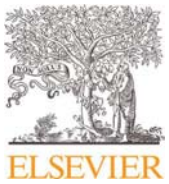
Journal of Health Economics 83 (2022) 102615



Contents lists available at ScienceDirect

Journal of Health Economics

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

Effects of dialogue meetings on sickness absence—Evidence from a large field experiment☆

Matteo Alpino[a], Karen Evelyn Hauge[b], Andreas Kotsadam[b,*], Simen Markussen[b]

[a] Bank of Italy, Italy
[b] Ragnar Frisch Centre for Economic Research, Norway

ARTICLE INFO

Keywords:
RCT
Sickness absence
Employment
Field experiment

ABSTRACT

Sickness absence entails large individual and societal costs. Dialogue Meetings (DMs) where the absentee, the employer, and the physician discuss arrangements for full or partial work resumption have been in place in Norway since 2007. In collaboration with the Labour and Welfare Administration, we conducted a large-scale, pre-registered, randomized field experiment to evaluate aspects of the Norwegian DMs policy. We do not find statistically significant effects of summoning to a meeting and we can reject even small threat (notification) effects of sending out letters. We also conduct an extensive search for heterogeneous treatment effects but find no evidence of these.

1. Introduction

Sickness absence implies high costs for individuals, firms, and society. Studies of work absence due to illness find that it is possible to design policies that reduce absenteeism. In particular, studies show that financial incentives (Böckerman et al., 2018; Duflo et al., 2012; Fevang et al., 2014; Johansson and Palme, 2005; Kostol and Mogstad, 2014; Ziebarth and Karlsson, 2010), monitoring (De Jong et al., 2011; De Paola et al., 2014; Godard et al., 2020; Hartman et al., 2013), peer and socialization effects (Bradley et al., 2007; Dahl et al., 2014; Hesselius et al., 2009; Ichino and Maggi, 2000; Markussen and Røed, 2015), and employment protection (Ichino and Riphahn, 2005; Olsson, 2009) affect absence. In addition, early identification and intervention are generally considered important in reducing sickness absence, especially with increased employer responsibility (Autor and Duggan, 2010).

One way of potentially reducing sickness absence is dialogue meetings (DMs). In such meetings, the absentee, the employer, the social insurance administrator, and often the physician discuss whether arrangements can be made at the workplace that make full or partial work resumption possible (e.g., to alter the nature of the tasks at work or changing teams). Norway introduced such a policy in 2007, and since then, around 50.000 meetings are held each year, typically when the absence spell has lasted around 26 weeks. Similar systems of meetings between the employer, the employees and the social security administrators exist in Austria, Denmark, Finland, Germany, the Netherlands, and Sweden (Belin et al., 2016).

The existing literature on the effects of such meetings on work resumption or reduced absenteeism is mixed. Markussen et al. (2017) evaluate DMs using observational data and by exploiting differences across Norwegian counties in the

☆ We would like to thank Ole Rogeberg, Pathric Hägglund and Per Johansson for valuable suggestions. We thank the Norwegian Labour and Welfare Administration for the cooperation in conducting the experiment and providing data and Tao Zhang for programming the randomization page. Funding from the Norwegian Research Council (grant number 259512) is acknowledged. An analysis plan is pre-registered at the AEA RCT registry (number AEARCTR-0002730) and all deviations from the plan are noted in the text. The pre-analysis plan can be found **here**. The views expressed in this paper do not necessarily reflect those of the Bank of Italy.

* Corresponding author.
E-mail address: andreas.kotsadam@frisch.uio.no (A. Kotsadam).

https://doi.org/10.1016/j.jhealeco.2022.102615
Received 28 April 2021; Received in revised form 16 March 2022; Accepted 28 March 2022
Available online 7 April 2022

propensity to use such meetings at different times. They find that DMs substantially reduce sickness absence. Furthermore, they find that this reduction works via both a pure notification effect and an attendance effect. Based on these two effects, they find that holding a DM speeded up complete return to work by approximately ten days, where around half of this reduction stemmed from absentees returning to work before the meeting date. A field experiment conducted in Sweden on individuals with weak labor market attachment analyzed the effects of meetings that provide information about the rehabilitation procedure and the sickness benefit claimants' rights and duties (Johansson and Lindahl, 2013) The authors find that such meetings reduced sickness absence. In a different randomized field experiment, Engström et al. (2016) test the effects of two types of meetings in Sweden. As for the first type, the caseworker and the absentee meet to assess the absentee's working abilities and rehabilitation needs. The authors find that this intervention caused a lock-in effect with increased take-up of both sickness and disability benefits. The second type of intervention, the most similar to the one analyzed in the present work, is a meeting between the absentee, the social insurance administrator, and the employer to discuss the possibility of alternative working tasks. They find no statistically significant effects of this intervention, but the intervention was quite weak in the sense that it only affected the likelihood of meetings by 4.4 percentage points.

In order to evaluate the Norwegian DM policy and in particular to explore treatment heterogeneity with respect to both the intervention design (timing of letters and meetings), and individual characteristics, we teamed up with the Norwegian Social Insurance Administration, called the Norwegian Labour and Welfare Administration (NAV), to conduct a large-scale randomized field experiment. This real-time experiment included more than 10 000 absence-spells and was conducted at 13 different local social insurance offices starting from September 2016 and ending in July 2018. Part of the motivation behind the experiment was to investigate whether early meetings would reduce absence more than late meetings and whether the timing of the meeting notification matters. We evaluate the experiment using administrative data such that there is no attrition, self-reporting, or other sources of bias involved. Individuals were randomly assigned to either the treatment group or control group. The treatment group received a letter summoning them to a meeting at a specific time some weeks ahead. In contrast, the letter to members of the control group stated that they were not summoned to any meeting but that they had the right and opportunity to ask for a meeting to be organized.[1] We further randomized the timing both of when letters were sent and meetings were held, in order to explore heterogeneity in these dimensions.

Our first contribution is to conduct a large-scale, pre-registered randomized controlled trial of dialogue meetings. While designing the experiment, we expected to learn whether the letters had a pure notification (threat) effect even before any meeting was held, and whether it was possible to target individuals based on observable characteristics. In addition, we expected to learn whether the timing of meetings and letters matters for effectiveness. We pre-specified three main hypotheses (AEARCTR-0002730): 1) summoning to a DM via letter causes a reduction in sickness absence; 2) letters by NAV summoning to a DM induce a threat effect, such that absentees return to work before the meeting is held; and 3) letters by NAV not summoning to a DM induce a reversed threat effect, such that absentees are less likely to return to work.

As we will document below, we cannot reject the null-hypothesis for these three hypotheses. Furthermore, our results can be summarized as follows: First, our experiment increased the likelihood of participating in a meeting: across all treatment arms, the fraction participating in a DM was 14 percentage points higher among individuals summoned to a meeting, than among those who were not. Second, individuals summoned to a DM (treatment group) had on average 3,1 days shorter absence spells than those not invited (control group). However, the difference is not statistically significant from zero (p-value: 0.176). Using an equivalence testing approach of two one-sided t-tests (TOST), and a 5 percent significance level, we can reject reductions larger than 6.9 days and increases larger than 0.7 days.[2] Third, we find no evidence of a notification (threat) effect: the return-to-work rate does not increase among workers summoned to a DM (but before the DM is held), relative to those who have not received any letter yet. Finally, we do not find any evidence of a reversed threat effect: the return-to-work rate does not decrease among individuals who receive a letter that does not summon to a meeting, relative to those who have not received any letter yet.

Our second contribution concerns treatment effect heterogeneity. For policy design, it is important to establish which groups benefit (or are hurt) most from a policy, and what details of the intervention (such as its timing) matter for effectiveness, and for whom. In our setting, such heterogeneity is highly relevant because the number of meetings is rationed due to caseworker capacity and because NAV has the possibility to target individuals at different points of time during the absence spell. Although randomized controlled trials are an excellent tool for recovering average treatment effects without bias, testing for treatment effect heterogeneity is a much harder problem (Athey and Imbens, 2017). The main challenge is that as the number of heterogeneity tests increases, one can be almost certain to find statistically significant differences in treatment effects along some dimensions. As practitioners of applied econometrics, we know by first-hand experience how easy it is to convince oneself and others that such "effects" have meaningful interpretations. Correcting for multiple hypotheses testing is helpful only to some extent, as it is impossible for readers (and sometimes for the econometrician) to know how many tested hypotheses are not reported in the final paper. In addition, multiple hypothesis testing is likely conservative when used across more than a few dimensions of heterogeneity (Davis and Heller, 2017).

In the present work, we follow two strategies to test for treatment heterogeneity. Our first approach is to rely on a detailed pre-analysis plan. Before obtaining all the experimental data, we specified how we would search for treatment effect heterogeneity, along which dimensions, and with what specifications. The corresponding results show no - or very weak - evidence of treatment heterogeneity. As an alternative approach, we rely on a split-sample methodology: first, we divide the sample in two random sub-

[1] The experimental design reflects that all individuals on sick-leave have the legal right to have a DM.

[2] This estimate has an intention-to-treat (ITT) interpretation, so it is not directly comparable to the finding in Markussen et al. (2017), who estimate the effect of holding a meeting. When scaled by the first stage to increase comparability, our findings are very imprecise but not inconsistent with the previous ones.

samples; then, we search extensively for heterogeneity in one sub-sample; finally, we try to replicate the findings in the hold out sub-sample. If the results do not replicate, it is likely that the heterogeneity dimensions found to be significant in the training sample capture idiosyncratic noise in that sample (overfitting). By using machine-learning techniques, this approach can be generalized and repeated extensively. In our case, we use honest random forests (Athey et al., 2019; Wager and Athey, 2018) to confirm the results described above, which point to no robust evidence of heterogeneity in the treatment effects of dialogue meetings.

The remaining of this paper is organized as follows. In Section 2, we present the experiment and the data used. In Section 3, we first present the differences between the treatment and control groups in conducting meetings and then the main results for the effects of dialogue meetings for absence duration and return to work. Section 4 concerns treatment effect heterogeneity while Section 5 concludes.

2. The field experiment and data used for evaluation

2.1. Institutional setting

Norwegian workers enjoy a relatively generous sick leave insurance system. All workers are entitled to sick leave benefits for up to 1 year, after a qualifying period of four weeks of employment. The first few days (3 or 8) are self-certified, whereas a physician must certify the remaining period. The income replacement rate is 100 percent up to an income threshold. The employers cover the first 16 days, whereas the social insurance administration pays for the remaining period. In addition, the insurance system allows combining working less than 100 percent with sick leave, called graded sick leave. For example, a graded sick leave of 30 percent would imply that the employee works 70 percent of his/her work hours and is on sick leave for the remaining 30 percent. Norway has consistently one of the highest sick-leave rates in Europe, around 6 percent. Measures for reducing absenteeism and promoting a return to work has thus been high on the political agenda for many years. The use of dialogue meetings (DMs) is one such measure.

2.2. The use of dialogue meetings

A system of DMs has been in place in Norway since 2007, and the intention of DMs is explicitly to induce long-term absentees to return fully or partly to work. The Norwegian Labour and Welfare Administration (NAV) initiates and organizes DMs by inviting the sick-listed employee, the employer, and the physician to a meeting. In addition, a caseworker from NAV attends and leads the meeting. NAV caseworkers work towards reducing sick leave and report on sick leave spells and length of sick leave spells, and have incentives to find solutions helping the sick-listed workers back to work. The dialogue meetings take place within 26 weeks of the sickness spell. In these meetings, the parties examine the situation and plan how to help the absentee get back to work. The meetings are mandatory, and the absentee may lose her sickness benefits if not attending. While the meeting does not result in a binding agreement, the NAV caseworker writes a summary of the meeting and there is a general goal that the participants should agree on how to move forward, who is supposed to do what, and a general time frame.

All absentees have the right to meet before week 27 of the sickness spell, but there are many exceptions in practice. First of all, most health conditions improve naturally such that absentees return to work before week 27 or before the meeting date and therefore do not have a meeting. Second, certain more critical medical reasons may exempt an absentee from attending DMs (such as receiving cancer treatment or being admitted to health institutions). Third, the capacity and procedures of calling into DMs differ between NAV offices across the country. At the time of the experiment, NAV did not summon all absentees to DMs. Around 50.000 DMs take place each year. In 2012 this amounted to an estimated cost of 40.5 million USD (237.5 million NOK).

2.3. The field experiment

We conducted a large-scale, pre-registered, randomized field experiment in cooperation with the Norwegian Labour and Welfare Administration (NAV). We collaborated with 13 NAV offices (10 in the capital city Oslo) who were willing to change their targeting procedures for the dialogue meetings. We started the trial in September 2016 with one office and then expanded it. The experiment ended in July 2018, and we received the final data in December 2018.

The experimental procedure consisted of several steps. Every week, each NAV office identified employees on sick leave who were in their eighth week of the sickness spell, and assigned them to caseworkers.[3] After that, the caseworkers logged in to a secure internet page; for each absentee, they wrote in the identification number and answered two questions about the case. At that point the internet page randomly assigned the absentee in one of eight different treatment arms, shown in Fig. 1. Finally, the caseworker sent the absentee a letter whose content and timing was determined by the outcome of the random draw.

Ideally we would have liked to randomize whether absentees attended a meeting or not. However, all absentees have the right to attend a meeting, and NAV did not want to violate this right. Thus we opted for the following design: in the treatment group, absentees were summoned via letter to a meeting, as in the standard NAV procedure; in the control group, absentees and their employers instead received a letter from NAV with the following information: a) NAV was aware that the employee was on sick leave; b) NAV decided

[3] The NAV offices had different methods for assigning cases to caseworkers. While some offices assigned the same number of cases to every caseworker based on birth dates, other offices assigned cases unequally so that some caseworkers had more cases than others, while others still assigned specific industries or firms to caseworkers. Our caseworker survey shows that around 29 percent of the cases were assigned based on firms.

Letter is sent in:	not summoned to a meeting	summoned to a meeting to be held in:
week 9		week 13
week 15		week 19
week 15		week 26
week 22		week 26

Fig. 1. Summary of the treatment arms.

not to summon her/him to a meeting; c) employees, employers and physician could request a DM if they found it necessary.[4] Note that being assigned to the control group did not have any consequences for the entitlement to sick leave benefits, neither in terms of amount nor length. The design of the experiment allows us to estimate the effect of being summoned to a dialogue meeting relative to not being summoned, while still having the possibility to ask for one. Although our purpose was to influence the probability of having meetings, the two types of letters could also influence the character of the meetings, as we discuss in the conclusion.

Furthermore, the experiment features two additional (random) treatment variations in terms of when the letters were sent, and when the meetings were held. NAV sent both type of letters in weeks 9, 15, or 22 of the absence spell (1, 7, and 14 weeks after the random draw); for individuals in the treatment group, meetings were scheduled in weeks 13, 19, or 26 of the absence spell (5, 11, or 18 weeks after the draw, see Fig. 1). Note that our design features treatment arms with some overlap in terms of timing: letters sent in week 15 summoned to a meeting either in week 19 or 26, while for meetings held in week 26, letters were either sent in week 15 or 22. The overlap allows us to test whether receiving the letter summoning to a meeting itself had an effect before the DM takes place (what we term notification or threat effect). In fact, if there is a strong and immediate notification effect, we would expect people who get earlier letters to have shorter sickness spells. Conversely, the variation in letter timing in the control group is useful to test whether people receiving the letter that does not summon to a meeting stay on sick leave longer, perhaps because they feel less monitored (what we term a reverse notification effect).

The NAV offices had different capacities and preferences regarding how many meetings to arrange. We had as our baseline case an equal probability of being assigned to each of the eight treatment arms,[5] i.e. a 12.5 percent chance of each cell (see Fig. 1). Some offices, however, wanted different combinations. In particular, many offices wanted a slower rollover with only late meetings to begin with. Other offices needed to change the number of meetings over time due to personnel shortages. Only one office wanted more than 50 percent of the individuals in the treatment group.

2.4. Data and coding of variables

Our final analysis sample was received from NAV in December 2018 and contains 10,235 individuals. In addition we used a smaller sample of 1627 individuals for an early training sample. The split-sample approach is described in detail in Appendix Section A.2.

2.4.1. Data sources

This study has three main data sources: i) the randomization data set, ii) caseworker surveys, and iii) outcomes and covariates delivered by NAV. The randomization data set contains information about the outcome of the random draw, that is, in which treatment arm each participant ended up. There are two types of caseworker surveys. First, at randomization (week 8 of the sick leave spell) but prior to learning the outcome of the draw, each caseworker is asked two questions about each absentee: 1) “Based on the information you have, how many more weeks do you think the sickness absence will continue?”. The answer alternatives are: less than 4 weeks, between 4 and 11 weeks, between 12 and 20 weeks, more than 20 weeks, or do not know. 2) “In your opinion, how important is it that the absentee is invited to a meeting?”. The answer alternatives are: Very important, important, neither important nor unimportant, not

[4] The letter to the employee read: “NAV has registered that you are on sick leave. NAV is responsible for considering whether dialogue meetings are necessary. In your case, we have decided not to summon you to a dialogue meeting. If you have questions related to this decision, you are welcome to contact us. Further, we need to emphasize that if you, your employer, or your physician consider a dialogue meeting as necessary, you can contact NAV XXX to make an appointment for a meeting.” The letter sent to the employer read: “NAV has registered that the person concerned is on sick leave. We have, in this case, chosen not to summon to a dialog meeting 2. If you nevertheless consider a dialogue meeting as necessary, or if it is necessary for clarification purposes, please contact NAV XXX.”

