Effects of Dialogue Meetings - Pre-analysis plan Phase 1*

Matteo Alpino¹, Karen Evelyn Hauge¹, Andreas Kotsadam¹, and Simen Markussen¹

¹Ragnar Frisch Centre for Economic Research

February 12, 2018

Abstract

Sickness absence entails large individual and societal costs. Dialogue Meetings (DMs) where the absentee, the employer, and often the physician, discuss arrangements for full or partial work resumption have been in place in Norway since 2007. In collaboration with the Norwegian Labour and Welfare Administration (NAV) we are conducting a large scale randomized natural field experiment to investigate if and for whom DMs have effects. The experiment is still ongoing and is conducted in several phases to test for mechanisms iteratively. While the experiment is still ongoing, it is possible to analyze data on the individuals that have already been involved in the experiment so far (early sample). To do so we proceeded as follows: first, we randomly divided the early sample into an early training sample and an early test sample. NAV then merged the early training sample with outcome data and covariates and handed the data over to us. Based on the results in the early training sample, we here register some key decisions that will be followed in analyzing the rest of the data (the analysis sample, which comprises the late sample and the early test sample). The results from the early training sample are presented in a companion document and will be subject to a direct replication using the early test sample. Based on the findings in the early test sample we will decide whether and how to change the intervention. Such changes will be pre-registered at a later stage.

Keywords: RCT; Sickness absence; Employment; Causal tree and forest; Machine learning

*We would like to thank Patrik Hägglund and Per Johansson for valuable suggestions. Funding from the Norwegian Research Council (grant number 259512) is also acknowledged.
1 Introduction

We conduct a large scale randomized field experiment to investigate the effects of mandatory dialogue meetings (DMs). In such meetings, the absentee, the employer, 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 number or the nature of tasks at work). A system of DMs have been in place in Norway since 2007 and it is explicitly intended to induce long-term absentees to fully or partly return to work. DMs constitute a combination of nudging, monitoring and facilitation of appropriate working conditions and tasks.

In order test the causal effect of the DMs we team up with the Norwegian Labour and Welfare Administration (NAV) to conduct a large scale randomized field experiment. In particular, we randomize: 1) whether people receive letters calling them in to mandatory or voluntary meetings; 2) when in the sickness spell they receive the letter; 3) and when the meeting is scheduled (if mandatory). Calling in people to a mandatory meeting may have a notification effect in addition to the meeting itself being successful. The notification effect is similar to the "threat effect" in the unemployment insurance literature with reference to the effect of being summoned to an activation program. In the case of DMs, however, the notification effect is a mix of behavioral responses by the employee, the employer as well as by the physician. There may also be a reversed notification effect if people that expect to be called in to a meeting are instead told that a meeting is voluntary for them.

The experiment enables us to answer a wide set of policy relevant questions, such as: Can mandatory dialogue meetings reduce sickness absence and promote work resumption? Is it sufficient to nudge people with the threat of a meeting to make them come back to work faster? Are there adverse effects of calling people in to voluntary meetings? Randomized controlled trials are an excellent tool for recovering average treatment effects without bias. Underlying the average effect there may be substantial heterogeneity and for some
sub-populations the effects may be the opposite of the average effects. From a theoretical perspective, treatment effect heterogeneity may help us understand the mechanisms behind why a treatment works or not and facilitates the generalization of findings from an experiment to other settings. From a policy perspective, improved targeting is also important for cost-efficiency.

However, identifying 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 and spurious correlations 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, Niederle, and Wilson, 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) and 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 heterogenous 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 (2017) label "honest", reduce the problem of multiple hypothesis testing by building replication in the process. A similar idea has already been suggested for randomized experiments as well whereby researchers are encouraged to only use a portion of the data for the analysis and replicate the results using the other portion (Anderson and Magruder, 2017; Fafchamps and Labonne, 2017). We will employ "honest" machine learning techniques to search for heterogeneous treatment effects and we also conduct a sample split approach on our early data to derive some early hypotheses and to guide our efforts in changing the experiment while it is still ongoing.

