate the leaves of the tree after the splits have been made, and for calculating estimates. This is required for a property known as honesty, which ensures consistency and asymptotic normality of forest estimates (Athey et al., 2019).

2. The full set of sickness absence spells in the splitting sample is considered.

    a. It is split into child nodes ("left" and "right") in turn at every possible threshold value of each included characteristic. The number of possible threshold values can be large for variables such as annual earnings, or just one for a binary variable such as gender.

    b. The criterion of interest is evaluated for each possible partition into child nodes. In the case of regression trees, it is heterogeneity with regard to an outcome y, while in the case of causal trees, it is heterogeneity with regard to estimated treatment effects τ. The split which maximises the criterion is selected.

    c. Steps a-b are repeated, but each of the child nodes is now in turn considered as the parent node. Each child node is split according to the variable and threshold value which maximise the criterion of interest. Splitting continues until the

[26] Half and half in the default settings of the `grf` package.

observations have been grouped into “leaves” with similar outcomes or treatment effects.

3. The splitting sample is not used for predicting the outcome or the size of the treatment effect in the “leaves”. The observations in the estimation subsample are “pushed down” into the tree, and sorted into “leaves” based on their characteristics. These estimation subsample observations are used for making predictions using the tree. When making predictions for an observation, that observation is “pushed down” analogously to the estimation subsample observations, ending up in a particular leaf. The outcomes or treated-control differences among observations in the leaf are then used to predict the outcome or treatment effect for that observation.

It is this splitting procedure which results in the GRF being nonparametric. It is also able to correctly handle ordinal variables, as the distance between two values of a variable has no effect on estimates. What matters is only whether a split was made between the two values or not; if a split was made, observations with different values of the variable will be placed in different leaves, otherwise not. The tree’s estimates are based on within-leaf neighbours, without adjusting for covariate distance between them. Furthermore, trees are able to smoothly handle missing values. When evaluating what split to make, those for whom the variable is missing are first grouped together with those with high values, then with those with low values and finally as a group by themselves.

c. Regression Forests and Causal Forests

While a single tree finds the best fit for treatment effect heterogeneity among the sample considered, the estimates of single trees can be non-robust, and their standard errors are difficult to estimate. For this reason, regression and causal forests, which are large ensembles of regression and causal trees respectively, are constructed. Each tree is constructed based on a random subsample of absence spells. This enables the forest to reveal relationships that hold consistently across random subsamples of the data. To further increase randomness, not all spell characteristics $\boldsymbol{x}$ are evaluated when choosing which splits to make; only a random sample is used, and this sample is redrawn for every split in a tree.

The forest combines the output of each tree to be predict an outcome or treatment effect for different combinations of $\boldsymbol{x}$. In effect, a forest provides a highly flexible kernel for matching observations for which predictions are to be made to observations used in its training. The weight of each training set observation in this kernel is determined by how many times it appears in the same leaf of a tree as the observation for which predictions are to be made.

In a regression or causal forest of B trees, with the number of observations in the appropriate leaf l_b of each tree given by N_b, a training observation's weight is given by:

$$\alpha_j = \frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{1}\{j \in l_b\}}{N_b} , \qquad j \in i, w$$

The estimated values of $\hat{e}_x$, $\hat{y}_x$ and $\hat{\tau}_x$ for observations with a given combination of covariates $\boldsymbol{x}$ are calculated as:

$$\hat{e}_x = \frac{1}{N}\sum_{w=1}^{N}\alpha_w W_w \ , \qquad \hat{y}_x = \frac{1}{N}\sum_{i=1}^{N}\alpha_i y_i \quad \text{and} \quad \hat{\tau}_x = \frac{\sum_{i=1}^{N}\alpha_i \tilde{y}_i \widetilde{W}_w}{\sum_{i=1}^{N}\alpha_i \widetilde{W}_w^2}$$

The values $\tilde{y}_i = y_i - \widehat{y_x}$ and $\widetilde{W}_w = W_w - \widehat{e_x}$ are residual values of the outcome and treatment propensity after the regression forest estimates have been subtracted. This is known as Robinson's transformation and makes GRF an efficient R-learner (Nie and Wager, 2019).

Overlap in terms of characteristics $\boldsymbol{x}$ between the treated and control groups is required for valid estimation. Thus, for each combination of characteristics among workers in the treated or control groups, there must be a corresponding subpopulation in the other group. Without overlap, the estimates would effectively involve extrapolation based on subpopulations with similar characteristics. This condition is shown to be satisfied in Figure 2 in the paper, as the distributions of $\widehat{e_x}$ are very similar for treated and control workers.

When making predictions for observations in the held-out test set, all trees in the forest are used. However, for training set observations, out-of-bag estimation is employed. This entails only using those trees into which the observation was not sampled. Out-of-bag estimation mitigates the overfitting inherent when using information about a particular observation when predicting outcomes or treatment effects for that observation. Because of this, on average half of the trees in the forests are used for making predictions for training set observations.

d. Practical Implementation of GRF

The data contain natural clusters in the form of sickness spells experienced by the same worker and his or her partner. To account for this structure, family-level clusters rather than single spells are drawn when selecting the sample used for constructing each tree and when dividing into splitting and estimation subsamples according to the honesty procedure. Furthermore, family clusters are reweighted when estimating $\hat{\tau}_x$ so that families who have had different numbers of sickness spells during the experiment get equal weight. Standard errors of the predicted $\hat{\tau}_x$ are cluster-robust. For computational feasibility reasons, all forests are ensembles of 5000 trees.

The models in this paper are constructed using the `grf` package in R. There are a number of parameters involved in constructing a GRF which can be varied. In the main model, all of these parameters are tuned using 50 small GRF models containing 200 trees each. The parameters selected as optimal by this tuning procedure are: fraction of data sampled into each tree = 0.47 (default = 0.50), number of variables randomly available for each split = 28 (default = 27 with 56 variables in $\boldsymbol{x}$), minimum leaf size = 1 treated and 1 control observation (default = 5 treated and 5 controls), share of splitting subsample = 0.60 (default = 0.5), maximum split imbalance = 0.04 (default = 0.05), soft imbalance penalty = 0.91 (default = 0). The effects of monitoring predicted by the default and tuned models are however very highly correlated, as can be seen in Table B2 in Appendix B.

e. Comparison of GRF and OLS and LASSO estimates

Tables B2 and B3 and Figure B14 in Appendix B compare the predictions of GRF to those of LASSO and OLS. Comparisons to parametric models are problematic, as they are not able to handle missing values. As variables related to colleagues and partners are missing for those at single-worker establishments and for singles respectively, some variables or observations have to be excluded when implementing parametric estimation. I exclude variables related to partners (partner's previous days of sick leave, number of short spells and treatment status), as 55 percent of spells in the sample involve single workers.[27] The age of the youngest child is missing for individuals without children under 18 years of age in the household, and is also excluded. Those at single-person establishments (two percent of observations) are dropped, as it is impossible to assess variables relating to colleagues' behaviour for them. For workers without a fixed establishment (nine percent of observations), workplace size is set to the number of workers without a fixed establishment within the firm, and distance to work is set to the sample mean.

I include quadratic terms and first-order interactions between variables, as well as a quartic in age, when estimating the OLS and LASSO models. The optimal penalty parameter for LASSO is selected through grid search and cross-validation across five folds.

[27] The high share is partly explained by cohabiting couples without common children not being identifiable in the data.

As shown in Table B2 in Appendix B, GRF estimates are quite strongly correlated with LASSO estimates (ρ=0.73), but quite weakly correlated with OLS estimates (ρ=0.39). This is expected, as both GRF and LASSO are designed to minimise overfitting, while OLS estimates are not regularised. I then estimate how well the different models perform when assessing the sensitivity of absence spells of test set workers. Estimates of rank-weighted average treatment effects (RATE, Yadlowsky et al., 2021) are shown in Table B3 in Appendix B. These suggest that GRF outperforms both LASSO and OLS, although the 95 percent confidence intervals of the RATE estimates overlap. Finally, the gains of targeted monitoring based on the GRF, LASSO and OLS models, as compared to random monitoring, are visualised in Figure B14 in Appendix B. Targeting based on the LASSO model does somewhat worse than GRF for the vast majority of shares of targeted workers. Targeting based on OLS does clearly worse than either using GRF or LASSO, and is similar to targeting based on the simple rules in Figure B13.