[5] Actually there are only seven groups, but we made the control group for letter week 15 double as it is used as a comparison for the two groups summoned to meetings.

Table 1
Descriptive statistics by treatment status.

	Total days		Summoned (DM)		Control	
	Mean	SD	Mean	SD	Mean	SD
Dependent variables						
Total days	163.51	(113.2)	162.56	(112.9)	164.48	(113.5)
Days (within spell)	120.49	(108.7)	120.87	(109.0)	120.09	(108.3)
Graded days	112.10	(102.3)	111.83	(98.3)	112.37	(106.2)
Share returning before...						
13 weeks	0.30	(0.5)	0.30	(0.5)	0.29	(0.5)
19 weeks	0.49	(0.5)	0.49	(0.5)	0.49	(0.5)
22 weeks	0.56	(0.5)	0.55	(0.5)	0.56	(0.5)
26 weeks	0.62	(0.5)	0.62	(0.5)	0.63	(0.5)
Baseline control variables						
Female	0.63	(0.5)	0.63	(0.5)	0.62	(0.5)
Birth year	1974.02	(11.9)	1974.14	(11.9)	1973.90	(12.0)
Days before	139.27	(107.5)	140.16	(108.3)	138.37	(106.6)
Grade	74.95	(26.4)	74.45	(26.4)	75.45	(26.4)
nr employees	516.42	(1535.1)	502.48	(1442.1)	530.49	(1623.6)
Other main heterogeneity variables						
Symptoms	0.40	(0.5)	0.39	(0.5)	0.40	(0.5)
Caseworker predictions						
CW predicted meeting important	0.51	(0.5)	0.52	(0.5)	0.50	(0.5)
CW predicted long spell	0.11	(0.3)	0.12	(0.3)	0.10	(0.3)
N	10,235		5141		5094	

Notes: The sample consists of the analysis sample. *Total days* is measured as the total number of days of sickness absence between the draw and the date of data extraction. *Days (within spell)* is the number of days of sick leave within the current sick leave spell. *Graded days* measures the full-day equivalents of sickness absence between the draw and the date of data extraction. *Days before* measures the total number of days on sickness absence since 2015 up until the date of the draw. *Grade* is the grade of sickness absence at the time of the draw. *nr employees* is the number of employees at the absentee's workplace (at the time of the draw). *symptoms* is the share of absentees classified by the physician as having symptoms rather than diagnoses according to International Classification of Primary Care (ICPC-2). *CW predicted meeting important* is the share of absentees for which the caseworkers (CW) predicted a meeting would be "important" or "very important". *CW predicted long spell* is the share of absentees for which the caseworker predicted having a long spell (over 20 weeks).

that important, not important at all. We include these questions to investigate how caseworker discretion may improve the targeting of the system. Answers to these questions provide information about the caseworkers' prediction of how long each absentee would be absent and the caseworkers' belief about the importance of having a meeting for that specific absentee.

In addition, a sub-sample of the caseworkers provided background information about themselves (gender, age, how long they had worked at NAV, education type) and about how the DMs were conducted (whether a physician participates and where the meeting is held). Caseworkers only answered these questions once and, as we conducted this survey during the middle of the data collection period, we do not have information from all caseworkers. In total, 53 caseworkers answered these survey questions, and the corresponding absentees are 9,147.

Finally, NAV provided data on outcomes (the total number of sickness days, the length of current sickness spell) and covariates (gender, birth year, the number of sickness days since 2015, the number of employees at the workplace of the absentee, graded sickness absence, diagnoses) for the absentees.

2.4.2. Coding of variables

Our main dependent variable is *Total days*; it counts the total number of days of sickness absence between the draw and the date of data extraction (at the end of the experiment). *Days (within spell)* is instead a measure of days within the current sick leave spell. We chose *Total days* as our main variable of interest since there may be longer-term effects on absences beyond the current spell. For example, the DM might induce the absentee to return to work before having fully recovered, but this could worsen her/his condition and so end up to generate a new sickness spell. Alternatively, returning to work earlier may reduce the likelihood of sick leave in the future, for instance by preventing the development of habits.

Furthermore, dialogue meetings could potentially affect the sick leave grade (i.e. the percentage of work hours on sick leave; recall that sick leave is not always full time), by reducing a full-time sick leave to a graded sick leave. Since *Total days* would not capture this, we rely on the variable *Graded days*, which is a measure of full-day equivalent days on sick leave between the draw and the date of data extraction. For instance, if a person is on half-time sick leave for a period, this measure counts each calendar day as a half day during this period. We label people summoned to meetings *Summoned (DM)*, and people not summoned to meetings as *Control*.

Table 2
Balance tests. The dependent variable is being summoned to a dialogue meeting.

	(1) Summon	(2) Summon	(3) Summon	(4) Summon	(5) Summon	(6) Summon	(7) Summon	(8) Summon	(9) Summon
Female	0.018*								0.015
	(0.011)								(0.011)
Birth year		0.00034							0.00043
		(0.00044)							(0.00044)
Days before			0.000040						0.000032
			(0.000048)						(0.000049)
Grade				−0.00032					−0.00036*
				(0.00020)					(0.00020)
nr employees					−0.0000011				−0.0000019
					(0.0000032)				(0.0000033)
Symptoms						−0.0075			−0.010
						(0.010)			(0.011)
CW predicted meeting important							0.023**		0.022*
							(0.012)		(0.012)
CW predicted long spell								0.034**	0.034**
								(0.017)	(0.017)
Share of individuals treated	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
Controls	Block	Block	Block	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects (a combination of week of draw and office). Robust SE in parentheses. *P*-values are $\leq 0.01^{***}$, $\leq 0.05^{**}$, and $\leq 0.1^{*}$. The F-value from an F-test of the significance of all variables in Column 9 is 6.23 ($p = 0.0126$) and of all the baseline covariates the F-value is 2.05 ($p = 0.1518$).

We present descriptive statistics for the analysis sample in Table 1, and we see that all three outcome measures are similar for both groups.

Our baseline background variables are gender, birth year, days before (the absence history measured as the total number of days on sickness absence since 2015 up until the date of the draw), grade of sickness absence (at the time of the draw), and number of employees at the workplace (at the time of the draw). These background variables are used as controls and for heterogeneity analysis. In addition, for all individuals in the data, we have the diagnoses classified according to the second edition of the International Classification of Primary Care (ICPC-2). This classification uses a letter for each broad type of diagnosis, such as a bodily part (e.g. F = Eye) or type (e.g. P = Psychological), and a number within each letter. There is a distinction between symptoms/complaints on the one side and established diagnoses on the other within each letter. Table 2 presents the share of individuals classified as having symptoms (as opposed to established diagnoses) in this classification system.

We present the share of people for whom the caseworkers predicted the meeting to be “important” or “very important” and the predicted share having a long spell (over 20 weeks). We see that the baseline variables are similar across treatment and control, while there is a small difference for the caseworker predictions.

2.4.3. Balance tests

In Table 2 we present balance tests based on estimating an equation where individual characteristics predict treatment status (being summoned to a meeting or not). We do so within cells of NAV office and weeks, as treatment probabilities are fixed within these cells, but not necessarily across cells. In Columns 1–8, we test whether treatment status correlates with variables describing the absentee, the workplace, the health problem, and the caseworker. In Column 9, we include all variables simultaneously and test their joint significance.

In Column 1, we see that women are treated somewhat more often than men. In contrast, birth year, absence history (measured as the total number of days on sickness absence since 2015 up until the date of the draw) and absence grade (full time or part-time absence) are completely unrelated to treatment status, as shown in Columns 2–4. In Column 5, we consider firm size (number of employees). Firm size and illness severity[6] are also unrelated to treatment status (Columns 5–6). On the contrary, there seem to be small, but not ignorable, differences between treated and controls in therms of caseworkers’ predictions, as shown in Columns 7 and 8. Finally, in Column 9, we include all the covariates jointly and test their joint significance. The correlation between caseworkers’ predictions and treatment status is also present in Column 9. Although we find no reason for why such an imbalance should exist because these questions were answered before the draw, we will present results with and without controlling for caseworker predictions.[7] The remaining covariates seem unrelated, individually or jointly, to treatment status. We present a similar balance table (in Appendix Table A.1) for the more fine-grained definitions of the treatment.

[6] We use a diagnosis-based measure of illness severity, distinguishing between “symptoms/complaints” or an actual diagnosis.

[7] Whether or not we include caseworkers’ predictions does not affect any of the results presented in this paper. In addition, we conduct an analysis using a double robust LASSO procedure to select control variables. This procedure chooses variables that are correlated with both the treatments and the outcomes. Also these results produce similar estimates.

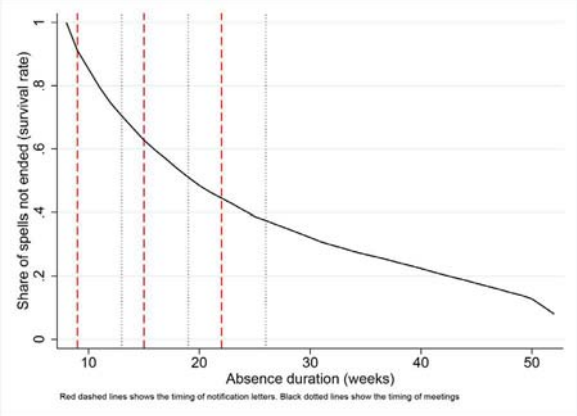


Fig. 2. Share of absence spells not ended (Survival rate) by week over the absence spell.

3. Results

3.1. The effect on meeting attendance

In our experiment the treatment consists in summoning absentees to a dialogue meeting, and the control in receiving a letter that states the absentee will not be summoned to a meeting (but that she has the right to ask for a meeting if she likes). The first question to ask is thus: did our intervention increase the probability of having a DM? In general, most absentees will never attend dialogue meetings because, as time passes by, health conditions naturally improve, and people return to work before they receive an invitation or after receiving an invitation but before the meeting date. This is also the case in our data, and we see that most absentees return to work before attending a meeting. Fig. 2 displays the share of sick-leave spells not ended (the survival rate) by week, conditional on eight weeks of absence. In week 13, when the earliest meetings are to take place, around 25 percent of the absentees have returned to work. In week 26, when the latest meetings are to take place, as many as 60 percent of the absentees have returned to work (or at least ended the sick leave spell).

To test whether absentees in the treatment group attended dialogue meetings more often than in the control groups, we estimate by OLS the following equation:

$$Y_i = \beta Summoned_i + \theta Block_{ow(i)} + \lambda X_i + \varepsilon_i \tag{1}$$

Here the outcome Y_i is $Had\ Meeting_i$ which is an indicator variable for whether or not the absentee i, summoned by office o in week w, participated in a dialogue meeting. $Summoned_i$ is a dummy equal to one for individuals summed to a meeting (the treatment group). $Block_{ow(i)}$ are fixed effects for all unique combinations of NAV office and week. Recall that within each combination of office and week, the treatment $Summoned_i$ is assigned at random. As such, in order to identify β, it is not necessary to control for the length of each spell (or for any other variable), even though the probability of a meeting occurring is increasing in the length of the spell.

In Table 3, Column 1, we see that absentees in the treatment group on average had a 14 percentage point higher propensity to attend a DM than those in the control groups. Overall, 11 percent of those in the control group attended a DM, meaning that the treatment groups' participation rate was 2.3 times higher than in the control group.

In Columns 2–4, we split the sample in three, based on the week in which the letters were sent (either 9, 15 or 22); in each sub-sample, we estimate Eq. (1) separately. In Column 2, we compare control and treatment groups who received letters in week 9: those summoned to DM (to be held in week 13) had a 24 percentage point higher propensity to attend dialogue meetings than absentees receiving the other type of letter (control group). In Column 3, we see that among those who were sent a letter in week 15, those summoned to a meeting to be held in week 19 and 26 had respectively a 16 and 13 percentage point higher propensity to attend a dialogue meeting than the control group. Among those sent letters in week 22 (for a meeting to be hold in week 26), the treatment group had a 5 percentage point higher propensity to attend DMsn (Column 4). Finally, we revert to the full sample in Column 5, where we estimate an equation that includes three different treatment dummies variables, discriminating inviduals summoned to a meeting (the treatment group) based on the week in which the meeting was supposed to take place (either 13, 19 or 26). This specification shows that the increase in propensity to attend a DM relative to the control group is higher for those summoned to earlier meetings.

To further inspect these patterns we display DM participation by week in Fig. 3. Panel (a) shows the fraction who attended a dialogue meeting each week for each of the treated groups (the control group is lumped together). Despite some within-group variation, each group has a spike in participation exactly in the week when they were summoned to a DM. Instead, Panel (b) shows the cumulative share that attended a dialogue meeting by week. Note that the participation rate is higher among people summoned to earlier meetings; furthermore, around 11 percent of the control groups attended a DM, and these meetings mostly took place between weeks 18 and 35.

Table 3
Treatment effects on probability of attending a meeting.

	(1) Had meeting	(2) Had meeting	(3) Had meeting	(4) Had meeting	(5) Had meeting
Summoned	0.14***				
	(0.0078)				
Meet week 13		0.24***			0.22***
		(0.019)			(0.015)
Meet week 19			0.16***		0.15***
			(0.015)		(0.014)
Meet week 26			0.13***	0.050***	0.098***
			(0.014)	(0.016)	(0.0094)
Mean dep. var in C.	0.11	0.10	0.10	0.15	0.11
No. of observations	10,235	2351	5122	2762	10,235
R-squared	0.11	0.31	0.17	0.25	0.12
Letter sent in week	9,15,22	9	15	22	9,15,22

Notes: All regressions control for block fixed effects. *P*-values are $\leq 0.01^{***}$, $\leq 0.05^{**}$, and $\leq 0.1^{*}$. The outcome variable, *Had meeting*, is an indicator variable for whether or not the absentee participated in a dialogue meeting. *Summoned* is an indicator variable for whether the absentee is treated or not. *Meet week w* is an indicator variable for whether the absentee was summoned to meet in week w. C is short for Control group. Robust SE in parentheses.

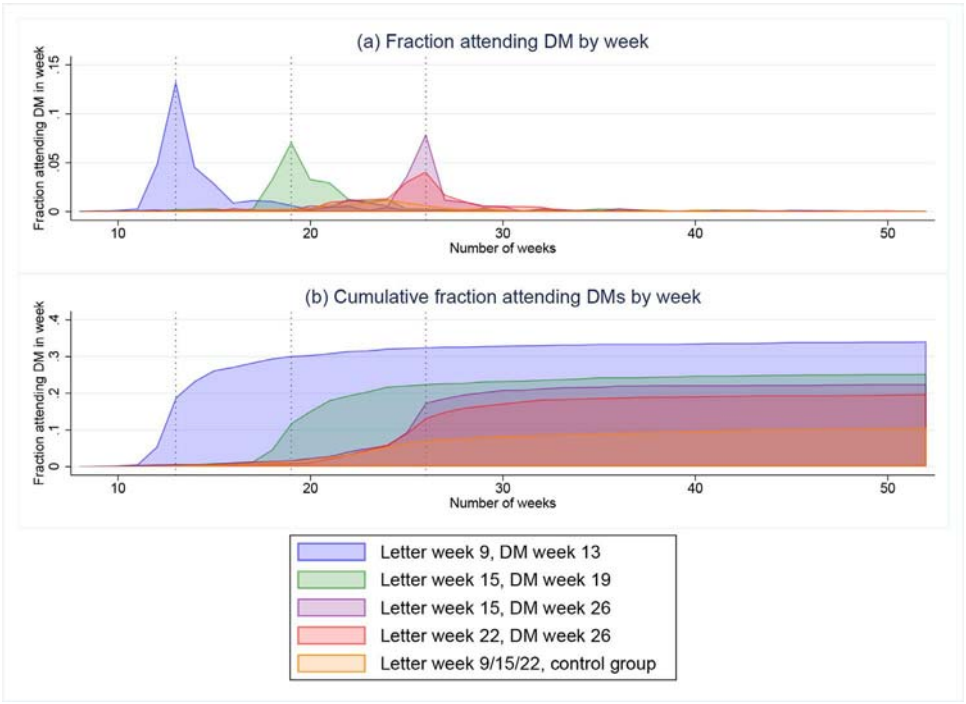


Fig. 3. Share and cumulative share participating in dialogue meetings per week over the absence spell.

3.2. The effect on absence duration

To estimate the effect of our intervention on absence duration, we estimate (1), using *Total days* as the outcome variable, i.e., the total number of absence days from the random draw and until the data extraction.[8] Results are reported in Table 4, which has the same structure as the previous table. In Column 1, we compare treatment group to control group in the full sample without exploiting information on when the letters were sent (the dummy $Summoned_{i,ow}$ is equal to one for all individuals summoned to a meeting irrespective of when the letters were sent); we find that absentees summoned to meetings have 3.1 fewer absence days after the draw, a reduction of approximately 1.8 percent. The difference is not statistically significant. Using an equivalence testing approach of two

[8] Using the outcome measures *Days* and *Graded days* instead, gives equivalent results, see Table A.3 in the Appendix. The results are also very similar if we add control variables, see Appendix Table A.4.

Table 4
Treatment effects on sickness absence.