In particular, we want to test for mechanisms iteratively but in order to do so we must know what the results are. We therefore conduct an analysis on a set of individuals that were included in the experiment early on. To reduce the risks of finding false positives, yet still being able to learn from the data in an explorative fashion, we conduct an analysis on a random sample of the early sample (early training sample). The results of this analysis is presented in the companion document "Effects of Dialogue Meetings - results from the early training sample". Based on the results in this early training sample, we specify some key hypotheses. Some of these hypotheses will be replicated on the early test data (that is the remaining part of the early data). This replication is especially powerful as the data come from the same experiment and the same time period. The results from the replication will guide us in refining the experiment and test additional hypotheses. Had we conducted the analysis on the whole early data we would have been more worried about our results being false positives and had we specified the analysis before analyzing the early data we would have learnt less from the data and some hypotheses would have been left untested. The refinements and new hypotheses that arise from the early test data will be pre-specified at a later stage.

This plan also outlines decisions to guide us in the analysis of the the analysis sample,
which consists of the early test sample as well as all individuals that are included in the experiment after the date of the early sample (late sample). As the analysis sample is larger, some of the methods and tests are more suited to be tested on that sample, in particular the machine learning techniques which are very data intensive and the more fine grained heterogeneity (e.g. by offices, occupations, and diagnoses).

2 The field experiment

The regular procedure of calling in to dialogue meetings differs across the country. In principle, all absentees should have a meeting before week 27 but in practice there are many exceptions to this rule. We work with 13 offices that are willing to change their procedures of targeting for the meetings. Most of these offices are in Oslo (the Norwegian capital), but three of them are outside of Oslo. We started the trial in September 2016 with one office and it has been expanding since then and is still ongoing.

The new procedure consists of several steps and absentees are randomly assigned to seven different treatments. Each week a caseworker at each office checks whom of the absentees are in their 8th week of the sickness spell and are employed. These individuals are assigned to different caseworkers. The assignment of individuals to caseworkers is not randomized and is different in different offices. In some of them, caseworkers are assigned equally many individuals and assignment is by birth dates. At other offices, some caseworkers are assigned more individuals. At yet other offices, some caseworkers are assigned absentees by industry or by employers. When the caseworkers have received their cases they log in to a secure internet page administered by us and write in the id number of the absentee.

Once the absent person has been added to the system, the caseworker is asked two questions: 1) "Based on the information you have, how many more weeks do you think the sickness absence will continue?". The 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 called in to a meeting?”. The alternatives are: Very important, important, neither important nor unimportant, not that important, not important at all. The idea of including these questions is to investigate how caseworker discretion may improve the targeting of the system.

The absentees are then randomly assigned to different treatments, shown in Figure 1. The variations of the treatments are of different types. First of all, we vary whether the meeting is mandatory or not. In principle, everyone has a right to a meeting and NAV does not want to infringe on this right. The individuals not being called in to a mandatory meeting therefore receive an offer to have a voluntary meeting. That is, NAV sends them a letter explaining that they know that the person is on sickness absence but that they have decided not to call in to a meeting. The letter further explains that if the person herself, her doctor, or employer thinks there is a need for a meeting, she can contact the office and a meeting date will be arranged. They also send a similar letter to the employer of the absentee.

For the mandatory meetings, we vary the timing so that they can be held in week 13, 19, or 26 of the absence spell (that is 5, 11, or 18 weeks after the draw). By comparing people randomly assigned to having meetings at different times we can test whether earlier meetings are better than later meetings. The timing of the letter is also varied to test for notification effects and more broadly to test if there is a difference in the effects of early or late letters. The design is set up so that we have overlapping groups of letters and meetings with different combinations. In particular, we randomly assign the timing of letters for people with meetings in week 26 to be either in week 15 or in week 22. If there is a strong notification effect of receiving the mandatory meeting letter we should expect people getting earlier letters to have shorter sickness spells. We also vary the timing of the meeting week to be either in week 19 or in week 26 for people getting letters in week 15. The timing of letters for the voluntary meeting are also randomized to be sent out in week 9, 15, or 22. This
is useful in order to test for a type of reversed notification effect whereby people receiving
the letter with a voluntary meeting may feel less monitored and perhaps stay on sick-leave
longer.