Appendix B: Additional Results

TABLE B1. BALANCING TABLE FOR CHARACTERISTICS OF TREATED AND CONTROL WORKERS

	Mean, controls (N=108,321)	**Mean, treated (N=103,977)**	**Difference**
Days of sick leave in past 2.5 years	47.0	46.5	-0.542
Number of short spells in past 2.5 years	4.35	4.32	-0.036*
Days in hospital in past 2.5 years	0.553	0.575	0.022
Age	38.7	38.6	-0.036
Female	0.476	0.475	-0.001
Native	0.873	0.877	0.004***
Immigrant:			
Nordic	0.051	0.051	0.000
Other Europe	0.050	0.049	-0.001
Rest of World	0.026	0.023	-0.003***
Married	0.446	0.444	-0.003
Never married	0.434	0.436	0.003
Divorced	0.105	0.105	0.000
Widowed	0.014	0.015	0.000
Share of family earnings	0.765	0.766	0.001
N children	0.717	0.715	-0.002
Age of youngest child	7.95	7.95	-0.008
Sick child days	1.21	1.17	-0.034
Share of family's sick child days	0.096	0.096	0.000
Partner's sickness absence, past 2.5 years	47.6	47.4	-0.116
Partner's short spells, past 2.5 years	3.89	3.88	-0.013
Partner treated	0.492	0.486	-0.006*
Education level	10.7	10.7	-0.001
Education field:			
General	0.419	0.418	-0.001
Teacher	0.028	0.028	0.000
Administration, law, social science	0.160	0.165	0.005***
Science and engineering	0.225	0.224	-0.001
Health	0.113	0.111	-0.002
Services	0.045	0.044	-0.001
Annual labour income	113,139	113,207	68
Social payment share	0.082	0.083	0.000
Self-employment share	0.049	0.048	-0.001
Main job share	0.927	0.927	0.000
Distance to work	15.6	15.5	-0.04
Number of workers at plant	914	921	6.782
Tenure	2.63	2.64	0.002
Earnings rank at plant	0.552	0.552	0.001
Mean sick leave of colleagues, past 2.5 years	45.2	45.1	-0.056
Mean N short spells of colleagues, past 2.5 years	4.44	4.44	-0.004
Share of colleagues treated	0.490	0.489	-0.001
Local public sector	0.297	0.295	-0.002
Industry:			
Primary	0.021	0.021	0.000
Manufacturing	0.208	0.207	-0.001
Construction	0.062	0.064	0.002*
Utilities	0.086	0.086	0.000

	Mean, controls (N=108,321)	Mean, treated (N=103,977)	Difference
Sales	0.209	0.208	-0.001
Business services	0.099	0.101	0.002
Health	0.228	0.228	0.000
Education	0.033	0.031	-0.001*
Public administration	0.036	0.036	0.000
Population density in municipality	748	749	0.986
Gothenburg	0.773	0.774	0.001
Neighbourhood:			
Mean days of sick leave in previous 2.5 years	48.4	48.3	-0.122*
Mean N short spells in previous 2.5 years	3.78	3.77	-0.003
Share with post-secondary education	0.214	0.215	0.001
Mean annual earnings	106,995	107,144	149**
Mean social payment share	0.087	0.087	-0.000*
Immigrant share	0.131	0.130	-0.001
Distance to medical establishment	1.59	1.58	-0.019
Outcomes during experiment:			
Total absence	17.4	18.1	0.684**
N absence spells	1.24	1.22	-0.018***
Mean spell duration	17.7	18.6	0.884**

Note: Statistics for treated and control workers who fulfil the restrictions on being included in the main analysis, that is are aged 18-64, have annual earnings at least three times a “minimum” monthly wage (defined as the tenth percentile among blue-collar workers) and do not work for the central government. To make sure that workers were exposed to the experiment for its entire duration, only those who lived in Gothenburg or Jämtland in 1987, 1988 and 1989 are included. Workers who did not have any sickness absence spells during the experimental period are retained. Variables pertaining to partners or colleagues missing for singles and those at single-employee workplaces respectively. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

TABLE B2. CORRELATIONS BETWEEN RESPONSIVENESS TO MONITORING ACCORDING TO DIFFERENT MODELS

Model	ρ between model and baseline
GRF with default hyperparameters	0.97
GRF on spells of all durations (duration censored at 30 days), tuned hyperparameters	0.78
GRF on spells of all durations, Pr(spell ends on days 8-14) as outcome, tuned hyperparameters	0.86
LASSO (sample with non-missing colleague characteristics only)	0.73
OLS (sample with non-missing colleague characteristics only)	0.39

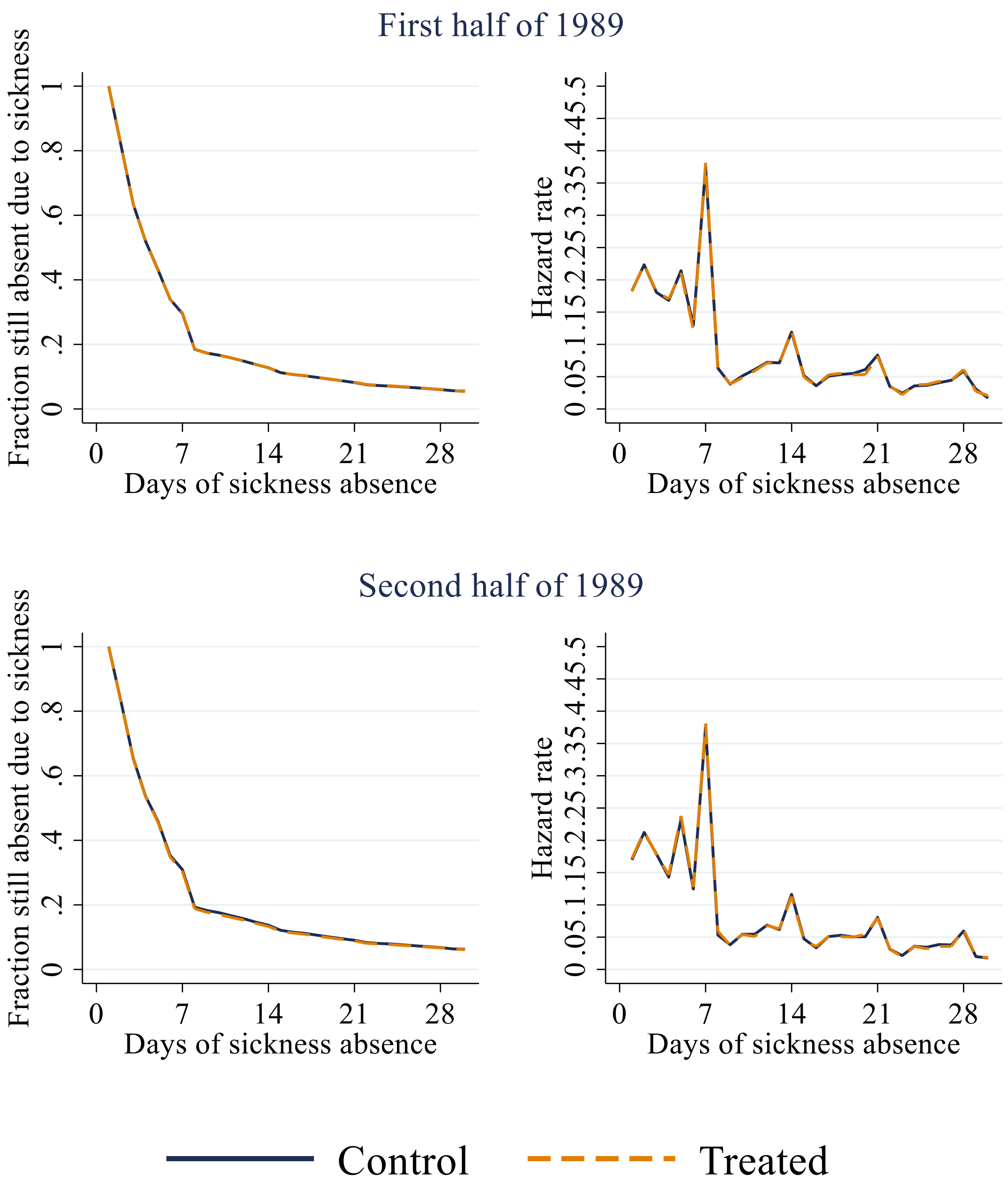
Note: Correlations with estimates from the baseline model (GRF on duration of spells 4-21 days in length, tuned hyperparameters) for spells of training set workers with durations of 4-21 days.

TABLE B3. RATE ESTIMATES OF GRF MODEL AND DIFFERENT SIMPLE TARGETING RULES ON THE TEST SET

Model	QINI coefficient (x10)	SE (x10)
GRF (baseline model)	2.04	0.16
Past sick leave	0.92	0.15
Social payment share of income	0.90	0.15
Health industry and social payment share of income (optimal simple rule)	1.30	0.15
LASSO (sample with non-missing colleague characteristics only)	1.64	0.14
OLS (sample with non-missing colleague characteristics only)	1.58	0.15