	(1) Total days	(2) Total days	(3) Total days	(4) Total days	(5) Total days
Summoned	−3.11				
	(2.30)				
Meet week 13		−2.26			−6.34*
		(5.47)			(3.80)
Meet week 19			2.84		3.76
			(4.27)		(3.79)
Meet week 26			−6.88*	−7.67	−4.59*
			(3.97)	(4.84)	(2.72)
Mean dep. var in C.	164.48	160.49	164.44	167.97	164.48
No. of observations	10,235	2351	5122	2762	10,235
R-squared	0.10	0.22	0.16	0.25	0.10
Letter sent in week	9,15,22	9	15	22	9,15,22

Notes: All regressions control for block fixed effects. *P*-values are ≤ 0.01***, ≤ 0.05**, and ≤ 0.1*. The outcome variable, *Total days* measures the total number of days of sickness absence between the draw and the date of data extraction. *Summoned* is an indicator variable for whether the absentee is treated or not. *Meet week w* is an indicator variable for whether the absentee was summoned to meet in week w. C is short for Control group. Robust SE in parentheses.

one-sided *t*-tests (TOST) and a 5 percent significance level, we can reject reductions larger than 6.9 total days and increases larger than 0.7 total days.

In Columns 2–4, we discriminate between individuals who received their letter at different times, exactly as we did in Table 3. In Column 2, we include only absentees who were sent letters in week 9; individuals summoned to a DM seem to have 2.3 fewer total days of sickness absence compared to the control group, yet the difference is not statistically significant. Among those who were sent letters in week 15, those summoned to a meeting in week 19 on average had a longer sickness absence than the control group, while those summoned to a meeting in week 26 on average had a shorter sickness absence length (Column 3). In Column 4 we see that among absentees who were sent a letter in week 22, those in the treatment group returned to work 7.7 days before the control group. Finally, in Column 5, we include the full sample, but discriminate between individuals who received letters at different points. We find that both those summoned to early and late meetings tend to have shorter spells than the control group, but the difference is significant at the 10-percent level only. On the contrary, workers summoned to meetings to be held in week 19 have longer spells than the control group (not statistically significant at any conventional level). We find this conspicuous pattern hard to explain. It is, however, worth noticing that the confidence intervals are large and that none of the differences between coefficients are statistically different from zero at the 5 percent level. Part of the motivation behind the experiment was to investigate whether early DMs would reduce absence more than late meetings. We find no support for this hypothesis.

We also provide a graphical representation of the absence patterns in the various treatment arms in terms of survival curves (a) and hazard rates (b) displayed in Fig. 4. The survival curve describes the fraction of all spells active in week eight that is still ongoing, whereas the hazard rate is the fraction of spells exiting in a given week, conditional on lasting up to that week. If DMs reduce total absence, we should expect the survival curve to be lower in the treatment groups than in the control groups. Likewise, we should expect that the weekly hazard rates (the share of absence spells that end or workers returning to work) to be higher in the treatment group than in the control group. Starting in Panel (a), we jointly plot the survival rates for the control and treatment groups. There are at least two things worth noticing in this panel. First, the survival curves fall substantially with time such that at week 22, just around 45 percent of the spells are still active. This is also visible in Panel (b), where we draw the weekly hazard rates for the same groups. Every week, between 4 and 8 percent of active spells end. Second, the treatment and control groups are very similar in terms of absence behavior. In Panel (a), the two survival rates are hard to distinguish from each other.

In Appendix Fig. A.1 we also present a graph where we differentiate between the different treatment arms; there we see that there is no consistent evidence of any treatment effect. In Appendix Sections A.3–A.6 we provide further robustness tests and duration analyses.

3.3. Threat effects and reversed threat effects

In Table 5 we report results on the threat and the reversed threat effects. The dependent variable in these regressions is a dummy equal to one if the absentee returned to work before a specific week (week 13 in Column 1, week 19 in Columns 2–3 and week 22 in Columns 4). The experimental design allows us to test whether receiving a letter summoning to DM increased the probability of returning to work before the meeting, by using as a comparison group all individuals who did not receive any letter yet until then.

In Column 1, the outcome is return to work prior to week 13. Keep in mind that two out of seven treatment arms were sent a letter in week 9, either summoning to a DM in week 13 or not. The remaining five arms have not heard anything from the social insurance administration by week 13, which is business as usual at this stage. We can thus use these five groups for comparison and estimate the threat effect and the reversed threat effect in the same regression. Receiving a letter summoning to a meeting leads to a slightly higher, not statistically significant, probability of returning before the meeting is supposed to take place (week 13) compared to not having been sent any letter. The coefficient is 0.0066, that corresponds to less than a 2.3 percent increase in the return-to-work

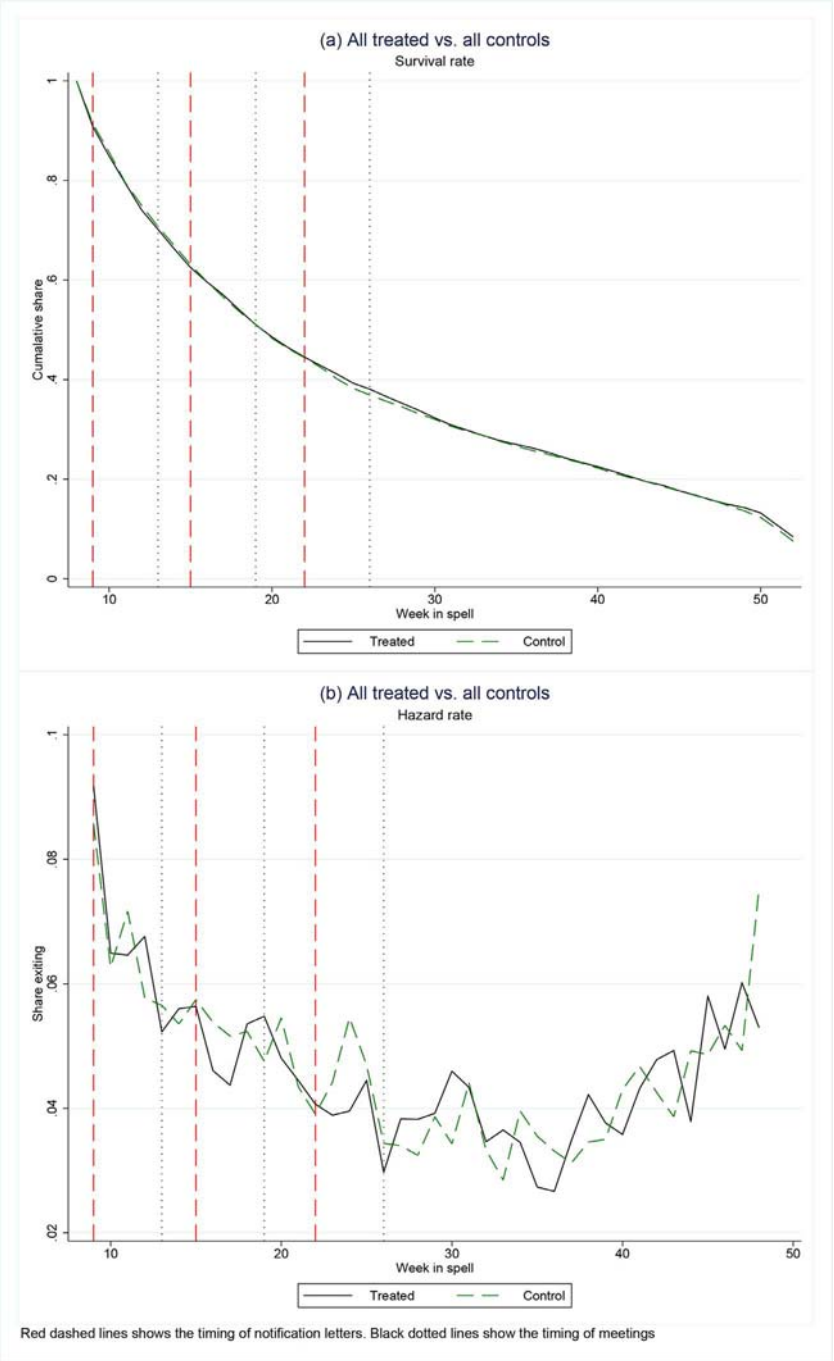


Fig. 4. Survival and hazard rates for treatment and control.

probability. Using the same equivalence test used in the previous subsection, we can reject that the increase in the likelihood of returning to work is equal or larger than 3.1 percentage points. Individuals who received a letter without being summoned to a DM have a slightly lower return-to-work probability before week 13, yet not close to being statistically significant at any conventional level. We see the same qualitative pattern in Column 2, where we only include individuals who received a letter later than week 14 and where we use as outcome a dummy for returning to work before week 19.

In Columns 3 and 4 we use as outcome a dummy for returning to work before week 19 and 22 respectively. In both cases we include all absentees in the regression, so that only individuals who received letters in week 22 are included in the reference category. Again, none of the coefficients are statistically significant. We thus conclude that threat effects are modest at best, and most likely

Table 5
The effects of receiving letters on returning to work.

	(1) Return before Week 13	(2) Return before Week 19	(3) Return before Week 19	(4) Return before Week 22
Threat effects				
DM 13, letter 9	0.0066		0.0034	−0.0078
	(0.015)		(0.018)	(0.018)
DM 19, letter 15		0.0083	0.0067	−0.0020
		(0.018)	(0.018)	(0.018)
DM 26, letter 15		−0.0027	−0.0043	0.0072
		(0.017)	(0.017)	(0.017)
Reversed threat effects				
Letter 9	−0.0091		0.030	0.023
	(0.015)		(0.018)	(0.018)
Letter 15		−0.0021	−0.0049	−0.00048
		(0.014)	(0.014)	(0.014)
Mean in excl. group	0.30	0.49	0.49	0.55
No. of observations	10,235	7884	10,235	10,235
R-squared	0.08	0.10	0.08	0.07
Sample	All	Letter after week 14	All	All

Notes: All regressions control for block fixed effects. *P*-values are $\leq 0.01^{***}$, $\leq 0.05^{**}$, and $\leq 0.1^{*}$. The outcome variables, *Return before Week w* are indicator variables for whether the absentee returns to work before week w of the sickness spell. Robust SE in parentheses.

of limited importance. In addition, we find no evidence for reversed threat effects. Our coefficients are not directly comparable to the estimates in Markussen et al. (2017) as they model exit from sick-leave and its dependence on (a predicted) risk of receiving a letter. In Appendix Section A.8 we provide some back-of-the-envelope calculations and discuss potential comparisons of both the threat effects and the total effects.

4. Pre-specified exploratory analysis of heterogenous treatment effects

The effect of a treatment may differ substantially between treated individuals. It is important to investigate such heterogeneity, as this makes it possible to understand for whom a treatment works best. This information would allow policy makers to target the interventions and thus increase welfare.

As specified in the pre-analysis plan, we first search for treatment effect heterogeneity by interacting the main treatment dummy *Summoned* with individual covariates. The variables we use for interactions are gender, birth year, the number of sickness days between 2015 and the date of the draft, a dummy for graded sickness absence, and the number of employees at the workplace. For ease of interpretation, the continuous variables are standardized, with mean zero and standard deviation one.

Table 6 displays the results using total days of sick leave as outcome; we test for treatment effect heterogeneity in a stepwise fashion (Columns 1–5). None of the treatment effect interactions are statistically significant. In Column 6 we include all the interactions jointly and the results are very similar. For completeness, we present additional pre-specified heterogeneity results in the Appendix Section A.5. These findings are in line with Table 6, and point to, at best, very limited treatment effect heterogeneity.

Treatment effect heterogeneity may also exist along dimensions unobserved to the econometrician, such as motivation, personal traits or health. Caseworkers may observe such factors, at least partially. In our study, caseworkers were asked (before the random draw) to predict the duration of the sickness spell for each absentee, and to assess the absentee-specific effectiveness of a DM. Results in Table 7 show that caseworkers' predictions on absence duration are strongly correlated with actual duration. In addition, the absentees for which caseworkers believe DMs are important tend to be absent longer than others. This evidence suggests that caseworkers observe important factors that are unobservable to the econometrician.

Using models with interactions, we do not find evidence that DMs are differentially effective on individuals who are predicted by the caseworkers to stay sick longer (Columns 1–4 of Table 8). More interestingly, we do find that DMs are more effective on the individuals for whom the casework thinks that this intervention might be useful: for this group the effect is equal to 9,5 days (Column 5 of Table 8). However, given the large number of secondary hypotheses tested, we urge for caution in interpreting any one weakly statistically significant effect as evidence of true heterogeneity.[9] Overall, we find very limited evidence of heterogeneous treatment effects[10].

[9] One way to judge the plausibility of the coefficients that are statistically significant is to consider the whole set of conducted tests. In Appendix A.7 we plot many of our tests. We exclude all triple interactions, such as the instances where e.g. gender differences for specific heterogeneities are tested and we also only show the results for the main treatment variable and the main outcome variable. Even with these restrictions, we see that we have 47 tests, out of which only 5 are statistically significantly negative at the 5 percent level. None of the effects are statistically significant if we were to adjust the *p*-values for the number of tests using for instance the false discovery rate method developed by Benjamini and Hochberg (1995) (which we pre-specified that we would use in case any of our three main hypotheses were to be statistically significant). If we do the same kind of analysis with a placebo outcome asking how long a person has been on sick leave in the years before the treatment we find three negative effects (see Fig. A.8). Hence, the evidence for any real heterogeneity in the treatment effects using standard methods is weak.

Table 6
Heterogenous treatment effects on sickness absence.

	(1) Total days	(2) Total days	(3) Total days	(4) Total days	(5) Total days	(6) Total days
Summoned (DM)	−1.68	−2.90	−3.29	−0.81	−3.13	0.23
	(3.76)	(2.29)	(2.29)	(3.43)	(2.30)	(4.22)
Female	2.79					5.23
	(3.35)					(3.38)
Female*DM	−2.33					−1.85
	(4.70)					(4.75)
Birth year		−12.6***				−11.5***
		(1.62)				(1.64)
Birth year*DM		1.25				0.92
		(2.26)				(2.27)
Days before			13.4***			12.2***
			(1.58)			(1.59)
Days before*DM			−1.13			−0.67
			(2.17)			(2.19)
Graded				−13.2***		−14.5***
				(3.25)		(3.27)
Graded*DM				−3.62		−3.52
				(4.56)		(4.59)
nr employees					−1.87	−1.99
					(1.64)	(1.65)
nr employees*DM					−1.23	−1.72
					(2.36)	(2.38)
Mean dep. var in C.	164.48	164.48	164.48	164.48	164.48	164.48
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.10	0.11	0.11	0.10	0.10	0.12
Controls	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. *P*-values are ≤ 0.01***, ≤ 0.05**, and ≤ 0.1*. The outcome variable, *Total days* measures the total number of days of sickness absence between the draw and the date of data extraction. *Days before* measures the total number of days on sickness absence since 2015 up until the date of the draw. *Grade* is the grade of sickness absence at the time of the draw. *nr employees* is the number of employees at the absentee's workplace (at the time of the draw). Robust SE in parentheses.

Table 7
Association between caseworker predictions and sickness absence.

	(1) Total days	(2) Total days	(3) Over 20 days	(4) Total days
CW predicts the absentee to be away... (excluded category is do not know)				
Less than 4 weeks	−42.3***			
	(5.72)			
Between 4–11 weeks	−12.1***			
	(3.61)			
Between 12–19 weeks	−0.48			
	(3.57)			
More than 20 weeks	24.8***			
	(4.42)			
Other measures				
CW predicted long spell		30.7***	0.13***	
		(3.61)	(0.016)	
CW predicted meeting important				16.0***
				(2.55)
Mean in excl. group	163.55	159.03	0.47	154.07
No. of observations	10,235	10,235	10,235	10,235
R-squared	0.18	0.17	0.16	0.17
Controls	Baseline	Baseline	Baseline	Baseline

Notes: Robust SE in parentheses. *P*-values are ≤ 0.01***, ≤ 0.05**, and ≤ 0.1*. The outcome variable, *Total days* measures the total number of days of sickness absence between the draw and the date of data extraction. *CW predicted meeting important* is the share of absentees for which the caseworkers (CW) predicted a meeting would be "important" or "very important". *CW predicted long spell* is the share of absentees for which the caseworker predicted having a long spell (over 20 weeks)

Table 8
Treatment effects and caseworker predictions. Dependent variable is total days of absence.

	(1) Total days	(2) Total days	(3) Total days	(4) Total days	(5) Total days	(6) Total days	(7) Total days
Summoned (DM)	−2.83	−4.18*	−7.00**	0.10	1.37	2.47	−0.13
	(4.93)	(2.33)	(2.94)	(4.13)	(3.05)	(5.52)	(3.86)
CW predicts the absentee to be away... (excluded category is do not know)							
Less than 4 weeks	−38.1***						
	(7.80)						
Less than 4 weeks*DM	−8.82						
	(10.4)						
Between 4–11 weeks	−9.04*						
	(4.80)						
Between 4–11 weeks*DM	−6.17						
	(6.30)						
Between 12–19 weeks	−2.25						
	(4.71)						
Between 12–19 weeks*DM	3.48						
	(6.22)						
More than 20 weeks	22.5***						
	(6.21)						
More than 20 weeks*DM	4.48						
	(8.23)						
CW predicted long spell		27.7***	18.2***	40.4***			
		(5.22)	(6.79)	(9.18)			
CW predicted meeting important					20.8***	17.9***	21.1***
					(3.36)	(6.09)	(4.28)
Long spell*DM		5.87	9.58	−4.75			
		(6.98)	(9.16)	(11.9)			
Meeting important*DM					−9.54**	−5.29	−11.4**
					(4.38)	(7.78)	(5.60)
Mean dep. var in C.	164.75	160.27	161.17	158.82	152.88	153.42	153.90
No. of observations	10,235	10,235	6403	3832	10,235	3832	6403
R-squared	0.18	0.17	0.22	0.24	0.17	0.24	0.22
Controls	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
Sample	All	All	Women	Men	All	Women	Men

Notes: Robust SE in parentheses. *P*-values are ≤ 0.01***, ≤ 0.05**, and ≤ 0.1*. See Table 7 for variable definitions.