<table>
<thead>
<tr>
<th>Letter is sent</th>
<th>Offer of meeting</th>
<th>Mandatory meeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 9</td>
<td></td>
<td>Meeting in week 13</td>
</tr>
<tr>
<td>Week 15</td>
<td></td>
<td>Meeting in week 19</td>
</tr>
<tr>
<td>Week 15</td>
<td></td>
<td>Meeting in week 26</td>
</tr>
<tr>
<td>Week 22</td>
<td></td>
<td>Meeting in week 26</td>
</tr>
</tbody>
</table>

Figure 1: Different treatments

The offices have different capacities and different wishes of how many meetings to have. We have as our baseline case equal probability of each of the eight cells (there are in fact only 7 groups but we make the voluntary meeting group for letter week 15 double as it is used as a comparison to two mandatory letters), i.e. a 12.5 percent chance of each cell. Some offices, however, wanted other combinations. In particular, many offices wanted a slower start with only late meetings to start with. Other offices need to change the number of meetings over time due to personnel shortages. So far, only one office wanted more than 50 percent of the meetings to be mandatory. In some rare cases, individual caseworkers have had personalized probabilities as well. For instance, one caseworker works with one large employer only and
this employer wants to have meetings for all absentees. In that case, only the timing of the mandatory meetings and letters are randomly assigned.

3 Samples

There are several different samples in the study and we show them in Figure 2. We have an early sample consisting of individuals that have already been involved in the experiment long enough to have meaningful outcome measures. In order to reliably test for early stage heterogenous treatment effects, we randomly divide this sample further into an early training sample and an early test sample. We also have a late sample of people that become enrolled in the experiment at later stages. The final analysis sample will contain the late sample and the early test sample. In this section we describe how we divide the total sample into these sub-samples.

On October 30th (week 44) 2017 we took out all the data from the secured internet page. This data only contains information about the treatment status and not any information about outcomes. At this date, 7,619 absentees had been entered into the system. To start the analysis, we wanted to extract a random sample of individuals that have had the opportunity
to have a meeting. In week 44 of 2017 we could thereby include everyone who had been
drafted until 18 weeks before (as they are drafted in week 8 of their spell and have the
meeting in week 26 of the spell). Hence, we extracted everyone drafted before week 26 of
2017. This amounts to 3,742 individuals and these individuals constitute our early sample.
After some initial power calculations, suggesting that we would be able to detect medium-
sized effects for the mandatory versus voluntary meeting using half of this sample (the
minimum detectable effect was calculated to be 0.13 standard deviations), we decided to
have NAV add outcomes and covariate data for a random half for the early sample. This
half constitutes the early training sample and the other half of the early sample is called the
early test sample.

In taking out the early training sample we wanted to have as similar individuals as
possible in the training sample and in the test sample used for replication. We first stratified
the data based on caseworker and week and randomly assigned half of the observations to
each sample. In cases of odd numbers of absentees we stratified the rest by office and week.
The few remaining ones were randomly assigned within weeks only.

Individuals drafted after week 25 of 2017 up until we change or stop the experiment
constitute our late sample. The final analysis sample will consist of all individuals in the
late sample and the individuals in the early test sample.

4 Data and coding of main variables

Decision rules for dropping observations: For the early training sample we asked NAV
to add data on outcomes as well as background characteristics. We received this data on
January 26th 2018. Some individuals out of the 1871 had to be dropped since they were
drawn after their sickness spell had ended. We also drop one individual with an error in
the sickness absence date. In total, we have 1627 observations with 48 percent assigned to
mandatory meetings and 52 percent assigned to the voluntary meetings. The same decision
Primary and secondary outcome variables: We choose our primary outcome variable to be the total days of sickness absence (Total). This variable counts the total number of days of sickness absence between the draw and the date of data extraction. We have two secondary outcome variables that we expect will give us similar results but that will not be analyzed as extensively nor by themselves be seen as confirmatory. These variables are Days, which is a measure of days within the current sick leave spell and Graded days, which is the grade adjusted total number of days. People may be on sick leave 100 percent, which is the most common, but they may also be partly on sick leave. Graded is a measure of full-day equivalents for the total sick days. If a person is on half time sick-leave for a period, for instance, this measure counts each day as a half day during this period.