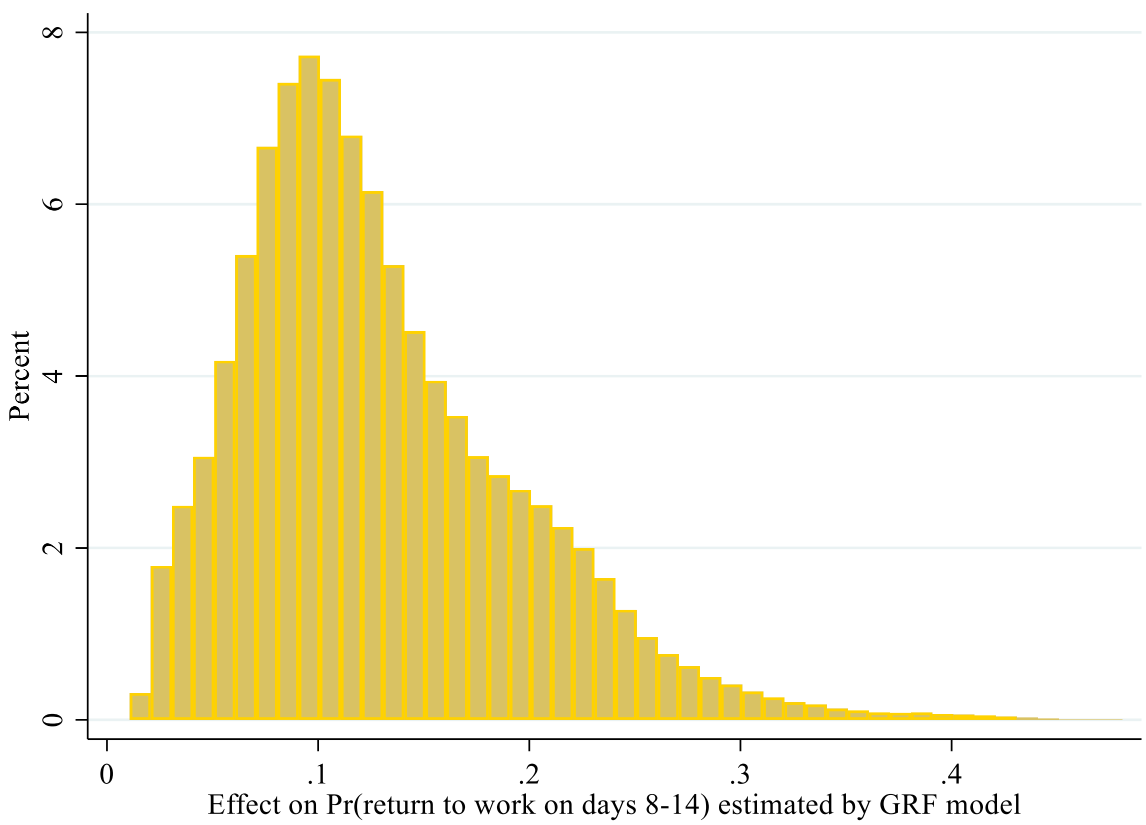
Note: Spells of test set workers ranked according to different rules, rank-average treatment effects estimated based on how treatment effects among those at or below different quantiles according to each ranking differ from the average treatment effect in the sample. Gains of each prioritisation rule assessed by QINI, which reweights the area under the targeting operator characteristic curve to give equal weight to observations at different quantiles. *GRF*: ordering spells from highest to lowest $\hat{\tau}_x$. *Past sick leave*: ordering spells based on worker's pre-experiment days of sick leave, highest first. *Social payment share of income*: ordering spells based on worker's social payment share of income, highest first. *Health industry and social payment share of income*: ordering workers based on whether they work in the health industry, and by social payment share within health industry and non-health industry, highest social payment share first within the industry groups. *LASSO*: ordering spells from highest to lowest $\hat{\tau}_x$ as estimated by a LASSO model with quadratic terms and first-order interactions between all variables, as well as a quartic in age. *OLS*: ordering spells from highest to lowest $\hat{\tau}_x$ as estimated by an OLS model with quadratic terms and first-order interactions between all variables, as well as a quartic in age. Only observations with non-missing colleague characteristics are included in the sample on which LASSO and OLS are estimated. These models excluding variables which capture partner characteristics, as discussed in Appendix A.

Figure B1. Survival and hazard rates for spells of workers in Gothenburg and Jämtland in the first and second halves of 1989.



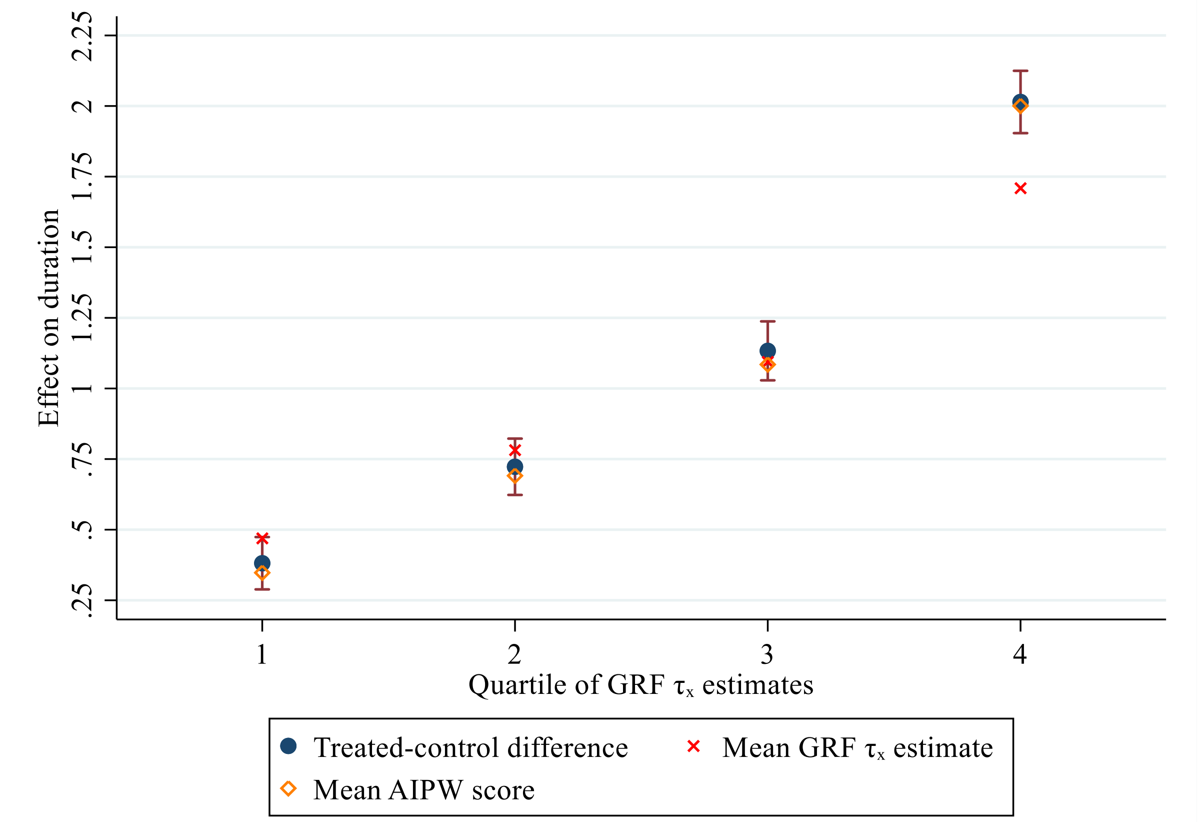
Note: Spells which began between January 1st and June 31st and July 1st and December 31st 1989 respectively. The hazard rate represents the probability that a spell which has been ongoing for a given number of days ends on the next day.

FIGURE B2. THE DISTRIBUTION OF PREDICTED TREATMENT EFFECTS ON THE PROBABILITY OF A SICKNESS ABSENCE SPELL ENDING ON DAYS 8-14.



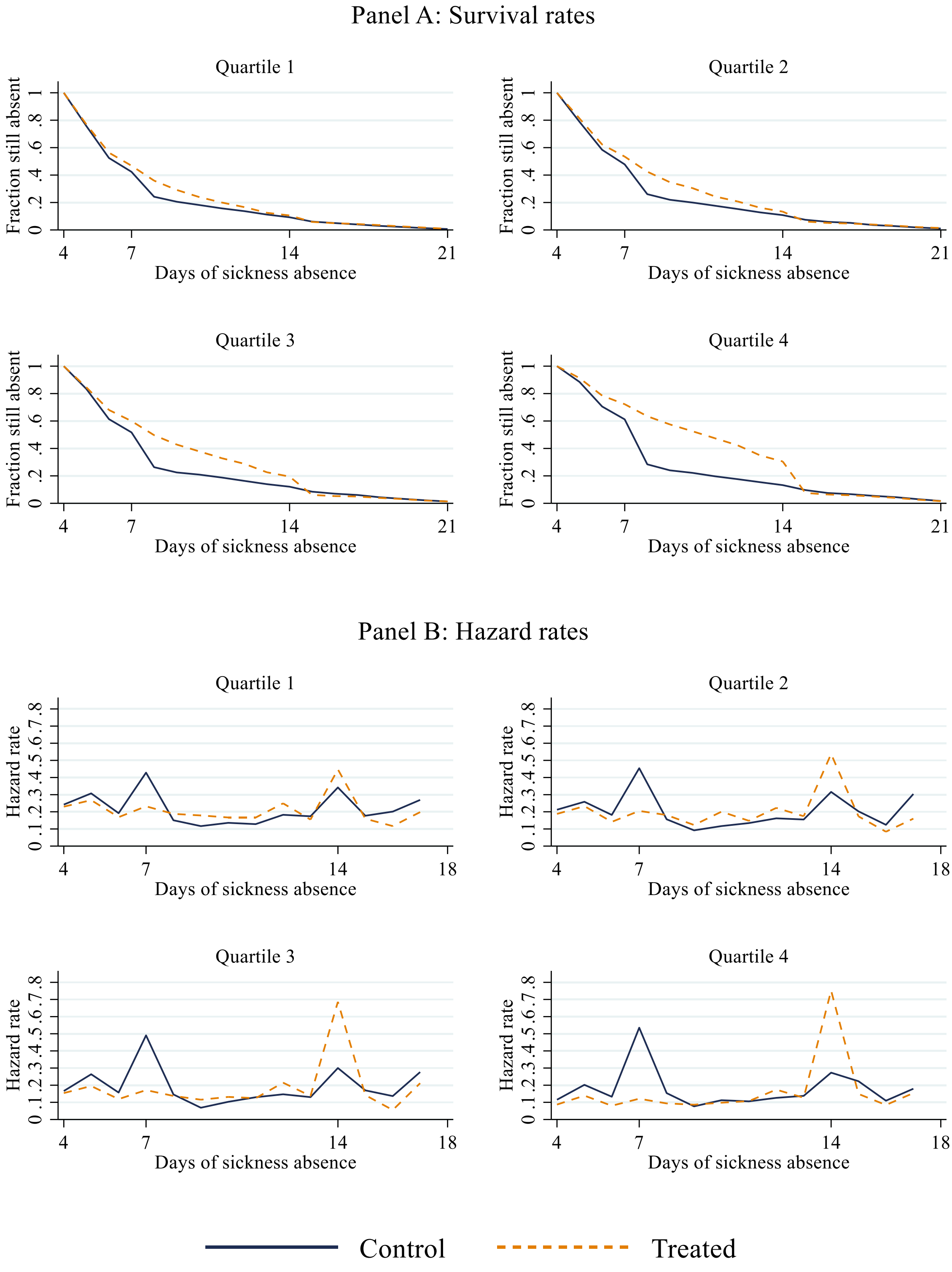
Note: Estimates for spells of training set workers. Bin width = 0.01.