We also use machine learning techniques, in particular random forests (Athey et al., 2019; Wager and Athey, 2018), to automate the search for heterogeneous treatment effects and to account for non-trivial interactions between 61 covariates.[11] We train our prediction model using only half of the sample (training sample), drawn at random; heterogeneous effects are then estimated and tested on the other half (test sample). Using this method, we are unable to detect any true treatment heterogeneity.

5. Concluding discussion

We teamed up with the Norwegian Labour and Welfare Administration to conduct a large-scale, pre-registered, randomized field experiment to evaluate aspects of the Norwegian DMs policy and to explore heterogeneous treatment effects. Our results can be summarized as follows: First, our experiment did increase the likelihood of participating in a meeting by, on average, 14 percentage points. Second, individuals summoned to a meeting had on average 3.1 days shorter absence spells than those not invited to a meeting. However, the difference is not statistically significant from zero (*p*-value: 0.176), and we can reject reductions larger than 6.9 days and increases larger than 0.7 days with a *p*-value of 0.05. Third, we find no evidence for a notification effect (or "threat effect") from being summoned to a meeting (before the meeting takes place). Fourth, our results do not support that early DMs reduce absence more than late meetings. Fifth, we extensively test for treatment effect heterogeneity using sample splits, interactions, and machine learning techniques. To our best judgment, we do not find any convincing support for such heterogeneity.

[10] Appendix Section A.9 presents additional not pre-registered heterogeneity analysis that confirm this conclusion.

[11] We pre-register the parameters and the specifications used. See the Appendix Section A.7 for more details and for the results. The covariates are gender, birth year, number of children, a divorced/separated dummy, a never married dummy, EU national dummy, Norwegian nationality dummy, days before (the absence history measured as the total number of days on sickness absence since 2015 up until the date of the draw), number of employees at work place (at the time of the draw), grade of sickness absence (at the time of the draw), a set of diagnoses dummies based on ICPC-2, a dummy for having symptoms as opposed to established diagnoses, two case workers predictions on whether the meeting was going to be important and the spell long, a dummy for the presence of the doctor in the meeting, a dummy for the presence of the case worker in the meeting, case worker gender, a dummy for case worker having high experience, a dummy for case worker being a professional social worker, a dummy for case worker having a university degree, a dummy for whether the meeting took place at the NAV office, and a set of regional office dummies.

Although we can not reject the effectiveness of holding a dialogue meeting on work resumption, our findings cast serious doubts on the usefulness of a generalized policy of summoning all absentees to a meeting, which is the current goal in Norway. In particular, our results severely undermine the potential for simply nudging absentees back to work using a letter invitation, thus ruling out a very cost-effective policy option.

There are several possibilities for why we were not able to detect meaningful effects of DM in our experiment. A first possibility is that the experiment was not powered enough to detect effects of meaningful importance. Throughout the paper we stress what effect sizes we can reject and according to power calculation conducted prior to the experiment, we deemed such effect sizes to be small. We can also reject small effects of merely sending out letters, which is important as NAV was under the impression that such an easy and cheap policy would be highly beneficial; this was also the consensus in the academic literature. That being said, we realize that our study is severely underpowered to detect effects of meetings actually held. The reason is that most people return to work before the meeting, so the first stage effect of sending letters on meeting attendance is low; as such all estimates have to be scaled by a large number in order to use the assigned treatments as instruments. Testing the effect of meeting attendance was not our initial intention and we admit that our results do not speak strongly to this question. We hope that future work can estimate this by, for example, using experiments where some meetings are cancelled at random.

A second possibility is that heterogeneous effects are at play. We think that this is not very likely in our case as we conduct an extensive search for heterogeneous treatment effects without finding any. Furthermore, as the average effect is also close to zero, the presence of heterogeneous effects would imply that any positive effects for some individuals would need to be exactly offset by negative effects on others, which seems odd at best.

A final possibility is of course that the DMs have no effect on sickness absence, despite previous findings. Markussen et al. (2017) report substantial effects of DMs, split approximately equally between a pre-meeting notification and a post-meeting participation effect. Our study has the advantage of being from a pre-registered randomized experiment, while the results in Markussen et al. (2017) are from analyses of observational data, which requires more assumptions. The advantage of the Markussen et al. (2017) is that they have more data and higher statistical precision. We cannot draw definitive conclusions from any one single study and we hope that future studies will be able to provide more conclusive evidence on the total effects of DMs.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

A1. More balance tests and figures

Table A.1 shows the balance across all seven detailed treatment variables.

A2. A split sample approach

Testing for treatment effect heterogeneity is more difficult than identifying average treatment effects. While it is important to establish which groups benefit or are hurt most by a policy, there are pitfalls of naively splitting the data to test for effects across subgroups. One issue is that multiple hypotheses are tested which is likely to lead to false positives if not corrected for. Without a pre-specified analysis plan it is impossible to know how many non-reported tests were conducted by the researchers (Olken, 2015). Specifying all possible tests is, however, difficult, especially in a setting where there is an inherent uncertainty with respect to treatment heterogeneity.

Hence, there is a tradeoff between pre-specifying hypotheses and learning about heterogeneity from the data. A solution to the problem is replication of research findings (Coffman and Niederle, 2015; Coffman et al., 2017) but often experiments are costly and difficult to replicate. Recently, economists have started to borrow methods from the machine learning literature to solve this issue. Machine learning methods are primarily aimed at predicting an outcome variable and use cross-validation to compare model predictions to actual outcomes in test samples not used to estimate the model (see e.g. Mullainathan and Spiess, 2017; Varian, 2014 for recent reviews of this literature from the perspective of economics).

Inspired by cross-validation, Athey and Imbens (2016) propose ways to estimate heterogeneous treatment effects relying on machine learning techniques and sample splitting: one sample is used to identify different subgroups based on all the available covariates, while the second sample is used to estimate treatment effects. Given that the second sample has not been used to select the subgroups, the subgroup structure is exogenous and standard inference can be used. These methods, which Athey and Imbens (2016) and Wager and Athey (2018) label “honest”, reduce the problem of multiple hypothesis testing by building replication into the process. We describe how we use machine learning techniques as a complement in Section A.7 but the split sample approach can be applied directly in any data collection as long as data is received in steps (Anderson and Magruder, 2017; Fafchamps and Labonne, 2017).

Here follows a more detailed account of our split sample procedure. On October 30 (calendar week 44) 2017 we received the treatment status (the outcome of the randomization) of 7619 individuals. In creating the early sample, we only wanted to include those who were drafted early enough to have had the possibility of having a meeting within that time (calendar week 44 in 2017).

Table A.1
Balance tests for all seven treatments.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	*T*-test																		
	DM 13 letter 9	Letter 9	DM 19 letter 15	DM 26 letter 15	Letter 15	DM 26 letter 22	Letter 22	Difference																		
Variable	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	(1)–(2)	(1)–(3)	(1)–(4)	(1)–(5)	(1)–(6)	(1)–(7)	(2)–(3)	(2)–(4)	(2)–(5)	(2)–(6)	(2)–(7)	(3)–(4)	(3)–(5)	(3)–(6)	(3)–(7)	(4)–(5)	(4)–(6)	(4)–(7)	(5)–(6)	(5)–(7)	(6)–(7)
Total days	147.273	147.624	154.365	148.081	151.004	148.534	153.558	−0.351	−7.092	−0.808	−3.731	−1.261	−6.285	−6.741	−0.458	−3.380	−0.910	−5.935	6.284	3.361	5.831	0.807	−2.923	−0.453		2.470		−5.024*
	(3.146)	(3.173)	(3.186)	(2.823)	(2.136)	(2.889)	(2.919)																		−5.477		−2.554	
Days	119.416	116.353	121.397	117.509	119.001	118.677	121.886	3.063	−1.981	1.907	0.415	0.739	−2.470	−5.044	−1.156*	−2.648*	−2.324	−5.533	3.888	2.395	2.719	−0.490	−1.492	−1.168		0.324		−3.209
	(3.053)	(3.044)	(3.142)	(2.787)	(2.095)	(2.906)	(2.841)																		−4.377		−2.885	
Graded	109.177	107.406	113.563	110.837	112.963	111.975	112.860	1.771	−4.386	−1.660	−3.786	−2.798	−3.683	−6.157	−3.431	−5.557*	−4.569	−5.454	2.726	0.600	1.588	0.703	−2.126	−1.138		0.988	0.103	−0.885
	(2.688)	(2.701)	(2.766)	(2.451)	(1.870)	(2.505)	(2.486)																		−2.023			
Female	0.637	0.638	0.643	0.603	0.615	0.652	0.606	−0.001	−0.006	0.034	0.021**	−0.016	0.031	−0.005	0.035	0.022	−0.015	0.032	0.040	0.027	−0.009	0.037	−0.013				0.010	0.046**
	(0.014)	(0.014)	(0.014)	(0.013)	(0.010)	(0.013)	(0.013)																	−0.049***	−0.003	−0.037***		
Birth year								0.353	0.128	−0.108	0.432	0.038	−0.139	−0.224	−0.461	0.080	−0.315	−0.492	−0.236	0.304	−0.091	−0.268	0.540	0.146		−0.395		−0.177
	1974.156	1973.803	1974.027	1974.263	1973.723	1974.118	1974.295																		−0.031		−0.571	
	(0.351)	(0.348)	(0.351)	(0.312)	(0.239)	(0.318)	(0.319)																					
Days before	144.126	141.569	139.906	136.346	136.608	140.833	138.688	2.557	4.220	7.780	7.518*	3.293	5.438	1.663	5.223*	4.961	0.736	2.881	3.559	3.297	−0.928	1.218	−0.262	−4.487		−4.225		2.146
	(3.312)	(3.259)	(3.143)	(2.771)	(2.077)	(2.916)	(2.840)																		−2.341		−2.079	
Grade	73.982	74.386	73.933	74.993	75.778	74.746	75.749	−0.404	0.049	−1.011	−1.796	−0.764	−1.767	0.453	−0.607	−1.392	−0.360	−1.363	−1.060		−0.813	−1.816	−0.785	0.246		1.032**	0.029	
	(0.761)	(0.773)	(0.791)	(0.704)	(0.524)	(0.702)	(0.712)													−1.845**					−0.757			−1.003**
nr employees	491.236	514.912	499.587	538.113	535.974	476.781	533.803	−23.677	−8.352	−46.877	−44.738	14.454	−42.567	15.325	−23.200	−21.061	38.131	−18.890	−38.526	−36.386	22.806	−34.216	2.139	61.332	4.310	59.193	2.171	−57.022
	(41.224)	(43.210)	(42.264)	(43.504)	(31.611)	(33.651)	(48.203)																					
Symptoms	0.380	0.383	0.397	0.397	0.405	0.403	0.410	−0.003	−0.016	−0.016	−0.025	−0.023	−0.030	−0.014	−0.014	−0.022	−0.020	−0.027	−0.000	−0.009	−0.007	−0.014	−0.009	−0.007		0.002		−0.007
	(0.014)	(0.014)	(0.014)	(0.013)	(0.010)	(0.013)	(0.013)																		−0.014		−0.005	
CW predicted	0.542	0.486	0.534	0.505	0.507	0.490	0.482	0.056**	0.008	0.037	0.035	0.052	0.059	−0.048	−0.019	−0.021		0.003	0.029	0.027	0.043	0.051	−0.002	0.014	0.022	0.017	0.025	0.008
meeting important	(0.014)	(0.015)	(0.015)	(0.013)	(0.010)	(0.013)	(0.014)										−0.004**											
CW predicted	0.129	0.104	0.107	0.111	0.103	0.116	0.101	0.026**	0.022	0.018	0.027*	0.013	0.029	−0.004	−0.007	0.001		0.003	−0.004	0.005	−0.009	0.007	0.008	−0.006	0.010	−0.014	0.002	0.016
long spell	(0.010)	(0.009)	(0.009)	(0.008)	(0.006)	(0.009)	(0.008)										−0.013**											
N	1183	1167	1175	1389	2556	1391	1370																					
F-test of joint significance (*F*-stat)								1.459	0.605	0.845	1.587*	0.563	1.431	0.493	1.457	1.149	1.528	0.703	0.510	1.112	0.693	0.601	0.674	1.190	0.957	1.981**	0.908	
																												2.546***
F-test, number of observations								2350	2358	2572	3739	2574	2553	2342	2556	3723	2558	2537	2564	3731	2566	2545	3945	2780	2759	3947	3926	2761

Notes: The value displayed for *t*-tests are the differences in the means across the groups. The value displayed for *F*-tests are the *F*-statistics. Fixed effects using variable Blocks are included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A.2
Baseline results with alternative block variable (caseworker and week).

	(1) Total days	(2) Total days
Summoned (DM)	−3.77	
	(2.63)	
Meet week 13		−6.44
		(4.36)
Meet week 19		3.42
		(4.40)
Meet week 26		−5.59*
		(3.11)
Mean dep. var in C.	164.48	164.48
No. of observations	10,235	10,235
R-squared	0.27	0.27
Controls	block2	block2

Notes: All regressions control for block2 fixed effects (caseworker and week). Robust SE in parentheses.

Therefore we only included individuals in the early sample who were drafted before calendar week 26 in 2017.[12] This gave an early sample consisting of 3722 individuals who were drafted before week 26 and thus already had the opportunity to have a meeting. The early sample was then randomly split into two groups: the early training sample and the early test sample (see Fig. A.2 for an overview of the procedure).[13]

On January 26th 2018, we received data on outcomes and covariates from NAV for the early training sample. This sample was used to derive and test some first hypotheses which were then registered in our pre-analysis plan submitted on February 12th 2018. The results from the early training sample were published together with the pre-analysis plan at the AEA registry. In this sample we could not reject that there was no effect of dialogue meetings but neither could we reject that the effects were medium sized. There seemed to be substantial effect heterogeneity with respect to the caseworkers' predictions of long absence whereby those that are predicted to be away for a long period have a much larger effect of being called in to a meeting. A duration analysis suggested that there was a threat effect such that individuals were more likely to end the sick leave spell in the week they got the letter and in the week before the meeting. There also seemed to be a reversed threat effect whereby individuals who received letters stating they would not be summoned to a meeting were less likely to end their sick-leave spell in the week they get the letter. This reversed threat effect was observed when running a regression of returning before week 13 on receiving a letter in week 9 not summoning to a DM. We pre-specified that we would suggest policy changes and changes to the experiment if the preliminary effects found in the early training data with respect to effect heterogeneity by caseworker predictions of long spells, threat effects and reversed threat effects, replicated in the early test sample.

On February 13th 2018 we received data on outcomes and covariates from NAV for the early test sample. All analyses described in the pre-analysis plan were now conducted on the early test sample. The results from the early training sample were not replicated using the early test sample. We found no support for any of the three pre-registered main hypothesis nor were the pre-registered heterogeneity results successfully replicated. The design of the experiment was therefore not altered (if we would have found the same results we would have tried to optimize the targeting of the treatment). In December 2018 we received all data on treatment status, outcomes and covariates from NAV for the late sample (all individuals drafted after week 26 in 2017). The late sample together with the early test sample constitutes the final analysis sample. The final analysis sample contains 10,235 individuals.

A3. Pre-registered robustness tests of the main specification

We conduct a series of robustness tests of the main specification. In Table A.2 we use an alternative blocking variable with similar results. In Table A.3 we show the baseline results for our different measures of sickness absence. That is, we run the same specification for the number of days within the sickness spell that an individual is absent, $Days_i$, and the grade adjusted total number of days, $Graded_i$. In Column 1 we see a small negative effect of being assigned to dialogue meetings (DM) on the number of days the individual is on sick leave within the spell active at the time of the draft. The reduction of around 0,5 days is not statistically significant. In Column 2 we see that the effect is larger in absolute numbers for our preferred measure, which measures the total days of sickness absence after treatment. The effect is still very small and it is not statistically significant. In Column 3 we investigate the effects on the grade-adjusted measure and we note that the effect lies between the other two measures. Dividing the treatment up by timing of the meeting shows no consistent pattern.

[12] Since it is at most is 18 weeks between the draft and the time of the meeting (in the experimental design the latest possible meeting is week 26 of the sickness spell and the draft is performed in week 8 of the spell).

[13] We choose to split the sample in two based on power calculations suggesting that we would be able to detect medium-sized effects for the summoned versus summoned to meeting using half of this sample. The minimum detectable effect was calculated to be 0.13 standard deviations with 80 percent power and 5 percent significance level for the main outcome, total days of absence.