The health related part of the Norwegian social insurance program consist mainly of three programs. For employees getting sick the first program is sickness leave (SL), as is mainly studied in this experiment. The worker can be on SL for up to one year, with full wage replacement (up to a threshold of 6 base amounts). If the worker is unable to return to work after 1 year, and her work ability is reduced by at least 50 percent, she can start on temporary disability insurance (TDI). On TDI the replacement rate is 66 percent and one can normally receive TDI for a period up to 4 years. During TDI one is supposed to participate in training and rehabilitation programs and the goal is to return to employment. If this does not succeed one can apply for permanent disability insurance (DI). For the early sample we will not investigate effects on TDI as most individuals have not been on leave for a long enough time to qualify for this. We will add this variable as a secondary outcome variable in later analyzes, however.

Based on the days within the spell we will dummy code whether an individual has returned before week 13, 19, 22, and 26. As described in Section 5, these variables will be used to test the different types of threat effects.
Treatment variables: We will use whether individuals are assigned to receiving a letter for a mandatory meeting (Treatment) or voluntary meeting (Control) as our main treatment variable of interest. We also dummy code each of the seven individual treatment variables and use them in some analyzes (see Section 5).

Other variables: The baseline background variables we use as controls and for heterogeneity are female, 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). In the balance tests these will either be coded as a dummy (female) or as continuous variables. Alternative codings for the heterogeneity analysis are described in Section 5.

We have the diagnoses of all the individuals in the data 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. A distinction is made between symptoms/complaints and established diagnoses within each letter. We create a dummy variable that equals one for individuals that are classified as having symptoms and zero otherwise.

We also create two variables based on the caseworkers predictions. These are CW predicted meeting important which equals 1 if the caseworkers predicted the meeting to be ”important” or ”very important” and CW predicted long spell which equals one if the caseworker predicted a spell of over 20 weeks. For the early training sample the summary statistics are presented in Table 1. We will create the same table for the analysis sample.

Balance tests: To test for balance we will regress our main treatment variable on the other variables described above both individually and together, while controlling for block fixed effects (a combination of week and caseworker). We will judge whether the randomization worked by conducting an F-test of whether the control variables jointly predict treatment.
Table 1: Descriptive statistics.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Treated</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Total days</td>
<td>123.80 (89.1)</td>
<td>121.37 (90.3)</td>
<td>126.44 (87.8)</td>
</tr>
<tr>
<td>Days (within spell)</td>
<td>103.24 (88.7)</td>
<td>101.76 (89.3)</td>
<td>104.85 (88.2)</td>
</tr>
<tr>
<td>Graded days</td>
<td>90.56 (77.9)</td>
<td>87.55 (77.7)</td>
<td>93.83 (78.1)</td>
</tr>
<tr>
<td>Returning before...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 weeks</td>
<td>0.55 (0.5)</td>
<td>0.56 (0.5)</td>
<td>0.53 (0.5)</td>
</tr>
<tr>
<td>19 weeks</td>
<td>0.65 (0.5)</td>
<td>0.66 (0.5)</td>
<td>0.65 (0.5)</td>
</tr>
<tr>
<td>22 weeks</td>
<td>0.69 (0.5)</td>
<td>0.70 (0.5)</td>
<td>0.69 (0.5)</td>
</tr>
<tr>
<td>26 weeks</td>
<td>0.74 (0.4)</td>
<td>0.75 (0.4)</td>
<td>0.74 (0.4)</td>
</tr>
<tr>
<td>Baseline control variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.64 (0.5)</td>
<td>0.64 (0.5)</td>
<td>0.64 (0.5)</td>
</tr>
<tr>
<td>Birth year</td>
<td>1974.92 (11.8)</td>
<td>1974.95 (12.0)</td>
<td>1974.88 (11.6)</td>
</tr>
<tr>
<td>Days before</td>
<td>116.85 (80.1)</td>
<td>115.13 (75.5)</td>
<td>118.72 (84.8)</td>
</tr>
<tr>
<td>Grade</td>
<td>72.79 (27.1)</td>
<td>71.92 (27.0)</td>
<td>73.72 (27.3)</td>
</tr>
<tr>
<td>nr employees</td>
<td>503.16 (1501)</td>
<td>529.38 (1657)</td>
<td>474.69 (1311)</td>
</tr>
<tr>
<td>Other main heterogeneity variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symptoms</td>
<td>0.41 (0.5)</td>
<td>0.40 (0.5)</td>
<td>0.42 (0.5)</td>
</tr>
<tr>
<td>Caseworker predictions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CW predicted meeting important</td>
<td>0.62 (0.5)</td>
<td>0.61 (0.5)</td>
<td>0.63 (0.5)</td>
</tr>
<tr>
<td>CW predicted long spell</td>
<td>0.16 (0.4)</td>
<td>0.17 (0.4)</td>
<td>0.16 (0.4)</td>
</tr>
</tbody>
</table>