FIGURE B3. CALIBRATION OF THE MODEL TRAINED ON WORKERS IN GOTHENBURG.



Note: Spells of trainings set workers (i.e., workers who lived in Gothenburg during the experiment). Quartiles defined by ranking spells based on treatment effects estimated by GRF. Q1 contains spells estimated to be least affected and Q4 spells estimated to be most affected. Treated-control differences in duration within each quartile estimated as $\hat{\tau} = \bar{y}_i|(W_w = 1) - \bar{y}_i|(W_w = 0)$. Effects for training set workers; out-of-bag estimates of $\hat{\tau}_x$ and AIPW scores. Confidence intervals at the 95 percent level for the treated-control differences.

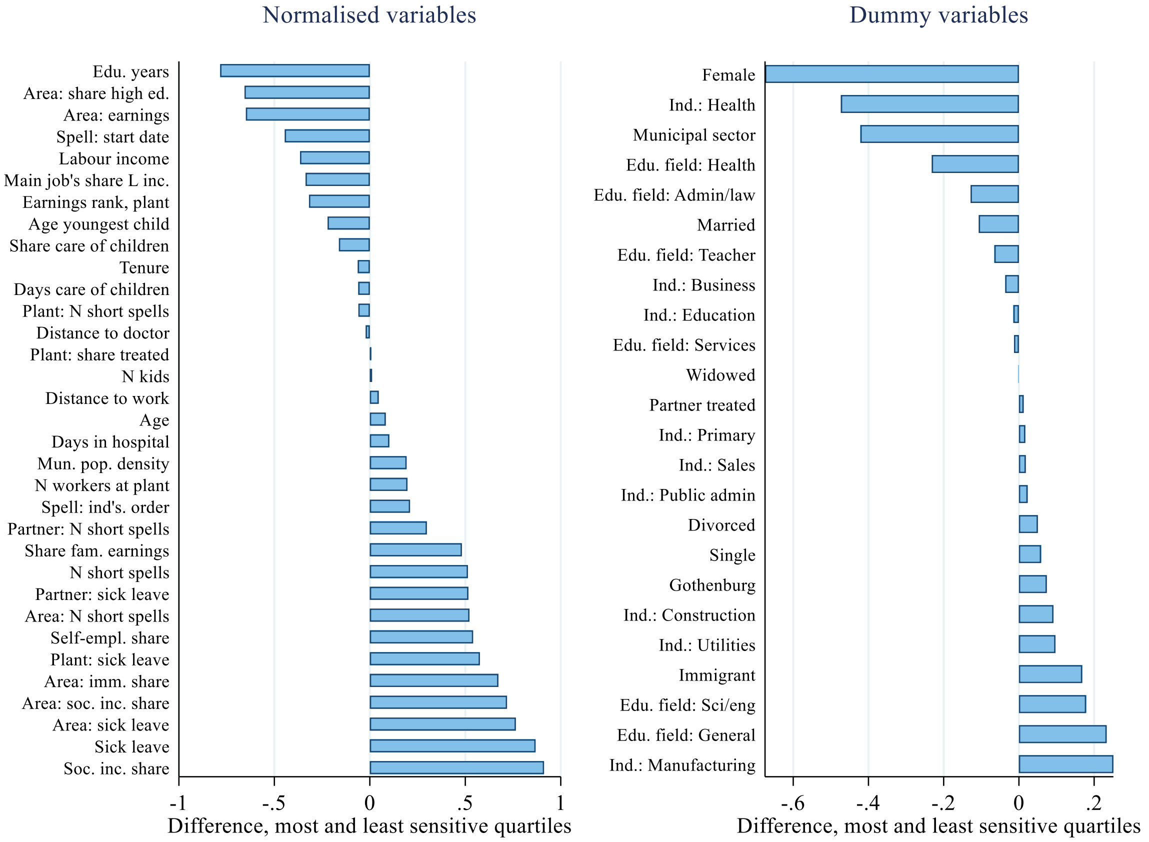
FIGURE B4. SURVIVAL AND HAZARD RATES OF ABSENCE SPELLS OF WORKERS IN JÄMTLAND, RANKED BY TREATMENT EFFECTS ESTIMATED USING THE MODEL TRAINED ON WORKERS IN GOTHENBURG


Panel A: Survival rates

Panel B: Hazard rates

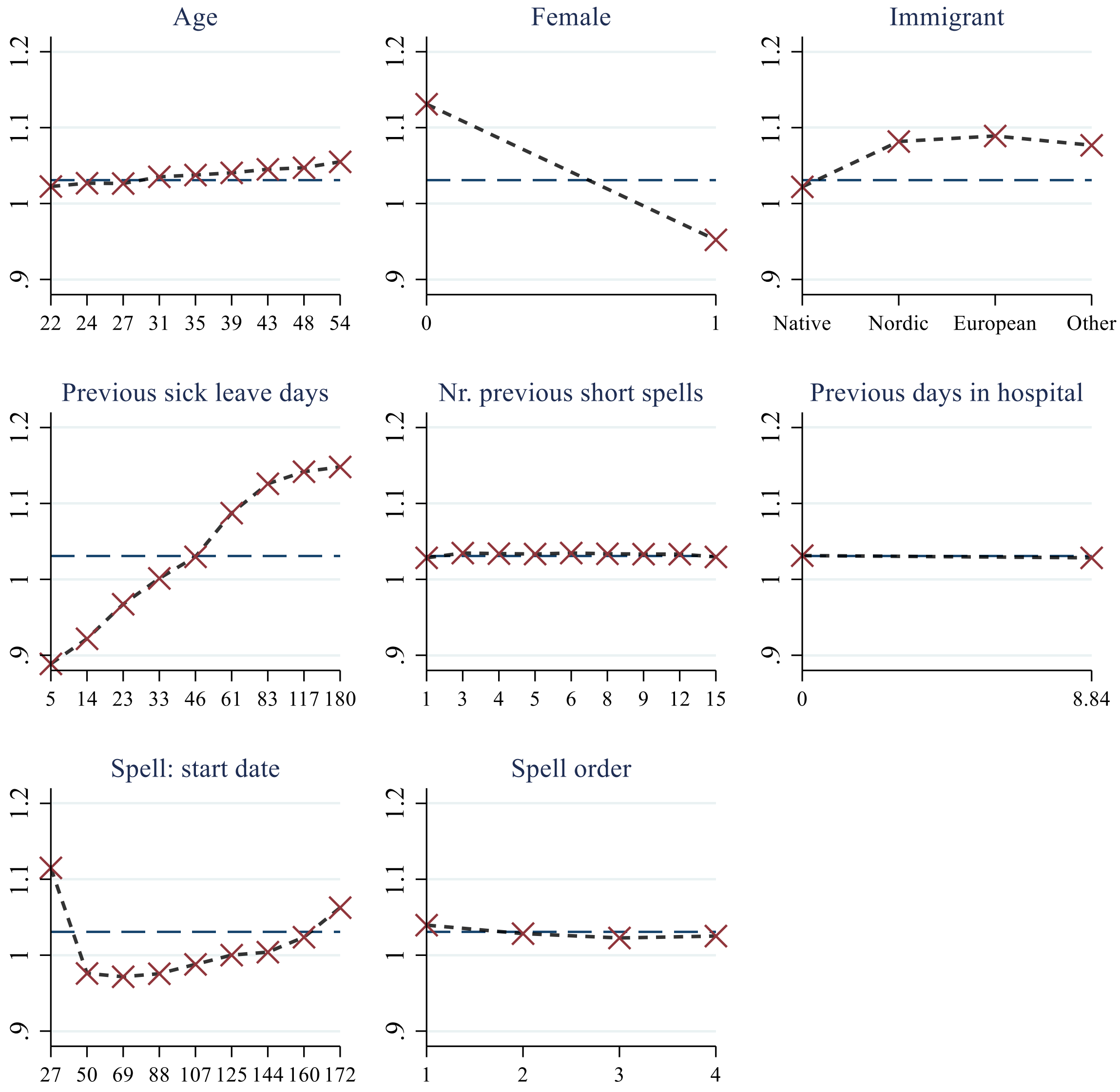
Note: Rates for absence spells of workers who lived in Jämtland during the experiment. The hazard rate represents the probability that a spell that has been ongoing for a given number of days ends on the next day. Workers divided into quartiles based on monitoring sensitivity predicted using the spells of workers who lived in Gothenburg. Q1 contains spells estimated to be least affected and Q4 spells estimated to be most affected. Controls born on odd dates, treated on even dates.