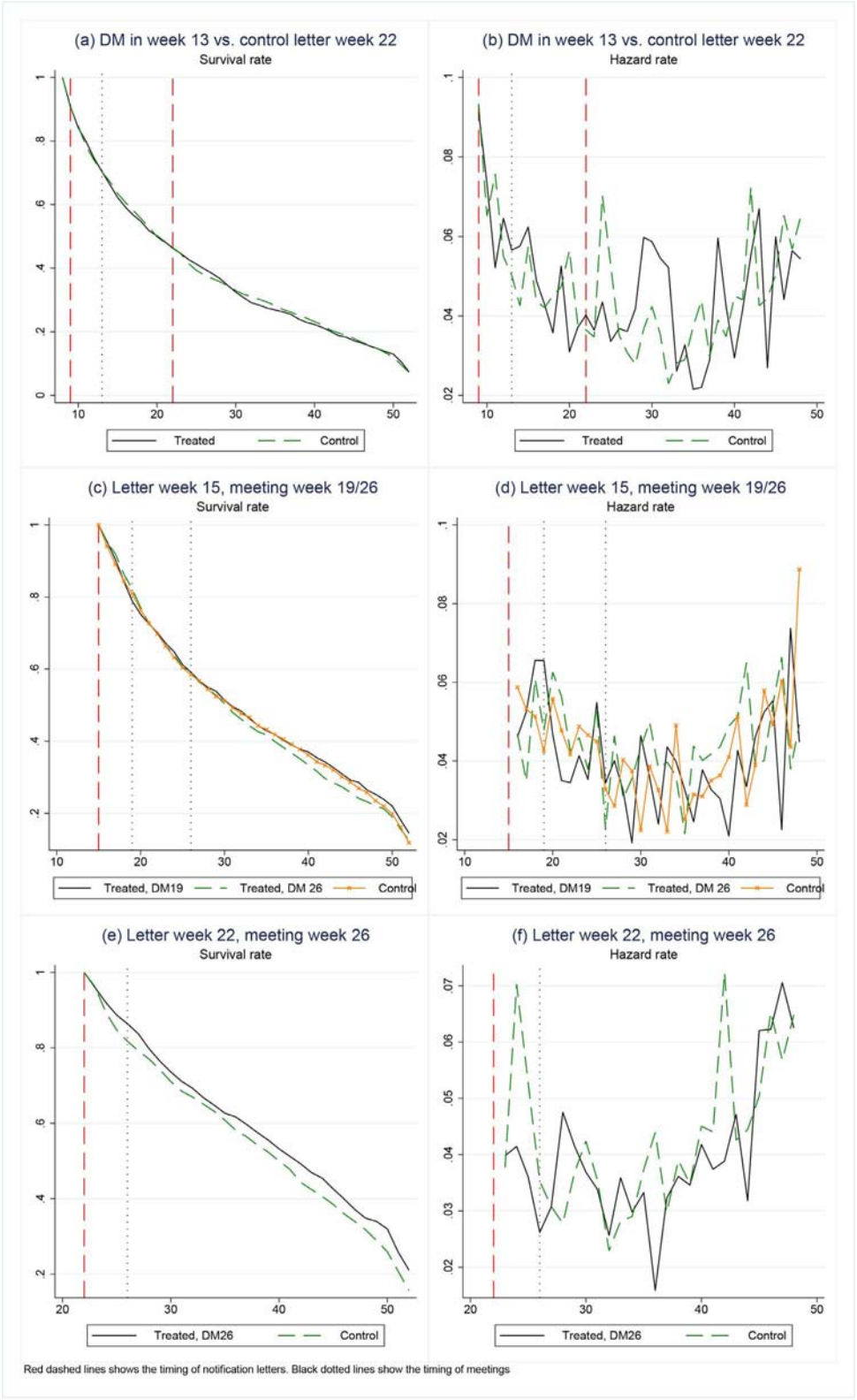


Fig. A.1. Survival and hazard rates for treatment and control.

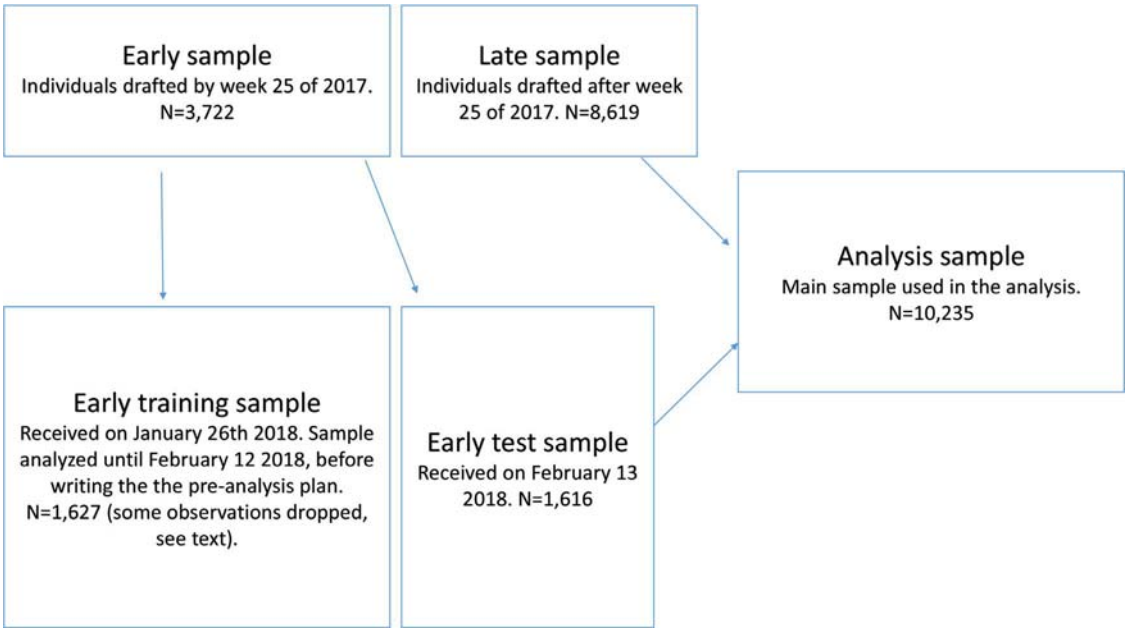


Fig. A.2. The different samples.

Table A.3
Baseline results for different measures.

	(1) Days	(2) Total days	(3) Graded days	(4) Days	(5) Total days	(6) Graded days
Summoned (DM)	−0.48	−3.11	−1.53			
	(2.23)	(2.30)	(2.22)			
Meet week 13				−1.62	−6.34*	−2.29
				(3.63)	(3.80)	(4.02)
Meet week 19				3.07	3.76	2.13
				(3.68)	(3.79)	(3.22)
Meet week 26				−1.51	−4.59*	−2.67
				(2.66)	(2.72)	(2.45)
Mean dep. var in C.	120.09	164.48	112.37	120.09	164.48	112.37
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.08	0.10	0.09	0.08	0.10	0.09
Controls	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.4
Baseline results with different control sets.

	(1) Total days	(2) Total days	(3) Total days	(4) Total days	(5) Total days	(6) Total days
Summoned (DM)	−3.11		−3.13		−3.23	
	(2.30)		(2.21)		(2.21)	
Meet week 13		−6.34*		−6.94*		−6.05
		(3.80)		(3.66)		(3.68)
Meet week 19		3.76		4.08		4.45
		(3.79)		(3.64)		(3.65)
Meet week 26		−4.59*		−4.52*		−3.90
		(2.72)		(2.61)		(2.63)
Mean dep. var in C.	164.48	164.48	164.48	164.48	164.48	164.48
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.10	0.10	0.17	0.17	0.17	0.16
Controls	Block	Block	Baseline	Baseline	Extensive	Extensive

Notes: All regressions control for block fixed effects. In addition we control for the baseline control variables in Columns 3 and 4 and the set of extensive controls in Columns 5 and 6. Robust SE in parentheses.

Table A.5
Effects when comparing to people with letters the same week. Total number of sick days as outcome.

	(1) Letter week 9	(2) Lw 15, Meet 19	(3) Lw 15, Meet 26	(4) Letter week 15	(5) Letter week 22
Meet week 13	−2.26				
	(5.47)				
Meet week 19		2.53		2.84	
		(4.42)		(4.27)	
Meet week 26			−6.72*	−6.88*	−7.67
			(4.07)	(3.97)	(4.84)
Mean dep. var in C.	160.49	164.44	164.44	164.44	167.97
No. of observations	2351	3733	3947	5122	2762
R-squared	0.22	0.21	0.19	0.16	0.25
Controls	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.6
Baseline results (cluster at the caseworker level).

	(1) Days	(2) Total days	(3) Graded days	(4) Days	(5) Total days	(6) Graded days
Summoned (DM)	−0.48	−3.11	−1.53			
	(2.79)	(3.17)	(2.90)			
Meet week 13				−1.62	−6.34	−2.29
				(4.89)	(5.20)	(5.11)
Meet week 19				3.07	3.76	2.13
				(4.01)	(4.47)	(3.92)
Meet week 26				−1.51	−4.59	−2.67
				(2.65)	(2.82)	(2.53)
Mean dep. var in C.	120.09	164.48	112.37	120.09	164.48	112.37
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.08	0.10	0.09	0.08	0.10	0.09
Controls	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects, and the variables used in the heterogeneity analysis. Robust SE in parentheses.

Table A.7
Baseline results (adding Xs and cluster at caseworker level).

	(1) Days	(2) Total days	(3) Graded days	(4) Days	(5) Total days	(6) Graded days
Summoned (DM)	−0.32	−3.13	−0.98			
	(2.57)	(2.88)	(2.54)			
Meet week 13				−1.83	−6.94	−2.42
				(4.47)	(4.73)	(4.45)
Meet week 19				3.29	4.08	2.58
				(3.65)	(4.07)	(3.41)
Meet week 26				−1.22	−4.52*	−1.78
				(2.45)	(2.56)	(2.26)
Mean dep. var in C.	120.09	164.48	112.37	120.09	164.48	112.37
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.16	0.17	0.21	0.16	0.17	0.21
Controls	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline

Notes: All regressions control for block fixed effects. Robust SE, clustered at the caseworker level, in parentheses.

In Table A.5 we conduct an analysis where we restrict the samples to individuals receiving letters the same week. In Column 1 we compare individuals with letters in week 9 that are either assigned to a meeting in week 13 or not to any meeting at all. For individuals with a letter in week 15 there are three groups: Meeting week 19, meeting week 26, or no meeting. In Column 2 we include people getting letters in week 15 and we exclude individuals that are assigned to a meeting in week 26 and in Column 3 we exclude those assigned to meetings in week 19. In Column 4 we include the two treatments together for the sample that gets a letter in week 15. For all of these specifications the effects are small and not statistically significant. In Column 5, where we look at people receiving late letters we see an effect that is statistically significant at the 10 percent level.

In Tables A.6–A.8 we show results where we cluster the standard errors at the caseworker level (but as randomization is done at the individual level this is not our main specification) with and without controls and we also present results where we use other functional forms. All these tests lead us to the same qualitative conclusions that there are at most minor effects of DM on absence.

Table A.8
Robustness with ln, ihs, and Poisson models for total days and days within spell.

	(1) Ln days	(2) Ln total	(3) IHS days	(4) IHS total	(5) Poisson days	(6) Poisson total
Main Summoned (DM)	−0.026	−0.024	−0.027	−0.025	−0.0040	−0.019
	(0.027)	(0.022)	(0.028)	(0.023)	(0.018)	(0.014)
Mean dep. var in C.	120.09	164.43	120.09	164.43	120.09	164.48
No. of observations	10,235	10,233	10,235	10,233	10,235	10,235
R-squared	0.08	0.09	0.08	0.09		
Controls	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.9
Baseline results with all treatments. Dependent variable is total days of sickness absence.

	(1) All	(2) No meeting	(3) Meeting
DM 13 letter 9	−1.95		
	(4.64)		
DM 19 letter 15	9.07*		8.82*
	(4.66)		(4.81)
DM 26 letter 15	−0.79		1.60
	(4.47)		(4.65)
Letter 15	5.75	5.76	
	(3.94)	(4.10)	
DM 26 letter 22	1.97		3.68
	(4.41)		(4.66)
Letter 22	8.01*	8.69*	
	(4.46)	(4.68)	
Excluded category	Letter 9	Letter 9	DM 13 letter 9
Mean in excl. group	160.49	160.49	160.11
No. of observations	10,235	5094	5141
R-squared	0.17	0.22	0.22
Controls	Baseline	Baseline	Baseline

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

A4. Pre-specified analysis investigating all seven treatments

In this section we investigate the effects of all seven treatments. As there may be a reversed notification effect of receiving a letter stating the absentee will not be summoned to a meeting, we also split the sample into summoned and not summoned to meetings. In Column 1 of Table A.9 the excluded category is receiving an early letter without meeting. We see that the effects go in different directions and that for one treatment, meeting week 19 and letter week 15, the effect is actually positive and statistically significant at the 10 percent level. Furthermore, we see in Column 2, when we only compare the control groups, that those randomly receiving early letters have shorter spells than those receiving later letters. This difference is only statistically significant for the latest letters and then only at the 10 percent level. In Column 3 we only look at people with letters summoning to meetings and we see that early meetings seem to be better than later ones. In Table A.10 we conduct the same analysis for women and men separately.

A5. Additional pre-specified heterogeneity results

We investigate the gender difference further in Tables A.11 and A.12 by looking at the people getting letters the same weeks. We note that one treatment for men stand out with a large positive coefficient. As we are testing many hypotheses, however, we do not put much emphasis on this particular result.

In order to further investigate the heterogeneity by age we present the relationship between birth year and total days of absence graphically for the treatment and control group in Fig. A.3. The regression line for the treated individuals always lie below that of the control individuals but the difference is very small. We also present it separately for men and women in Appendix Figs. A.4 and A.5.

We also conduct separate analyzes for people with only symptoms (Table A.13) and more than symptoms (Table A.14) and we note that the treatment effect seems larger for the former group.

We have also investigated heterogeneity by diagnoses (Appendix Tables A.15–A.18), offices (Appendix Tables A.19 and A.20) and occupations (Appendix Table A.21). We find some heterogeneity according to these aspects, in particular across offices, but not even the office heterogeneity survives a simple bonferroni correction for the number of offices tested. We also investigate heterogeneity in the treat effect in Appendix Table A.22

Table A.10
Results by gender with all treatments. Dependent variable is total days of sickness absence.

	(1) Women	(2) No meeting	(3) Meeting	(4) Men	(5) No meeting	(6) Meeting
DM 13 letter 9	−3.30			0.15		
	(5.85)			(8.37)		
DM 19 letter 15	2.72		5.46	22.3***		17.5*
	(5.81)		(6.21)	(8.51)		(9.45)
DM 26 letter 15	1.45		5.84	−5.26		−1.77
	(5.66)		(6.08)	(7.94)		(9.00)
Letter 15	7.00	7.56		5.34	−0.15	
	(4.98)	(5.37)		(7.10)	(7.99)	
DM 26 letter 22	0.25		5.58	5.90		1.78
	(5.53)		(6.07)	(8.08)		(9.16)
Letter 22	9.45*	11.4*		7.91	4.94	
	(5.62)	(6.05)		(8.06)	(8.97)	
Mean in excl. group	160.30	160.30	160.99	160.81	160.81	158.56
No. of observations	6403	3148	3255	3832	1946	1886
R-squared	0.21	0.30	0.29	0.24	0.35	0.38
Controls	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.11
Effects when comparing to people with letters the same week. Total number of sick days as outcome. Women only.

	(1) Letter week 9	(2) Lw 15, Meet 19	(3) Lw 15, Meet 26	(4) Letter week 15	(5) Letter week 22
Meet week 13	−0.35				
	(7.32)				
Meet week 19		−5.80		−4.91	
		(5.83)		(5.54)	
Meet week 26			−2.32	−3.23	−7.08
			(5.59)	(5.36)	(6.70)
Mean dep. var in C.	160.30	164.92	164.92	164.92	168.88
No. of observations	1498	2330	2412	3167	1738
R-squared	0.33	0.28	0.27	0.22	0.35
Controls	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.12
Effects when comparing to people with letters the same week. Total number of sick days as outcome. Men only.

	(1) Letter week 9	(2) Lw 15, Meet 19	(3) Lw 15, Meet 26	(4) Letter week 15	(5) Letter week 22
Meet week 13	−1.40				
	(12.0)				
Meet week 19		28.7***		21.4***	
		(8.77)		(8.06)	
Meet week 26			−15.0**	−10.1	−2.67
			(7.63)	(7.25)	(10.9)
Mean dep. var in C.	160.81	163.65	163.65	163.65	166.56
No. of observations	853	1403	1535	1955	1024
R-squared	0.50	0.37	0.36	0.31	0.46
Controls	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

As described in Section 2.3, every time the caseworkers perform the random draw of each person on sick leave, they are also asked two survey questions. The results are discussed in the main document and we here explore some additional issues. Table A.23 we investigate the characteristics that are correlated with the caseworkers thinking the absentees will be away for over 20 weeks. We see that caseworkers think younger people, people not on sick leave full time, and people with diagnoses characterized as symptoms are less likely to be absent for a long time and that people with more sickness absence before have a larger probability of being absent for a long time. As we show in Appendix Table A.24, these characteristics are indeed correlated with long absenteeism. In Appendix Table A.25 we analyze for whom the caseworkers think the meeting is most important. We see that they think meetings are more important for younger people, people with graded absence, people that have been sick for longer periods before, people from larger establishments, and for people with symptoms only.

Table A.13
Baseline results for people with only symptoms.

	(1) Total	(2) Total	(3) Total	(4) Total
Dialogue meetings	−6.27*	−6.33*		
	(3.77)	(3.62)		
Meet week 13			−12.3**	−11.7*
			(6.27)	(6.01)
Meet week 19			−1.03	−1.62
			(6.40)	(6.13)
Meet week 26			−5.89	−6.02
			(4.46)	(4.31)
Mean dep. var in C.	157.99	157.99	157.99	157.99
No. of observations	4075	4075	4075	4075
R-squared	0.19	0.26	0.19	0.26
Controls	Block	Baseline	Block	Baseline

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.14
Baseline results for people with more than symptoms.