**Notes:** The sample consists of the early training sample.

status. For the early training sample the corresponding table is seen in Table 2. We create a similar balance table for the more fine grained definitions of the treatment and as there are 21 different tests involved in comparing them to each other we will not judge the success of the randomization based on any single F-test in the table. A version of this table for the early training sample is provided in Table 3. We will create the same tables for the analysis sample.
Table 2: Test of balance.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
<td>Treated</td>
</tr>
<tr>
<td>Female</td>
<td>0.0049</td>
<td>0.0013</td>
<td>0.00053</td>
<td>0.00040</td>
<td>0.00013</td>
<td>0.00010</td>
<td>0.00034</td>
<td>0.00009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.028)</td>
<td>(0.0011)</td>
<td>(0.0011)</td>
<td>(0.00047)</td>
<td>(0.00049)</td>
<td>(0.0026)</td>
<td>(0.0027)</td>
<td></td>
</tr>
<tr>
<td>Birth year</td>
<td>-0.00057</td>
<td>-0.00051</td>
<td>-0.00013</td>
<td>-0.00010</td>
<td></td>
<td></td>
<td>-0.00060</td>
<td>-0.00034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00016)</td>
<td>(0.00016)</td>
<td>(0.00047)</td>
<td>(0.00049)</td>
<td></td>
<td></td>
<td>(0.0029)</td>
<td>(0.0030)</td>
<td></td>
</tr>
<tr>
<td>Days before</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade</td>
<td>-0.000057</td>
<td>-0.000055</td>
<td>-0.00013</td>
<td>-0.00010</td>
<td></td>
<td></td>
<td>-0.00060</td>
<td>-0.00034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000077)</td>
<td>(0.000078)</td>
<td>(0.000047)</td>
<td>(0.000049)</td>
<td></td>
<td></td>
<td>(0.0029)</td>
<td>(0.0030)</td>
<td></td>
</tr>
<tr>
<td>nr employees</td>
<td>-0.000057</td>
<td>-0.000055</td>
<td>-0.00013</td>
<td>-0.00010</td>
<td></td>
<td></td>
<td>-0.00060</td>
<td>-0.00034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000077)</td>
<td>(0.000078)</td>
<td>(0.000047)</td>
<td>(0.000049)</td>
<td></td>
<td></td>
<td>(0.0029)</td>
<td>(0.0030)</td>
<td></td>
</tr>
<tr>
<td>Symptoms</td>
<td>0.016</td>
<td>0.014</td>
<td>0.016</td>
<td>0.014</td>
<td>0.016</td>
<td>0.014</td>
<td>0.016</td>
<td>0.014</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.027)</td>
<td>(0.026)</td>
<td>(0.027)</td>
<td>(0.026)</td>
<td>(0.027)</td>
<td>(0.026)</td>
<td>(0.027)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>CW predicted meeting important</td>
<td>-0.00060</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0029)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CW predicted long spell</td>
<td>-0.00060</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0029)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share of individuals treated</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of observations</td>
<td>1627</td>
<td>1627</td>
<td>1627</td>
<td>1627</td>
<td>1627</td>
<td>1627</td>
<td>1627</td>
<td>1627</td>
<td>1627</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>Controls</td>
<td>Block</td>
<td>Block</td>
<td>Block</td>
<td>Block</td>
<td>Block</td>
<td>Block</td>
<td>Block</td>
<td>Block</td>
<td>Block</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: All regressions control for block fixed effects (a combination of week of draw and caseworker). Robust SE in parentheses. The F-value from an F-test of the significance of all variables in column 9 is 0.02 (p=0.8985).
Table 3: Balance tests for all seven treatments