FIGURE B5. DIFFERENCES IN TERMS OF INDIVIDUAL AND SPELL CHARACTERISTICS BETWEEN THE QUARTILES MOST AND LEAST RESPONSIVE TO MONITORING DEFINED BY PROBABILITY OF RETURNING TO WORK ON DAYS 8-14.



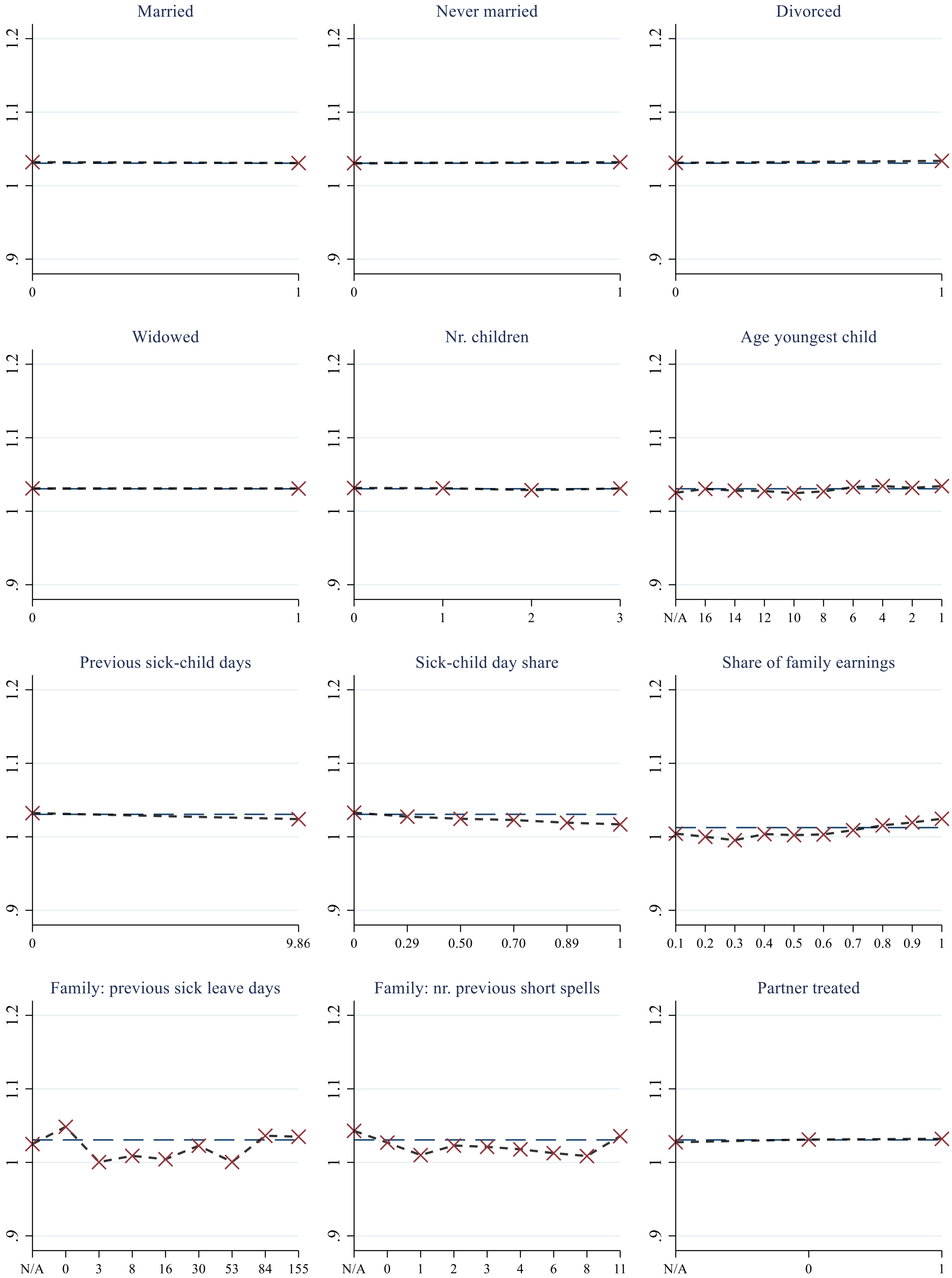
Note: Quartiles defined based on GRF $\hat{\tau}_x$ so as to include equal numbers of spells. Continuous variables normalised to have mean zero and standard deviation one. For dummies, differences are based on shares with the given characteristic. Positive differences mean higher values among the most sensitive (Quartile 4), negative differences mean higher values among the least sensitive (Quartile 1). Absence spells of training set workers.

FIGURE B6. PARTIAL DEPENDENCE PLOTS FOR DEMOGRAPHIC AND HEALTH COVARIATES.



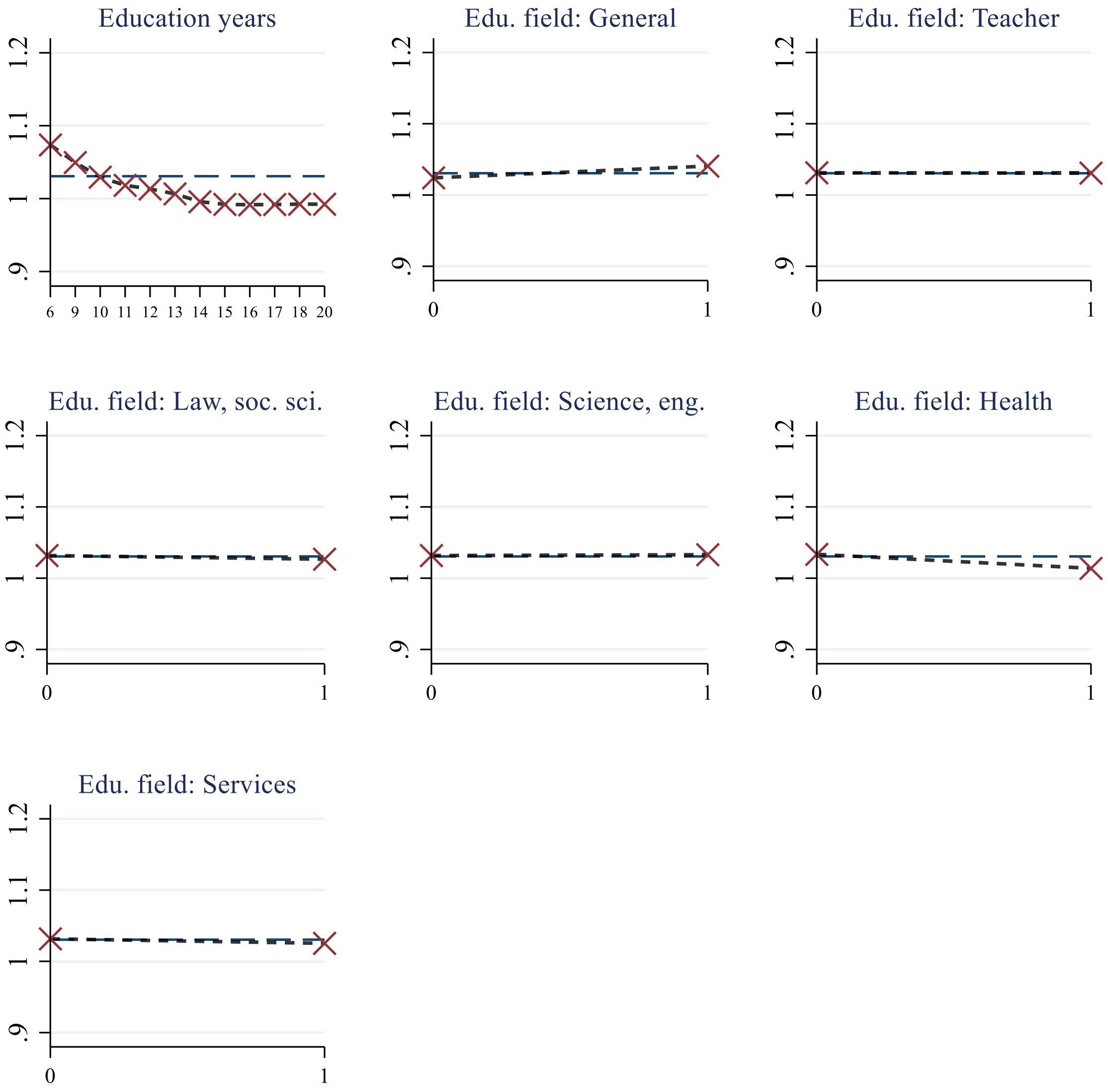
Note: Average increase in spell duration (y-axis) evaluated by the GRF model if the covariate is set to a given value (x-axis) for all workers and spells, while all other covariates are kept at their empirically observed values. Value choice is dictated by values at 1st – 9th deciles for continuous variables, values with five or more percent of observations for variables which are concentrated at a few mass points, and zero and one for binary variables. In the case of variables where most individuals have a value of zero, partial dependence functions evaluated at zero and at the average value among those with nonzero values. Dashed navy line indicates mean $\hat{\tau}_x$ estimate of the baseline model.