	(1) Total	(2) Total	(3) Total	(4) Total
Dialogue meetings	−2.10	−2.45		
	(3.09)	(2.98)		
Meet week 13			−6.26	−7.08
			(5.03)	(4.88)
Meet week 19			4.11	3.91
			(5.01)	(4.86)
Meet week 26			−2.77	−2.98
			(3.68)	(3.53)
Mean dep. var in C.	168.84	168.84	168.84	168.84
No. of observations	6160	6160	6160	6160
R-squared	0.13	0.20	0.13	0.20
Controls	Block	Baseline	Block	Baseline

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.15
Effects by main diagnoses1. Total number of sick days as outcome.

	(1) A	(2) D	(3) F	(4) K	(5) L	(6) Mult
Summoned (DM)	−12.7	11.8	−43.2	59.7*	3.32	−18.3
	(15.0)	(30.9)	(80.7)	(30.8)	(4.27)	(38.2)
Mean dep. var	163.73	166.61	164.98	152.67	161.76	218.37
No. of observations	657	376	116	317	3557	278
R-squared	0.55	0.69	0.87	0.78	0.19	0.70
Controls	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.16
Effects by main diagnoses2. Total number of sick days as outcome.

	(1) N	(2) P	(3) R	(4) S	(5) T	(6) W	(7) X
Summoned (DM)	−15.3	−6.77	19.5	26.2	−70.3	−1.92	10.4
	(19.6)	(4.86)	(60.9)	(49.7)	(51.3)	(11.0)	(89.4)
Mean dep. var	186.68	168.65	159.17	109.04	193.32	109.62	200.84
No. of observations	611	2665	216	226	208	426	110
R-squared	0.62	0.26	0.77	0.82	0.86	0.61	0.95
Controls	Block	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

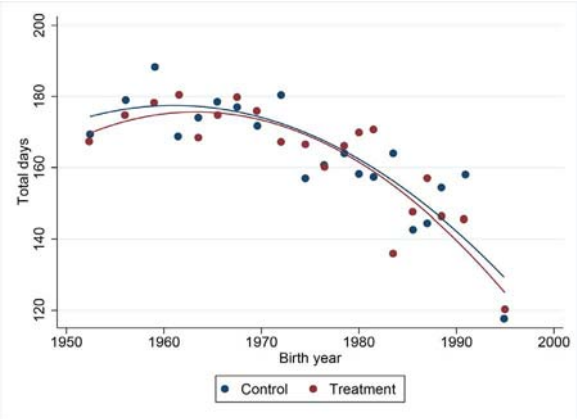


Fig. A.3. Absence by birth years. Total days.

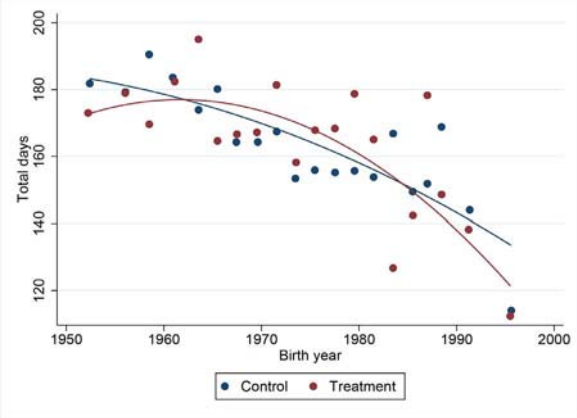


Fig. A.4. Absence by birth years. Total days. Men only.

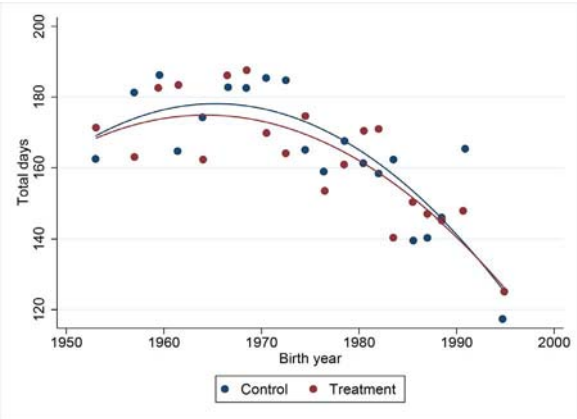


Fig. A.5. Absence by birth years. Total days. Women only.

We also investigate how well the predictions of long spells help predict heterogeneity using the more fine tuned measure of all seven treatments. We start with the sample of individuals that the caseworkers predict will be away for a long time in Appendix Table A.26. In Column 1 we see that receiving an early letter for an early meeting make people come back faster than everyone else. In Column 2 we see that there is no statistically significant relationship between getting letters early or late for people not called in to a meeting. Column 3 focuses on those called in to a meeting and all coefficients are positive, albeit not statistically significant, suggesting that early meetings may be better for this group. It should be noted that the group predicted to have long spells is very small and these

Table A.17
Effects by main diagnoses1. Total number of sick days as outcome.

	(1) A	(2) D	(3) F	(4) K	(5) L	(6) Mult
Meet week 13	−26.7	36.3	−12	2.82	−2.86	−76.7
	(25.2)	(56.6)	(110.7)	(59.7)	(6.86)	(71.1)
Meet week 19	−0.076	65.5	65.0	85.4	13.8**	−6.27
	(25.0)	(45.2)	(.)	(61.4)	(6.99)	(54.2)
Meet week 26	−12.1	−2.67	−114.5	66.0*	1.46	−5.67
	(17.4)	(32.6)	(132.0)	(33.8)	(5.07)	(40.9)
Mean dep. var	163.73	166.61	164.98	152.67	161.76	218.37
No. of observations	657	376	116	317	3557	278
R-squared	0.56	0.70	0.88	0.79	0.20	0.71
Controls	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.18
Effects by main diagnoses2. Total number of sick days as outcome.

	(1) N	(2) P	(3) R	(4) S	(5) T	(6) W	(7) X
Meet week 13	−17.5	−10.6	−18.6	12.8	−45.1	−28.8*	54.0
	(30.1)	(8.33)	(141.7)	(82.1)	(91.1)	(16.5)	(.)
Meet week 19	−16.2	−7.45	45.8	21.1	−33.5	2.76	−85.9
	(32.6)	(7.62)	(81.2)	(48.0)	(70.7)	(19.8)	(252.4)
Meet week 26	−14.1	−4.81	16.4	32.8	−99.6	9.45	26.2
	(21.7)	(5.79)	(94.3)	(62.2)	(63.0)	(12.7)	(158.7)
Mean dep. var	186.68	168.65	159.17	109.04	193.32	109.62	200.84
No. of observations	611	2665	216	226	208	426	110
R-squared	0.62	0.26	0.78	0.82	0.86	0.62	0.96
Controls	Block	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.19
Effects by office1. Total number of sick days as outcome.

	(1) 1	(2) 2	(3) 3	(4) 4	(5) 5	(6) 6	(7) 7
Summoned (DM)	−15.4	20.4	−17.9**	−12.5**	8.91	8.35	12.5*
	(19.3)	(16.1)	(8.57)	(5.68)	(12.7)	(11.5)	(6.84)
Mean dep. var	195.93	181.26	168.66	163.69	179.95	174.90	163.72
No. of observations	187	240	855	1733	368	461	1139
R-squared	0.08	0.10	0.08	0.11	0.05	0.11	0.05
Controls	Block	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.20
Effects by office2. Total number of sick days as outcome.

	(1) 8	(2) 9	(3) 10	(4) 11	(5) 12	(6) 13
Summoned (DM)	−11.2	1.19	−18.5	−20.9**	5.44	−3.93
	(8.61)	(6.20)	(11.3)	(9.53)	(5.19)	(10.4)
Mean dep. var	180.89	151.12	173.11	179.43	150.35	154.03
No. of observations	748	1274	437	558	1728	507
R-squared	0.10	0.08	0.13	0.11	0.08	0.14
Controls	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

results should therefore be seen as even more suggestive than the other results. For the group that is not predicted to be away for a long time we see that the treatment coefficients are generally smaller as seen in Appendix Table A.27.

Table A.21
Effects by occupations. Total number of sick days as outcome.

	(1) Managers	(2) Professionals	(3) Technicians	(4) Clerical	(5) Service	(6) Agricultural	(7) Plant	(8) Elementary	(9) Other
Summoned (DM)	0.82	−4.99	0.12	−13.9	−11.4**	13.6	−34.6	−8.10	34.5*
	(11.5)	(6.57)	(5.21)	(14.5)	(5.47)	(14.4)	(32.4)	(19.1)	(20.5)
Mean dep. var	171.26	154.60	163.23	173.74	173.76	160.15	166.22	171.24	137.30
No. of observations	894	1834	2517	693	2289	633	345	532	498
R-squared	0.48	0.33	0.24	0.58	0.27	0.56	0.65	0.61	0.64
Controls	Block	Block	Block	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.22
Heterogeneity in threat effects. Dependent variable is returning before week 13.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
DM 13 letter 9	0.032	0.0065	0.0077	0.013	0.0068	0.013	−0.034
	(0.025)	(0.015)	(0.015)	(0.022)	(0.015)	(0.016)	(0.021)
Letter 9	0.022	−0.0090	−0.0090	−0.020	−0.0090	−0.0094	−0.015
	(0.025)	(0.015)	(0.015)	(0.021)	(0.015)	(0.016)	(0.021)
Female*DM13	−0.040						
	(0.031)						
Birth year*DM13		0.0014					
		(0.015)					
Days before*DM13			0.0071				
			(0.013)				
Graded*DM13				−0.013			
				(0.029)			
nr employees*DM13					0.0056		
					(0.015)		
Long spell*DM13						−0.037	
						(0.039)	
Meeting important*DM13							0.078***
							(0.029)
Female*L9	−0.050						
	(0.030)						
Birth year*L9		−0.00094					
		(0.015)					
Days before*L9			0.012				
			(0.013)				
Graded*L9				0.018			
				(0.029)			
nr employees*L9					−0.013		
					(0.014)		
Long spell*L9						−0.0099	
						(0.043)	
Meeting important*L9							0.0076
							(0.029)
Excluded category	All other	All other	All other	All other	All other	All other	All other
Mean in excl. group	0.30	0.30	0.30	0.30	0.30	0.30	0.30
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.08	0.08	0.08	0.08	0.08	0.09	0.09
Controls	Block	Block	Block	Block	Block	Block	Block
Sample	All	All	All	All	All	All	All

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

A6. Duration analysis

We conduct a pre-specified duration analysis where we put more structure on the timing of the relationship between being called in to a meeting and returning to work. For this purpose, we start by restructuring the dataset. For each week in which the absentee is under risk for returning to work (ending the sickness spell) the spell is given an additional line in the dataset. The outcome variable is then whether or not the absentee actually returns to work this week. The advantage of this approach is that it enables us to test more specific behavioral hypotheses, such as *when* workers eventually respond to a letter (e.g. same week, the week after etc.), and *when* the meetings actually impact return to work, if at all, e.g. before the meeting (threat effect), same week or after the meeting.

All models are estimated using Cox proportional hazards models (stcox in STATA). This is a partial likelihood model, implying that we do not estimate the underlying hazard rate or duration dependence. We thus do not impose any parametric assumptions on the shape of the hazard function over the duration.

Table A.23
Caseworker predictions: whom do they think will be away for many weeks?.

	(1) Long spell	(2) Long	(3) Long	(4) Long	(5) Long	(6) Long	(7) Long
Female	−0.0088						0.0036
	(0.0065)						(0.0066)
Birth year		−0.023***					−0.021***
		(0.0033)					(0.0033)
Days before			0.012***				0.0095***
			(0.0032)				(0.0032)
Graded				−0.052***			−0.052***
				(0.0063)			(0.0064)
nr employees					−0.00058		−0.00061
					(0.0032)		(0.0032)
symptoms						−0.038***	−0.032***
						(0.0061)	(0.0061)
Mean dep. var in C.	0.10	0.10	0.10	0.10	0.10	0.10	0.10
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.11	0.12	0.11	0.12	0.11	0.12	0.13
Controls	Block	Block	Block	Block	Block	Block	Block

Notes: Robust SE in parentheses.

Table A.24
Actual heterogeneity in absenteeism. The spell lasts at least 20 weeks.

	(1) At least 20 weeks	(2) 20	(3) 20	(4) 20	(5) 20	(6) 20	(7) 20
Female	−0.00020						0.017
	(0.011)						(0.011)
Birth year		−0.035***					−0.032***
		(0.0052)					(0.0052)
Days before			0.032***				0.029***
			(0.0050)				(0.0050)
Graded				−0.073***			−0.075***
				(0.010)			(0.010)
nr employees					−0.013**		−0.014***
					(0.0051)		(0.0052)
Symptoms						−0.045***	−0.035***
						(0.010)	(0.010)
Mean dep. var in C.	0.49	0.49	0.49	0.49	0.49	0.49	0.49
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.08	0.08	0.08	0.08	0.08	0.08	0.09
Controls	Block	Block	Block	Block	Block	Block	Block

Notes: Robust SE in parentheses.

Table A.25
Caseworker predictions: whom do they think meetings are important for?.

	(1) Important	(2) Important	(3) Important	(4) Important	(5) Important	(6) Important	(7) Important
Female	0.014						0.014
	(0.0090)						(0.0092)
Birth year		0.013***					0.014***
		(0.0045)					(0.0045)
Days before			0.019***				0.021***
			(0.0044)				(0.0044)
Graded				−0.040***			−0.045***
				(0.0089)			(0.0091)
nr employees					0.013***		0.014***
					(0.0045)		(0.0046)
Symptoms						0.019**	0.019**
						(0.0090)	(0.0091)
Mean dep. var in C.	0.50	0.50	0.50	0.50	0.50	0.50	0.50
No. of observations	10,235	10,235	10,235	10,235	10,235	10,235	10,235
R-squared	0.31	0.31	0.31	0.31	0.31	0.31	0.31
Controls	Block	Block	Block	Block	Block	Block	Block

Notes: Robust SE in parentheses.

Table A.26
Results with all treatments for people with long predicted spells. Dependent variable is total days of sickness absence.

	(1) All	(2) No meeting	(3) Meeting
DM 13 letter 9	−3.09		
	(18.1)		
DM 19 letter 15	21.6		28.1
	(18.8)		(20.7)
DM 26 letter 15	10.4		17.9
	(18.8)		(22.8)
Letter 15	4.29	−6.76	
	(16.6)	(23.4)	
DM 26 letter 22	2.73		6.42
	(17.3)		(20.2)
Letter 22	7.33	16.8	
	(18.2)	(26.1)	
Excluded category	Letter 9	Letter 9	DM 13 letter 9
Mean in excl. group	196.00	196.00	190.29
No. of observations	1117	521	596
R-squared	0.49	0.67	0.66
Controls	Baseline	Baseline	Baseline

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

Table A.27
Results with all treatments for people with shorter predicted spells. Dependent variable is total days of sickness absence.

	(1) All	(2) No meeting	(3) Meeting
DM 13 letter 9	−2.69		
	(4.95)		
DM 19 letter 15	6.17		5.73
	(4.93)		(5.16)
DM 26 letter 15	−2.91		−1.75
	(4.68)		(4.96)
Letter 15	5.77	5.85	
	(4.16)	(4.35)	
DM 26 letter 22	0.31		1.47
	(4.70)		(5.06)
Letter 22	6.83	7.81	
	(4.71)	(4.95)	
Excluded category	Letter 9	Letter 9	DM 13 letter 9
Mean in excl. group	156.38	156.38	155.60
No. of observations	9118	4573	4545
R-squared	0.17	0.23	0.22
Controls	Baseline	Baseline	Baseline

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

All models are estimated using the same baseline covariates (and the block fixed effects) as in the linear models above and we also add dummies for the seven treatment groups. The data is right censored as we do not observe the end of all spells.

As these models use the exact timing of the letters and the meetings there will be measurement errors whenever the meetings and letters are not held/sent at the times they are supposed to be. To the extent that the errors in sending of letters or timing of meetings are uncorrelated with the treatment received this will bias our coefficients towards zero.

We estimate 5 different models. First we investigate whether receiving a letter impacts immediate return to work. To do so we construct a time varying variable which start out being equal to 0. One week before the letter arrives the variable is changed to 1. The same week as the letter arrives it equals 2, and the week after it takes the value 3. Then, for two weeks after and onwards it is again equal to 0. To separate between letters actually summoning for a meeting (the treatment group), and letters stating that the absentee will not be summoned to a meeting (the control group) we interact the variable with a dummy for the treatment group. We show the results for this model in Column 1 of Table A.28 and note that there does not seem to be much of an effect of either receiving a letter summoning to a meeting or not in the weeks precisely around receiving the letters. If anything, there is a lower probability of returning to work the week after having been summoned to a meeting.

We move on to compare early and late letters. We now ask whether the timing of the letter has any importance for its effect. We do so by constructing another time-varying variable taking the value of 1 the week the letter arrives and 0 otherwise. This variable is then interacted with a dummy for the week in the spell the letter is arriving (9, 15 or 22) and treatment. The results are shown in Column 2 of Table A.28. The results are imprecise.