FIGURE B7. PARTIAL DEPENDENCE PLOTS FOR FAMILY COVARIATES.



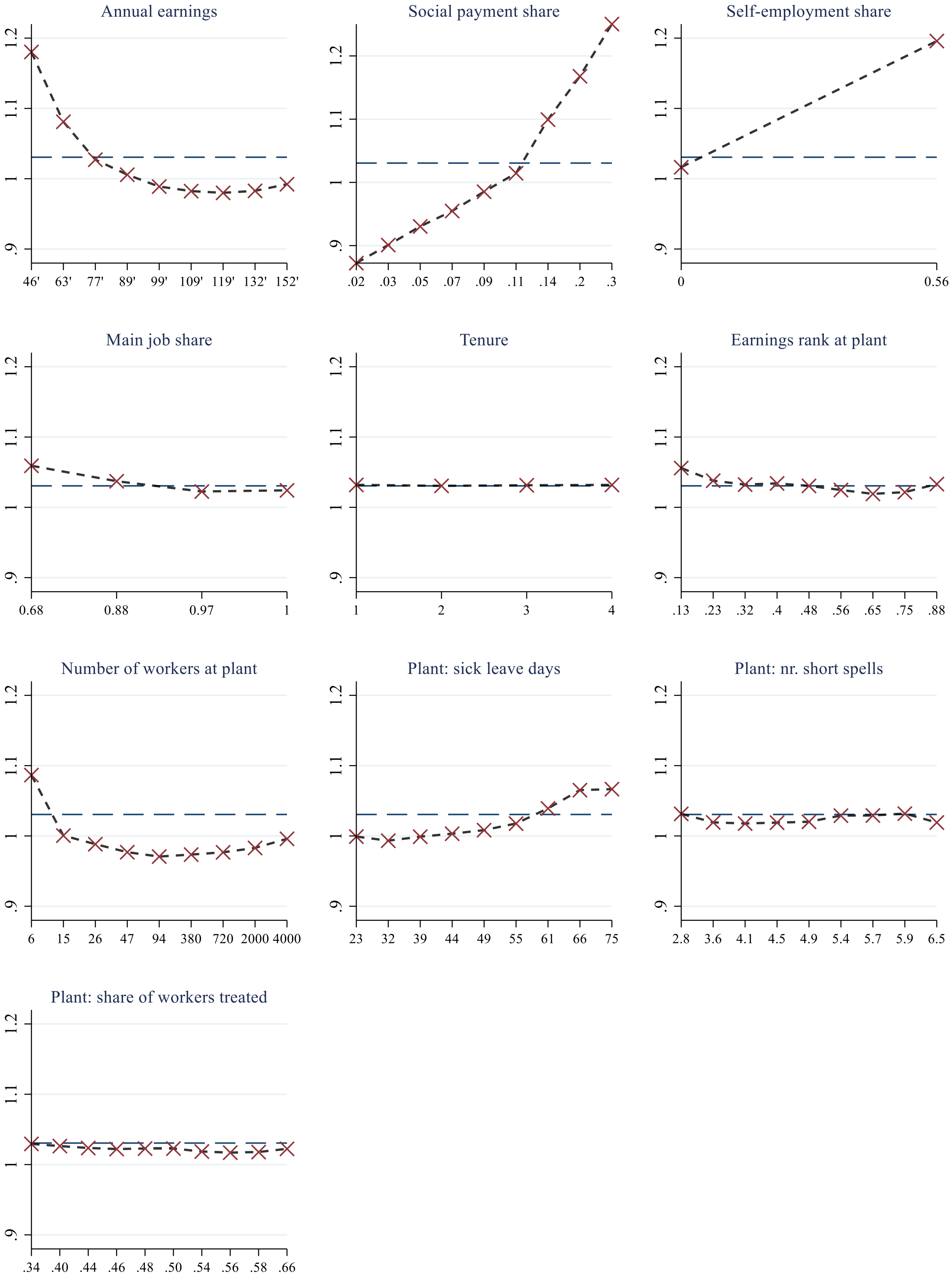
Note: Average increase in spell duration (y-axis) evaluated if the covariate is set to a given value (x-axis). Value choice dictated by values at 1st – 9th deciles for continuous variables, values with five or more percent of observations for variables concentrated at a few mass points, and zero and one for binary variables. For variables where most individuals have a value of zero, partial dependence functions evaluated at zero and at the average value among those with nonzero values. If a covariate is missing for many workers, the partial dependence function is also evaluated when it is set to missing. Dashed navy line indicates mean $\hat{\tau}_x$ estimate of the baseline model.

FIGURE B8. PARTIAL DEPENDENCE PLOTS FOR EDUCATION COVARIATES.



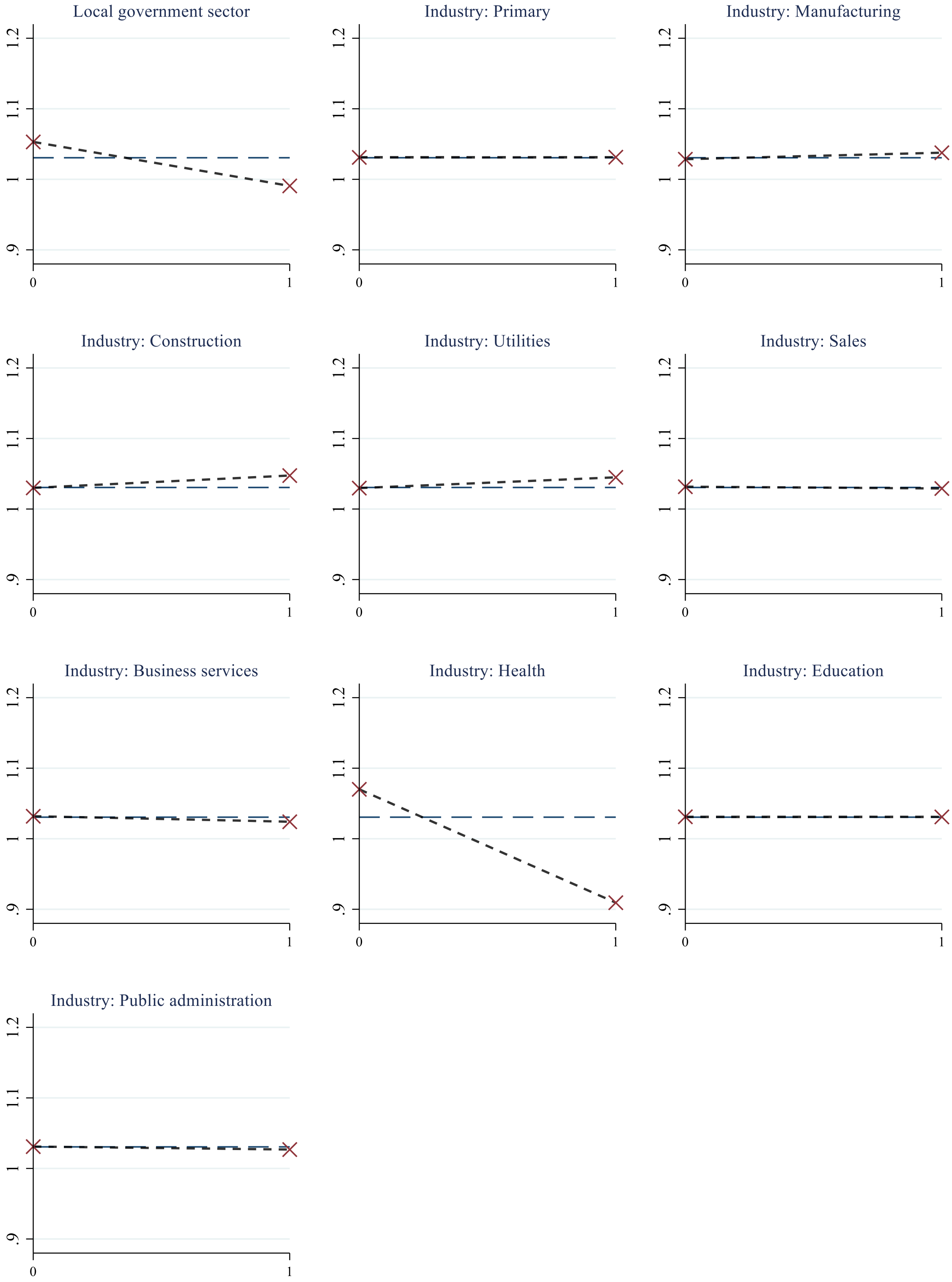
Note: Increase in spell duration in days (*y*-axis) evaluated when the covariate takes on its different possible values (*x*-axis). Individuals with >2 pre-school age children and >3 school age children present in the data, but effects not evaluated due to their small proportion. For population density, each category represents the density in one municipality, ordered from least densely populated to most densely populated. Dashed navy line indicates mean $\hat{\tau}_x$ estimate of the baseline model.

FIGURE B9. PARTIAL DEPENDENCE PLOTS FOR WORK COVARIATES.



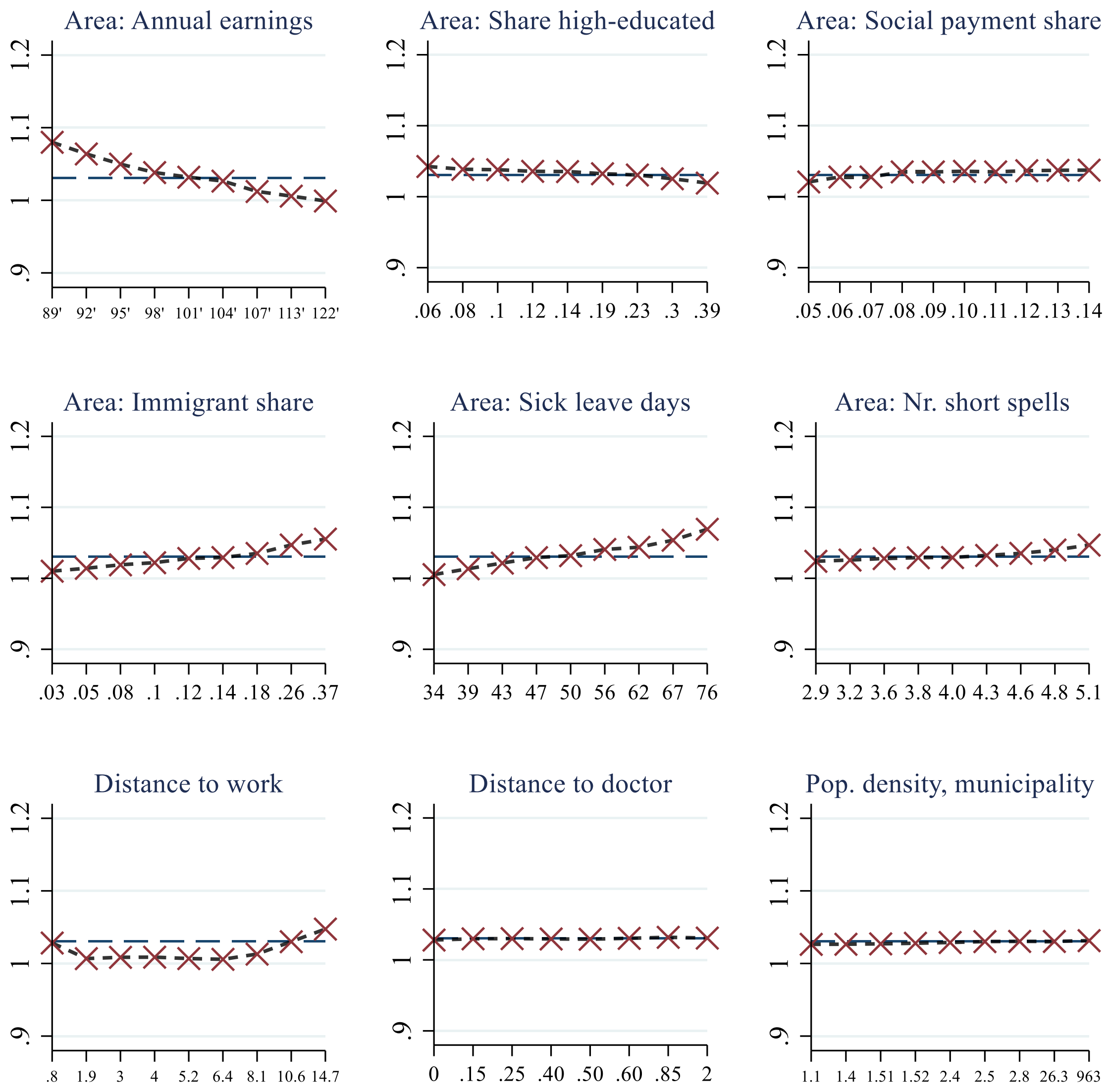
Note: Increase in spell duration in days (*y*-axis) evaluated when the covariate takes on its different possible values (*x*-axis). Individuals with >2 pre-school age children and >3 school age children present in the data, but effects not evaluated due to their small proportion. For population density, each category represents the density in one municipality, ordered from least densely populated to most densely populated. Dashed navy line indicates mean $\hat{\tau}_x$ estimate of the baseline model.

FIGURE B10. PARTIAL DEPENDENCE PLOTS FOR SECTOR COVARIATES.



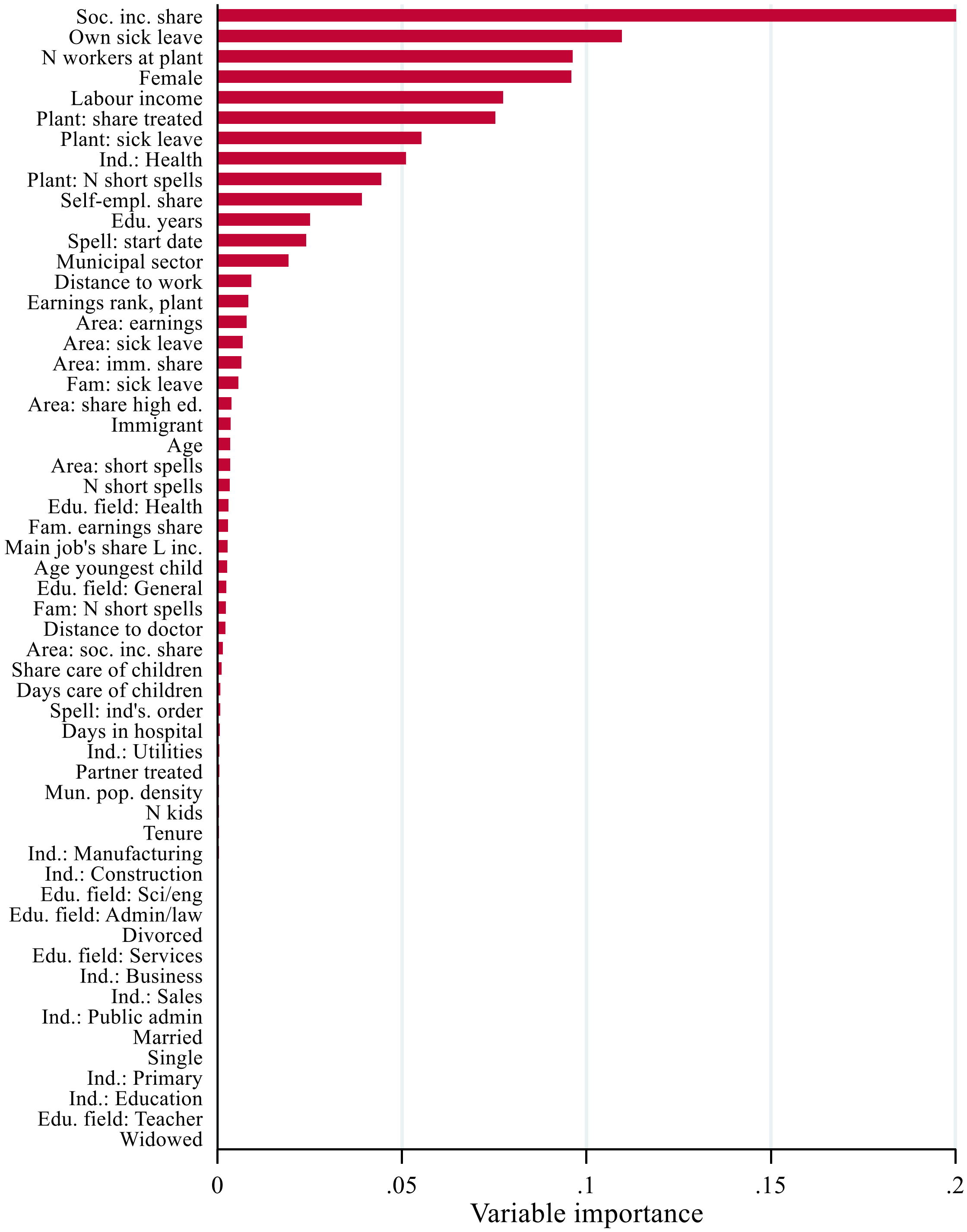
Note: Increase in spell duration in days (*y*-axis) evaluated when the covariate takes on its different possible values (*x*-axis). Individuals with >2 pre-school age children and >3 school age children present in the data, but effects not evaluated due to their small proportion. For population density, each category represents the density in one municipality, ordered from least densely populated to most densely populated. Dashed navy line indicates mean $\hat{\tau}_x$ estimate of the baseline model.