Table A.28
Cox models 1–3

	(1) All	(2) All	(3) All
Week before letter	−0.14*		
	(0.082)		
Letter week	−0.13		
	(0.089)		
Week after letter	0.075		
	(0.085)		
Week before letter*DM	0.14		
	(0.10)		
Letter week*DM	0.094		
	(0.11)		
Week after letter*DM	−0.23**		
	(0.11)		
Effect in the letter week of receiving a letter in week 9		−0.17	
		(0.14)	
Effect in the letter week of receiving a letter in week 15		−0.069	
		(0.14)	
Effect in the letter week of receiving a letter in week 22		−0.42*	
		(0.24)	
... in week 9 in the letter week*DM		0.34*	
		(0.18)	
... in week 15 in the letter week*DM		−0.12	
		(0.17)	
... in week 22 in the letter week*DM		0.29	
		(0.31)	
Control group			−0.13
			(0.089)
Mandatory meeting in 4 weeks			0.0017
			(0.096)
Mandatory meeting in 11 weeks			−0.13
			(0.17)
No. of observations	178,459	178,459	178,459
Controls	Baseline	Baseline	Baseline

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

In Column 3 of Table A.28 we separate between three types of letters and their effect in the week the letter arrives: those for a voluntarily meeting, those saying that there is a meeting in four weeks from now and those saying there is a meeting 11 weeks from now. There are no statistically significant effects of any of these different types of letters in this specification.

In Column 1 of Table A.29 we set up a model in which we can estimate effects of letters as well as meetings. The model contains a dummy variable indicating that the individual has received a letter. This variable is turned on the week the letter is sent, and never turned off again. It also includes a dummy indicating that you receive a letter this week. This variable takes the value 1 in the letter-week, and zero otherwise. The model also includes an interaction between DM and having received a letter as well as an interaction between DM and being in the letter week. The model further includes a set of dummy variables for being 1, 2, or 3 weeks before the meeting, a dummy for the meeting week and a dummy for all the weeks after the meeting. Analyzing the independent effects of the meetings is tricky as this involves mediation analysis. The main problem is that the only ones having meetings constitute a selected sample of individuals not induced to end their spell by receiving the letter. The variables for having received the letter and its interaction with DM are included to mitigate the potential selection problem arising from the fact that there is a positive/negative effect on return to work from receiving a treatment/control letter.

Focusing on the effects of the meetings we see that three weeks before the meeting is supposed to be held, return to work decreases by as much as 22 percent. This could be interpreted as a lock in effect but the estimates are not vary stable (for instance, we found very different results in the early train sample). The effect is also only statistically significant at the 10 percent level. Finally we want to investigate whether being called in to meetings have the largest effect early or late. We construct a variable which takes the value 1 the week before and same week as the meeting is held, and zero otherwise. This variable is then interacted with the timing of meeting (i.e. week 13, 19, or 26). The results in Column 2 of Table A.29 indicate that that there is not much going on.

As a supplement to these duration models, we have also specified models capturing essentially the same identifying variation within a linear probability model.[14] Starting with the original dataset, we first expand the dataset to contain one observation per week. Hence, for a spell lasting from week 8 to week 17 we get 10 observations, where the outcome (return to work) equals 0 on the first 9 lines and 1 on the last line. Second, we construct a set of time-varying variables capturing the treatments. These are:

"Letter": Equals 1 the week the letter is received, 2 the week after, 3 the second week after, and 0 all other weeks.

[14] We are thankful to one the anonymous referees for this suggestion and this was not pre-specified.

Table A.29
Cox models 4 and 5.

	(1) All	(2) All
Having received a letter	0.076	
	(0.052)	
Letter week	−0.15	
	(0.093)	
Having received a letter*DM	−0.050	
	(0.086)	
Letter week*DM	0.064	
	(0.13)	
3 Weeks before meeting	−0.22*	
	(0.12)	
2 Weeks before meeting	0.016	
	(0.12)	
Week before meeting	−0.015	
	(0.12)	
Meeting week	−0.043	
	(0.13)	
All weeks after meeting	−0.072	
	(0.082)	
Control group and letter is received		0.053
		(0.048)
Mandatory meeting and letter is received		−0.056
		(0.049)
Week before meeting and meeting week if meeting week is 13		0.043
		(0.12)
Week before meeting and meeting week if meeting week is 19		0.21
		(0.13)
Week before meeting and meeting week if meeting week is 26		−0.13
		(0.14)
No. of observations	178,459	178,459
Controls	Baseline	Baseline

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

"Letter x treat": Equals "Letter" for those with compulsory meetings and 0 for all others.
"Meeting": Equals 1 the week before the meeting, 2 the week of the meeting, 3 the first week after the meeting, 4 the second week after, and 5 all later weeks. For all other weeks (before), as well as for the control group, this variable equals 0.

Finally, we expand the model with a complete set of fixed effects for BLOCK x duration (week). We estimate the model with, and without individual fixed effects. The role of individual fixed effects is to capture the dynamic selection arising from the effect of preceding treatments. The results, presented in Table A.30 below, can be summarized as follows. (1) We find no effect of the letter informing of voluntarily meetings. (2) We find no effect of the meeting summoning the absentee to a compulsory meeting, in the same week as the letter is sent. (2) We find a negative effect (reduced return to work) the first week after the letter is sent. (3) This effect if (partly) offset by a positive effect the second week after the letter is sent. The relative size of these effects differ between the two specifications. 4) We find no effects of meetings, either the week before, same week, and we two first weeks after the meeting is held. In the model without individual fixed effects (column 1) we do however find a significant negative long term effect of meetings. In the model with individual fixed effects, this effect is not present.

A7. Machine learning results

The causal tree algorithm (Athey and Imbens, 2016) is a modification of the regression tree, popular in the machine learning literature for prediction purposes. It reformulates the usual problem of estimating a treatment effect as a prediction problem, and aims at estimating heterogeneous treatment effects along observable covariates. Starting with the initial sample, the causal tree considers all the possible ways to split the data in two partitions along the available covariates. It then select the single split that delivers the highest possible treatment heterogeneity between the two partitions resulting from the split, at the same time penalizing within partition variance. Each partition, called leaf, is further split in two sub-leaves, but only if splitting yields gains relative to stopping. To sum up, the algorithm partitions the data in groups that are heterogeneous in terms of treatment effects. In particular, honest causal trees are those where the structure of the tree is built using only a random subset of the available data, while the treatment effects within each leaf are estimated using the other subset of the data. This procedure ensures that the leaves structure is exogenous with respect to the second subset, and so standard inference can be used to estimate the treatment effects.

We noted that in our case the predictions by each honest causal tree are very sensitive to the initial seed chosen to split the data into $S1$ and $S2$. In other words, different trees yield different predictions. We therefore do not present any result obtained from individual trees but we immediately turn our attention to forests, whose predictions are obtained aggregating results from several

Table A.30
Linear probability model, return to work by duration.

Letter	(1)	(2)
Same week	−0.006	−0.007
	(0.005)	(0.004)
+1 week	0.006	0.003
	(0.005)	(0.005)
+2 weeks	−0.002	−0.005
	(0.005)	(0.005)
Letter × Treatment		
Same week	0.001	0.003
	(0.006)	(0.005)
+1 week	−0.015**	−0.010*
	(0.006)	(0.006)
+2 weeks	0.009	0.011*
	(0.007)	(0.006)
Meeting		
−1 week	−0.003	−0.000
	(0.005)	(0.005)
Same week	−0.001	0.002
	(0.005)	(0.005)
+1 week	−0.000	0.002
	(0.005)	(0.005)
+2 weeks	0.001	0.002
	(0.005)	(0.005)
+3 weeks or more	−0.006**	−0.003
	(0.003)	(0.003)
Treatment group FE	Yes	No
Individual FE	No	Yes
Block × duration	Yes	Yes
$R2$	0.162	0.29
N	173,851	173,049

Notes: Standard errors clustered at the individual level in parantheses.

Table A.31
Summary statistics of CATEs predicted by the causal forest.

	count	mean	sd	p50	p25	p75	min	max
CATE	5117	−4.5	6.4	−4.5	−8.7	−0.1	−27.8	19.6

Note: CATEs predicted on the test sample S2 based on the forest grown using the training sample S1.

trees. As expected, we find the honest causal forest to be less sensitive to the initial split between $S1$ and $S2$, especially when we grow sufficiently many trees. We first draw $S1$ and feed it to the *causal forest* algorithm from the *grf* package.[15] The algorithm uses a random 50% of $S1$ to grow each tree. For each tree a further random half of this data is used to chose the splits (the structure of the tree), and the rest is used to populate the leaves (classify observations into leaves) and calculate treatment effects at the leaf-level.[16] Next, the algorithm predicts conditional average treatment effects (CATEs) for each observation in $S2$ as follows. First, it classifies observations in $S2$ in the appropriate leaf for each tree. Based on this information, each observation in $S2$ gets assigned a list of neighboring observations in $S1$, weighted by how many times observations fall in the same leaf. As a last step, the algorithm calculates the treatment effect using the outcomes and treatment status of the neighbor observations from $S1$.[17]

The *grf* package is still under development, and the algorithm fails to grow very large forests using our data. As suggested in the documentation, we circumvent this problem by growing 600 different forests[18] with 5000 trees each, and then averaging CATEs across forests. Our final predictions are thus based on 3 million underlying trees, while we pre-specified that we would grow at least 25,000 trees.

Coming to the results, the distribution of the CATEs is plotted in Fig. A.6, and their summary statistics are reported in Table A.31. The distribution of CATEs is centered around 0 days. Taking the results at face value, the forest points to substantial heterogeneity in treatment effects.

[15] We also provide the algorithm with the true propensity scores from the experimental design.
[16] That is we use the honest mode, as pre-registered. Other pre-registered choices are the following: the algorithm only considered half of the available covariates at each split (parameter *mtry*), and poses little restriction on the minimum number of observations in each final leaf (parameter *min.node.size* = 4).
[17] See https://github.com/grf-labs/grf/blob/master/REFERENCE.md for more details on the algorithm.
[18] We select the parameter *seed* at random for each forest.

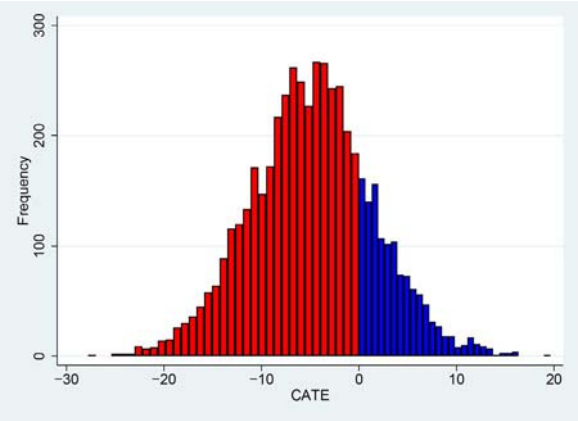


Fig. A.6. Histogram of CATEs in S2 as predicted by the causal forest.

In order to test whether our forest was able to detect true heterogeneity instead of picking spurious correlations, we follow Davis and Heller (2017). Using OLS, we regress the outcome on the following variables: a dummy for those observations that the forest predicts to have positive CATEs, a dummy for those predicted to have negative CATEs, their interactions with DM, and block fixed effects. We do not include the constant and DM in the model, so that the coefficient on each interaction captures the average treatment effect in the corresponding group. The mean outcomes in the control groups are captured by the other two coefficients. We run the regression separately for $S1$ and $S2$.

In sample $S1$, the predictions of the forest are accurate: the OLS estimates of the treatment effect in the two groups have the expected sign, and the magnitude is implausibly large, likely due to overfitting. The results in Column 1 suggest that in-sample the forest is able to classify observations in groups ranked by their treatment effect magnitude. Running the same regression in $S2$ (out of sample) yields very different results: the two treatment effects are both small, negative, not significantly different from zero, and not significantly different from each other (Column 3 of Table A.32).

These results suggest that our forest has not been able to detect any true treatment heterogeneity, and reiterates why it is important to test forest predictions out of sample. We further test the forest predictions by splitting the samples at the quartiles of CATEs, instead that at CATE equal to zero, and by estimating the treatment effects by OLS in $S1$ and $S2$ separately. We adopt an analogous specification as the one used above. Again, the forest predictions are accurate in sample $S1$ (Column 2): the treatment effect is equal to −152 days in the first quartile, −43 in the second, +31 in the third, and +146 in the fourth. On the contrary in sample $S2$ (Column 4) treatment effects are small, do not always have the expected sign and magnitude, and are not significantly different from zero.

We also experiment further with more analyses not included in the preregistration. Most notably, we split the distribution of CATEs at finer percentiles (deciles, quintiles, ...) and we estimate treatment effects by OLS in the different parts of the CATE distribution. We always fail to find convincing evidence of any significant treatment effect consistent with what predicted by the forest (results not reported). Overall, we conclude that the forest has been unable to detect any true treatment heterogeneity.

A8. Comparison with Markussen et al. (2017)

A8.1. Comparison of our reduced form effects with their meeting attendance effects

Our reduced form estimates on absence duration are measures the effects of being summoned to a meeting and, as such, are not directly comparable to the effects of a meeting held, as estimated in Markussen et al. (2017). Directly replicating the latter findings was not the primary goal of our experiment, and as such, the design was not tailored to this aim. However, we can use our treatment letter as an instrument for having a meeting to estimate its effect on the compliers. To do so, we need to assume that the exclusion restriction holds, which in this context amounts to assuming that there is no threat nor reversed threat effect; in the next section, we provide evidence that this is indeed the case in our experiment. When rescaling the coefficients in Table 4 (reduced form) using the effects on DM-participation presented in Table 3 (first-stage), we find that, on average, meetings reduce absence by 23 days per meeting held (the corresponding IV estimates are reported in Table A.33). If we conduct an IV-analysis using the same outcome variable as in Markussen et al. (2017), days of absence within the spell, we find that holding a meeting reduces absence by 3.5 days (see Column 4 of Table A.34). However, we cannot reject that the effects are much larger or even relatively large and in the other direction.

Two notes of caution are warranted: a) our IV estimate is very imprecise; b) the first-stage coefficient is small because most people go back to work nevertheless, and there is a long lag between the letter and the scheduled meeting; this implies that the magnitude of the IV effect hinges mostly on dividing a small reduced-form estimate by a small number.[19] We did not pre-specify a rescaling of the

[19] In an attempt to increase the power we also conducted analyses where control variables were selected through a double robust LASSO procedure (Belloni et al., 2014), but as can be seen in Table A.34 this did not help much. This analysis was not pre-registered.

Table A.32
Causal forest post estimation.

	Training sample S1		Test sample S2	
	(1) Total	(2) Total	(3) Total	(4) Total
DM * 1(NegCATE)	−59.52***		−2.67	
	(3.73)		(3.93)	
DM * 1(PosCATE)	135.52***		7.77	
	(5.43)		(6.73)	
1(NegCATE)	251.88***		4.95	
	(5.02)		(5.79)	
1(PosCATE)	163.48***		2.23	
	(5.43)		(6.73)	
DM * 1(1st quartile CATE)		−152.35***		1.76
		(5.67)		(6.99)
DM * 1(2nd quartile CATE)		−42.83***		−9.53
		(5.90)		(6.60)
DM * 1(3rd quartile CATE)		30.59***		−1.66
		(6.05)		(6.74)
DM * 1(4th quartile CATE)		146.63***		9.05
		(5.61)		(6.64)
1(1st quartile CATE)		300.35***		295.26***
		(5.67)		(7.22)
1(2nd quartile CATE)		237.69***		298.19***
		(5.76)		(6.93)
1(3rd quartile CATE)		203.80***		298.69***
		(5.59)		(6.76)
1(4th quartile CATE)		157.15***		292.95***
		(5.17)		(6.64)
p-value	0.00	0.00	0.41	0.38

Note: OLS regressions with block fixed effects and without constant term. The outcome variable is Total. Robust standard errors in parenthesis. 1(NegCATE) is a dummy for observations with predicted negative CATE, and 1(PosCATE) is analogously defined. 1(x quartile CATE) are dummies for observations in different quartiles of the CATE distribution. *P*-value refers to the test between the two interactions in Columns (1) and (3), and between the interaction with the bottom quartile, and the interaction with the top quartile in Columns (2) and (4).

Table A.33
IV estimates on the effects of meeting attendance.

	(1) Total days	(2) Total days	(3) Total days	(4) Total days
Had meeting	−22.5	−9.33	−17.0	−154.2
	(16.6)	(20.3)	(21.8)	(112.7)
Mean dep. var in C.	164.48	160.49	164.44	167.97
No. of observations	10,235	2351	5122	2762
First stage *F*-statistic	320.0	174.4	179.8	9.5
R-squared	0.03	0.20	0.11	0.30
Letter sent in week	9,15,22	9	15	22

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

effect as we thought the assumptions were too restrictive, and the total effect of summoning to meetings was the policy lever we were trying to optimize and target. An alternative and a not-pre-registered way to test for attendance effects is to restrict the sample to individuals summoned to a meeting in week 15 and then regress a dummy variable for returning between week 19 and 26 on having the meeting in week 19 instead of in 26. The effect of this exercise is not statistically significant, but we can reject that the earlier meeting increased the probability of returning in this period with more than 0.4 percentage points. However, to more credibly test for a pure meeting effect, another type of experiment should be run. Such an experiment could, for example, summon everyone and then randomly cancel some of the meetings close to the meeting date. This type of intervention would probably result in a first-stage coefficient close to one, but it is unfeasible in practice due to existing regulations.

A8.2. Comparison of threat effects

Our relatively precise zero threat effect is not directly comparable to the estimates in Markussen et al. (2017) (MRS) as they estimate an event-history model and model exit from sick-leave and its dependence on (a predicted) risk of receiving a letter. In order

Table A.34
Effects of DM.

	(1) OLS: Total days	(2) IV: Total days	(3) OLS: Days	(4) IV: Days	(5) IV: Days
Dialogue meetings (DM)	−3.24		−0.28		
	(2.13)		(2.05)		
Had meeting		−23.6		−3.46	−2.03
		(15.9)		(15.7)	(14.8)
Mean dep. var in C.	164.48	164.48	120.09	120.09	120.09
No. of observations	10,235	10,235	10,235	10,235	10,235
R-squared				0.07	
Controls	Optimal	Optimal	Optimal	Block	Optimal

Notes: All regressions control for block fixed effects. The optimal controls are selected from a wide set of controls, including diagnoses, by a double LASSO approach. Robust SE in parentheses.

to facilitate such a comparison, we can do the following back-of-the envelope calculation: From the weekly hazard rate presented in figure 5 we can calculate the probability of leaving sick-leave before a given week. The effect estimated by MRS implies a substantial increase in this hazard rate (223 percent), but for one week only. We can then alter one of these hazard rates to mimic the effect of receiving a letter on that week's hazard rate and then re-calculate the probability of leaving sick leave before a given week (i.e. week 13/19/22), and compare the difference between these two predictions to the estimates in Table 5. We then find that our estimates for the threat effect of early letters/meetings (letter 9/DM 13) are much smaller than the estimate in MRS (somewhere between a third and a fourth). For letters sent in week 15 summoning for meetings in week 19, the difference between our estimate and the implied estimate from MRS is much smaller (about 60 percent). Finally, for letters summoning to late meetings (letter 15/DM 26), our estimate and the (implied) estimate from MRS are very similar. The vast majority of the meetings (and letters) studied by MRS are actually late, with meetings typically taking place between week 22 and week 27.

To the extent that the effects are comparable there are other differences between the studies as well. Despite the evaluated policy being the same, some contextual differences might explain the differences between the two studies. Approximately 70% of our sample is composed by individuals in the Oslo area (the capital city), while Markussen et al. (2017) leverage quasi-random variation in meeting propensity across regions, with the Oslo area being one of those with the lowest propensity. As such, the treated populations in the two studies could be different, and even if we do not find evidence of treatment heterogeneity along observables (see next section), many potentially relevant characteristics remain unobserved (e.g. risk preferences). Furthermore, the two analysis were conducted almost ten years apart; even though the labor market conditions in the two periods were not radically different, the earlier study focused on a period when the policy was relatively new, while individuals in our analysis had more time to become accustomed to it, and so perhaps became less susceptible to being scared by a pure threat effect. Furthermore, in the previous study most meetings were held in week 26, and the letters were assumed to be sent on week 23, while our experiment allows only to test threat effects generated by letters sent as late as week 15 (recall that we always use individuals who have not yet received a letter yet as control group); it is thus possible that threat effects only materialize for long-term absentees. Finally, there is always the possibility that the previous study, as every observational one, suffers from some form of omitted variable bias, despite all efforts exerted by the authors. At the same time, we recognize that decisive conclusions can not be drawn from a single RCT, and replicating experiments is a useful practice that should apply here as well.

A9. Non pre-specified exploratory results

After having received the results from the early sample we added survey questions to the caseworkers in order to get a better sense of how the meetings are conducted and to have some information about the background of the caseworkers. The caseworkers only answered these questions once. As we conducted this survey during the middle of the trial period we do not have information from all caseworkers. In total 53 workers answered the survey questions and these workers have together assigned 9147 individuals to different treatment conditions during the total trial period in the analysis sample. The details of the analyses in this section were not pre-specified and we only wrote that we would survey the caseworkers and that one interesting question is whether meetings work better when the physician attends.

In Table A.35 we present heterogeneity results based on the answers to this survey. All standard errors are now clustered at the caseworker. We start by investigating whether there is heterogeneity in the effects of meetings where the physician participates or not. The survey question to the caseworkers is "Is a physician usually participating in the meeting?". We code the variable Doctor participate as equal to 1 if the answer is "Yes, physically" or "Yes, by phone", and to zero if the answer is "No", or "Do not know". The physician is participating in 68 percent of the cases. In Column 1 we see that there seems to be substantial heterogeneity with respect to this variable such that those assigned to meetings with a physician present have 13 days fewer days of absence as compared to those that were not assigned to a meeting.

In Column 2 we test if there is heterogeneity with respect to the education type of the caseworker. We code a dummy variable equal to one if the caseworker is educated as a social worker (18 percent of the sample) and zero otherwise. In Column 3 we add

Table A.35
Heterogeneity with respect to variables in the caseworker (CW) survey. Dependent variable is total days of sickness absence.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dialogue meetings (DM)	5.42	−4.58	−2.98	−3.26	3.07	12.3	12.5
	(4.27)	(3.73)	(3.44)	(12.4)	(5.98)	(15.1)	(15.2)
Doctor in meeting	13.8**					14.0**	15.5**
	(5.31)					(6.20)	(7.36)
CW socialworker		−0.080				−2.93	0.22
		(7.06)				(7.74)	(8.18)
CW male			2.71			2.17	3.37
			(3.96)			(4.23)	(4.30)
CW age				−0.10		0.046	0.17
				(0.20)		(0.26)	(0.26)
CW more than 5 years					1.31	0.85	−3.41
					(4.48)	(5.26)	(5.06)
Doctor in meeting*DM	−13.4**					−13.5**	−12.3*
	(5.93)					(6.60)	(6.31)
CW socialworker*DM		2.57				8.97	9.63
		(7.21)				(9.55)	(9.02)
CW male*DM			−4.63			−1.63	−2.83
			(8.47)			(7.99)	(7.47)
CW age*DM				−0.019		−0.033	−0.063
				(0.26)		(0.33)	(0.34)
CW more than 5 years*DM					−9.68	−9.45	−8.80
					(7.21)	(7.36)	(7.31)
Mean in control group	162.58	162.93	162.93	162.93	162.93	162.93	162.93
No. of observations	9301	9149	9149	9149	9149	9149	9149
R-squared	0.10	0.09	0.09	0.09	0.09	0.10	0.15
Controls	Block	Block	Block	Block	Block	Block	Balance variables

Notes: All regressions control for block fixed effects. Robust SE clustered at the CW level in parentheses.

Table A.36
Characteristics of people with caseworkers usually having the doctor on the meetings or not. Dependent variable is having the doctor on the meetings.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	−0.032***								−0.027***
	(0.0055)								(0.0055)
Birth year		0.00070***							0.00068***
		(0.00023)							(0.00023)
Days before			−0.000029						−0.0000093
			(0.000023)						(0.000023)
Grade				0.00048***					0.00042***
				(0.000091)					(0.000093)
nr employees					−0.0000065***				−0.0000052***
					(0.0000012)				(0.0000012)
Symptoms						−0.0023			−0.0012
						(0.0051)			(0.0052)
CW predicted meeting important							−0.017***		−0.016***
							(0.0053)		(0.0053)
CW predicted long spell								−0.023***	−0.022***
								(0.0072)	(0.0072)
Share of individuals treated	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
No. of observations	9301	9301	9301	9301	9301	9301	9301	9301	9301
R-squared	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.78
Controls	Block	Block	Block	Block	Block	Block	Block	Block	Block

Notes: All regressions control for block fixed effects. Robust SE in parentheses.

information about the caseworkers sex (25 percent are male) and in Column 4 we add information about the age of the caseworker. In Column 5 we add information about working experience and dummy code a variable to be equal to 1 if the respondents have worked as caseworkers for more than 5 years (74 percent of caseworkers). We see that there appears to be no significant heterogeneity with respect to these variables.

The heterogeneity with respect to the participation of the physician in the meetings is robust to controlling for the other survey questions as seen in Column 6. There are several important caveats to interpreting this heterogeneity as having to do with the

Table A.37
Heterogeneity with respect to variables in the caseworker (CW) survey. Sample is the early training data. Dependent variable is total days of sickness absence.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dialogue meetings (DM)	−6.30	−10.0	−15.1	2.12	2.02	−4.16	−14.9
	(18.3)	(8.67)	(10.4)	(33.8)	(18.1)	(56.3)	(54.4)
Doctor in meeting	45.1***					38.5**	29.9*
	(12.3)					(15.0)	(17.1)
CW socialworker		−4.27				−10.1	9.13
		(22.0)				(18.1)	(18.2)
CW male			−11.6			−22.6	−18.1
			(15.0)			(14.5)	(15.1)
CW age				0.087		−0.72	−0.80
				(0.42)		(0.66)	(0.62)
CW more than 5 years					17.7	49.9**	46.6**
					(11.4)	(19.8)	(19.3)
Doctor in meeting*DM	−7.46					−2.34	12.1
	(19.6)					(24.6)	(26.1)
CW socialworker*DM		−16.2				−7.12	−15.0
		(13.6)				(34.8)	(33.5)
CW male*DM			7.70			16.9	14.8
			(13.6)			(20.0)	(18.5)
CW age*DM				−0.32		0.29	0.27
				(0.70)		(1.46)	(1.37)
CW more than 5 years*DM					−18.7	−32.4	−32.6
					(19.7)	(26.9)	(26.7)
Mean in control group	182.78	182.35	182.35	182.35	182.35	182.35	182.35
No. of observations	1251	1229	1229	1229	1229	1229	1229
R-squared	0.10	0.10	0.10	0.10	0.10	0.11	0.18
Controls	Block	Block	Block	Block	Block	Block	Balance variables

Notes: All regressions control for block fixed effects. Robust SE clustered at the CW level in parentheses.

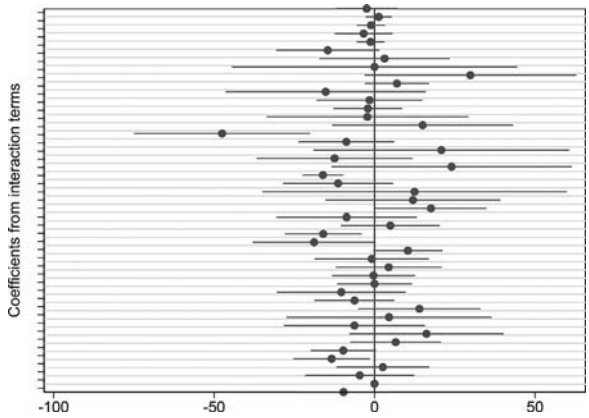


Fig. A.7. Plot of heterogenous treatment effects.

physicians presence, however. First of all, we are testing heterogeneity across many different dimensions and it is likely that some of them will turn out to be statistically significant by pure chance. In fact, even if we only consider the other variables asked to the caseworkers and adjust the *p*-values for multiple testing the heterogeneity is no longer statistically significant at the 5 percent level (adjusted to 0.01 to account for 5 different tests). Furthermore, we see that the heterogeneity seems to be driven by the base category. Taking the results at face value they would imply that individuals with caseworkers that usually include the physician in the meetings have longer sickness absence than other absentees, but only if they do not have meetings. This may reflect that the pool of absentees in the different categories are different. In Table A.36 we show that there are many differences in absentee characteristics across caseworkers with different meeting types. The results are similar if we control for caseworker covariates, however, as seen in Column 7 of Table A.35. To further check if the heterogeneity is indicates a real pattern we test whether we find similar patterns in our early training data (that has not been used to conduct a similar analysis before). The results are seen in Table A.37. We see that the treatment heterogeneity with respect to physician presence is negative also in this sample but it is smaller in magnitude and it actually turns positive when we add controls. Also in this sample do we observe that the base category with doctors present but without meetings have longer absences. A final caveat is of course that we have no way of knowing if there is some omitted variable, such as caseworker seriousness or other office practices (the variable varies mostly between offices and only varies within three offices),

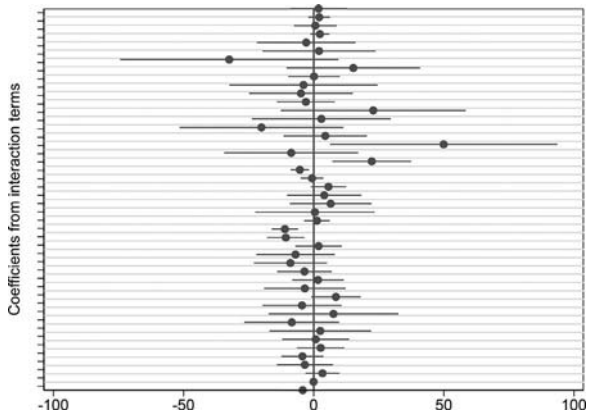


Fig. A.8. Plot of heterogenous treatment effects with placebo outcome (Days before).

that is correlated with both having the physician in the meetings and reduced absence. For instance, Granqvist et al. (2017) find that caseworkers attitudes towards rules and rehabilitation methods were strongly correlated with client work resumption in Sweden.

A10. Plots of heterogenous effects

References

Anderson, M.L., Magruder, J., 2017. Split-Sample Strategies for Avoiding False Discoveries. Technical Report. National Bureau of Economic Research.
Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proc. Natl. Acad. Sci. 113 (27), 7353–7360.
Athey, S., Imbens, G.W., 2017. The econometrics of randomized experiments. In: Handbook of Economic Field Experiments, vol. 1. Elsevier, pp. 73–140.
Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. Ann. Stat. 47 (2), 1148–1178.
Autor, D.H., Duggan, M., 2010. Supporting Work: A Proposal for Modernizing the US Disability Insurance System. Center for American Progress and The Hamilton Project.
Belin, A., Dupont, C., Oulès, L., Kuipers, Y., Fries-Tersch, E., 2016. Rehabilitation and Return to Work: Analysis Report on EU and Member States Policies, Strategies and Programmes. Report. European Agency for Safety and Health at Work.
Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. Rev. Econ. Stud. 81 (2), 608–650.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc.: series B (Methodological) 57 (1), 289–300.
Böckerman, P., Kanninen, O., Suoniemi, I., 2018. A kink that makes you sick: the effect of sick pay on absence. J. Appl. Econom. 33 (4), 568–579.
Bradley, S., Green, C., Leeves, G., 2007. Worker absence and shirking: evidence from matched teacher-school data. Labour Econ. 14 (3), 319–334.
Coffman, L.C., Niederle, M., 2015. Pre-analysis plans have limited upside, especially where replications are feasible. J. Econ. Perspect. 29 (3), 81–98.
Coffman, L.C., Niederle, M., Wilson, A.J., 2017. A proposal to organize and promote replications. Am. Econ. Rev. 107 (5), 41–45.
Dahl, G.B., Kostøl, A.R., Mogstad, M., 2014. Family welfare cultures. Q. J. Econ. 129 (4), 1711–1752.
Davis, J., Heller, S.B., 2017. Using causal forests to predict treatment heterogeneity: an application to summer jobs. Am. Econ. Rev. 107 (5), 546–550.
De Jong, P., Lindeboom, M., Van der Klaauw, B., 2011. Screening disability insurance applications. J. Eur. Econ. Assoc. 9 (1), 106–129.
De Paola, M., Scoppa, V., Pupo, V., 2014. Absenteeism in the italian public sector: the effects of changes in sick leave policy. J. Labor Econ. 32 (2), 337–360.
Duflo, E., Hanna, R., Ryan, S.P., 2012. Incentives work: getting teachers to come to school. Am. Econ. Rev. 102 (4), 1241–1278.
Engström, P., Hägglund, P., Johansson, P., 2016. Early interventions and disability insurance: experience from a field experiment. Econ. J. 127 (600), 363–392.
Fafchamps, M., Labonne, J., 2017. Using split samples to improve inference on causal effects. Polit. Anal. 25 (4), 465–482.
Fevang, E., Markussen, S., Røed, K., 2014. The sick pay trap. J. Labor Econ. 32 (2), 305–336.
Godard, M., Koning, P., Lindeboom, M., 2020. Application and award responses to stricter screening in disability insurance.
Granqvist, N., Hägglund, P., Jakobsson, S., 2017. Caseworkers' attitudes: do they matter? Empir. Econ. 52 (4), 1271–1288.
Hartman, L., Hesselius, P., Johansson, P., 2013. Effects of eligibility screening in the sickness insurance: evidence from a field experiment. Labour Econ. 20, 48–56.
Hesselius, P., Nilsson, J.P., Johansson, P., 2009. Sick of your colleagues 'absence? J. Eur. Econ. Assoc. 7 (2–3), 583–594.
Ichino, A., Maggi, G., 2000. Work environment and individual background: explaining regional shirking differentials in a large italian firm. Q. J. Econ. 115 (3), 1057–1090.
Ichino, A., Riphahn, R.T., 2005. The effect of employment protection on worker effort: absenteeism during and after probation. J. Eur. Econ. Assoc. 3 (1), 120–143.
Johansson, P., Lindahl, E., 2013. Can sickness absence be affected by information meetings? Evidence from a social experiment. Empir. Econ. 44 (3), 1673–1695.
Johansson, P., Palme, M., 2005. Moral hazard and sickness insurance. J. Public Econ. 89 (9–10), 1879–1890.
Kostol, A.R., Mogstad, M., 2014. How financial incentives induce disability insurance recipients to return to work. Am. Econ. Rev. 104 (2), 624–655.
Markussen, S., Røed, K., 2015. Social insurance networks. J. Hum. Resour. 50 (4), 1081–1113.
Markussen, S., Røed, K., Schreiner, R.C., 2017. Can compulsory dialogues nudge sick-listed workers back to work? Econ. J. 128 (610), 1276–1303.
Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. J. Econ. Perspect. 31 (2), 87–106.
Olken, B.A., 2015. Promises and perils of pre-analysis plans. J. Econ. Perspect. 29 (3), 61–80.
Olsson, M., 2009. Employment protection and sickness absence. Labour Econ. 16 (2), 208–214.
Varian, H.R., 2014. Big data: new tricks for econometrics. J. Econ. Perspect. 28 (2), 3–28.
Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. 113 (523), 1228–1242.
Ziebarth, N.R., Karlsson, M., 2010. A natural experiment on sick pay cuts, sickness absence, and labor costs. J. Public Econ. 94 (11–12), 1108–1122.