FIGURE B11. PARTIAL DEPENDENCE PLOTS FOR LOCATION COVARIATES.



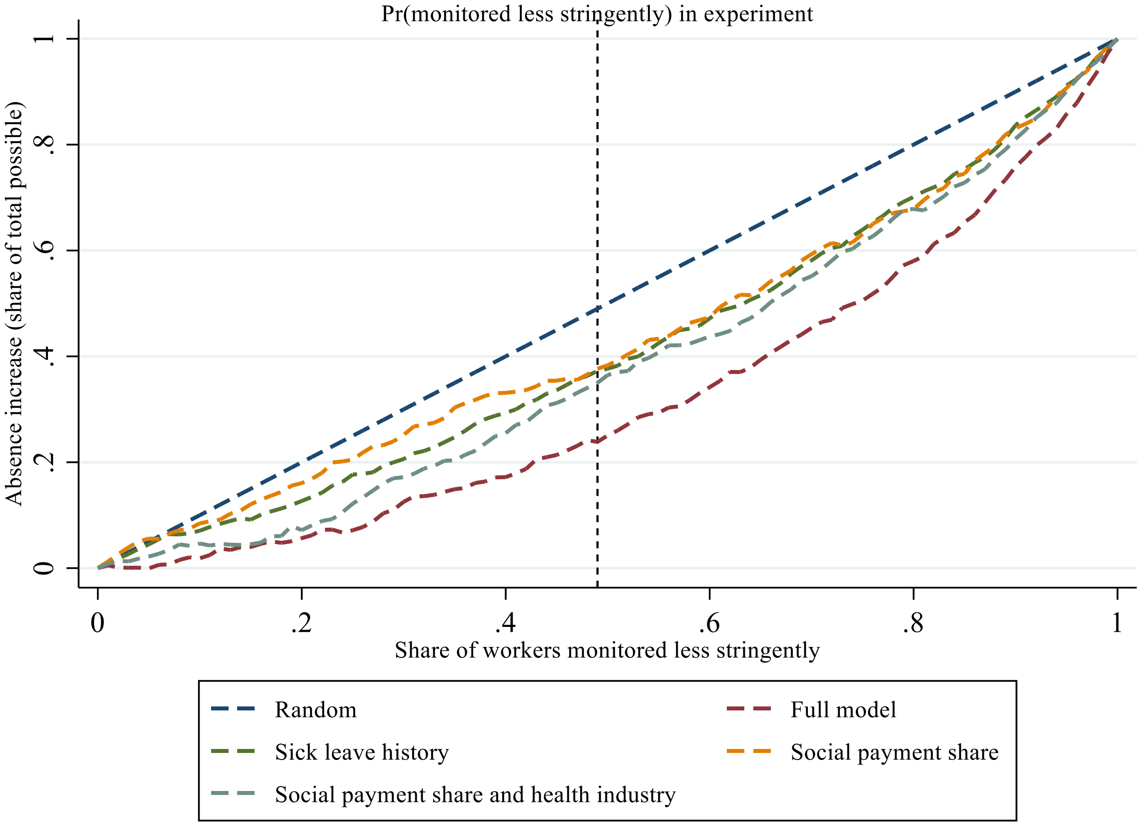
Note: Increase in spell duration in days (y-axis) evaluated when the covariate takes on its different possible values (x-axis). Individuals with >2 pre-school age children and >3 school age children present in the data, but effects not evaluated due to their small proportion. For population density, each category represents the density in one municipality, ordered from least densely populated to most densely populated. Dashed navy line indicates mean $\hat{\tau}_x$ estimate of the baseline model.

FIGURE B12. VARIABLE IMPORTANCE ACCORDING TO THE GRF PACKAGE'S SIMPLE BUILT-IN MEASURE



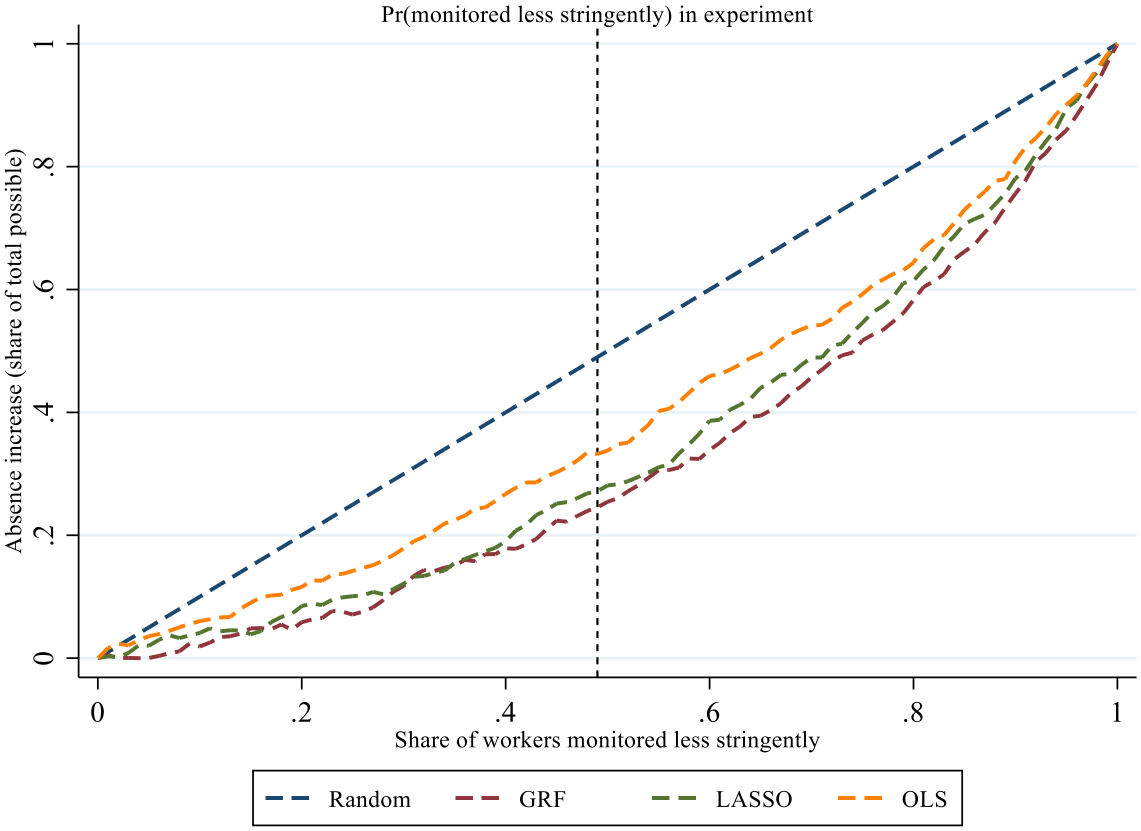
Note: Importance is measured as share of splits up to depth 4 within the trees (ignoring splits made at further depths). Splits at lower depth d given two times the weight of those at $d + 1$. Total importance sums to 1.

FIGURE B13. COMPARISON OF TARGETED MONITORING POLICIES BASED ON THE FULL MODEL AND ON DIFFERENT SIMPLE RULES



Note: Increase in sickness absence when monitoring is relaxed for different shares of workers according to different prioritisation rules. Spells of workers in the test set. Monitoring based on GRF assumes spells are ranked based on their estimated GRF $\hat{\tau}_x$ and monitoring is relaxed for those with lower estimated treatment effects first. Monitoring based on sick leave history and social payment share assumes workers are ranked from lowest to highest past sickness absence and social payment share respectively and monitoring is relaxed for those with lower past sickness absence or social payment share first. Order of monitoring among those with equal numbers of days of sick leave in the past or with equal social payment share is random. Monitoring based on social payment share and health industry is suggested by optimal policy trees of depth 2 (Athey and Wager, 2021) and involves relaxing monitoring first for health industry workers based on their social payment share, and then for non-health industry workers based on their social payment share. Pr(monitored less stringently) in experiment = 0.49.

FIGURE B14. COMPARISON OF TARGETED MONITORING POLICIES BASED ON THE GRF MODEL AND ON LASSO AND OLS MODELS



Note: Increase in sickness absence when monitoring is relaxed for different shares of workers according to different prioritisation rules. Spells of workers in the test set who are employed at establishments with more than one worker, as explained in Appendix A. Details of how the LASSO and OLS models are estimated are provided in Appendix A. Monitoring based on GRF, LASSO and OLS assumes spells are ranked based on their estimated $\hat{\tau}_x$ from these models and monitoring is relaxed for those with lower estimated treatment effects first. Pr(monitored less stringently) in experiment = 0.49.