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total days</td>
<td>121.120</td>
<td>133.984</td>
<td>129.493</td>
<td>120.389</td>
<td>123.939</td>
<td>114.833</td>
<td>123.991</td>
</tr>
<tr>
<td>T-test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Empirical strategies

5.1 Regression analysis

We will start by estimating the effects of mandatory versus voluntary meetings on the total number of days the absentee is absent after the draw. In doing so, we estimate versions of the following linear regression model:

\[ Total_i = \beta DM_i + \theta Block_{cw} + \lambda X_i + \epsilon_i \] (eq. 1),

where \( DM_i \) is a dummy that equals one for being summoned to a mandatory dialogue meeting for individual \( i \) and zero otherwise, \( Block_{cw} \) are fixed effects for the strata (a combination of caseworker (c) and week (w) fixed effects). These blocking variables are included in all regressions as randomization takes place within these strata. In particular, the assignment of cases to caseworkers is not always random, the treatment share and intensity varies over time, and the types of absentees are likely to change over seasons. \( X_i \) are baseline controls that can be added to possibly increase power. The baseline controls were chosen by testing different specifications in the early training data. These are: a dummy variable for female, dummy variables for age in 5 year groups, the number of days of sickness absence since 2015 and the number of employees at the workplace of the absentee as a continuous variables, and dummies for graded sickness absence in categories of 10. We will present results with and without the baseline controls. We use robust standard errors in all estimations.
The standard errors do not need to be clustered at any level as the randomization is at the individual level (see Abadie et al. (2017)).

The same specification is also run for the secondary outcome variables, i.e. the number of days within the sickness spell that an individual is absent, $Days_i$, and the grade adjusted total number of days, $Graded\ days\ i$. We will later on create a dummy variable for temporary disability insurance (TDI) and run a linear probability model with TDI as outcome. We also test for different treatment effects by replacing $DM_i$ with dummies for different mandatory meeting weeks.

We will also conduct tests where we run regressions for people with letters in the same week but with different content. That is, we will restrict the sample to people getting letters in week 9 and test whether the letter is one for a mandatory meeting in week 13 or for a voluntary meeting. The same type of analysis will be done for individuals receiving letters in week 22. For individuals with letters in week 15 we will run three different regressions: i) Excluding those with a mandatory meeting in week 19; ii) Excluding those with a mandatory meeting in week 26; and iii) Including both groups with mandatory meetings but with separate dummy variables.

We will also include all seven treatments in the same regression. To test for the independent effect of receiving voluntary letters and the timing effects of mandatory letters we will run regressions separately for people with voluntary and mandatory letters. If there is an independent reversed notification effect we expect that people getting voluntary letters early will have longer absence than people getting voluntary letters later.

To explore heterogeneity we will first interact the treatment variable(s) with the baseline control variables. For ease of interpretation we dummy code the Grade variable into 1 for people with any amount of graded sickness absence and zero for those with 100 percent sick leave. We call this variable Graded. We retain the dummy coding for female and transform the continuous variables so that they are in standardized form. We will further
split the sample by gender and into symptoms or diagnoses (i.e. symptoms=0) and we will run separate regressions for each main diagnosis (as defined by the letter in the ICPC-2 or a dummy for Multiple diagnoses in cases where the individual has several diagnoses). The sample will be split by caseworker predictions (long spell and importance of meeting), by offices as well as by occupations. We also apply machine learning techniques to search for heterogenous treatment effects more systematically (see Section 5.3).

5.1.1 Identification of threat effects

To identify the different types of threat effects we will use the dummy variables for having returned before week X (13, 19, 22) and regress this on the different sets of dummy variables for the treatments. To start with, we will run the following linear probability model:

\[
\text{Return before week 13}_i = \beta \text{ DM 13 letter } 9_i + \gamma \text{ Letter } 9_i + \theta \text{ Block }_{cw_i} + \lambda X_i + \epsilon_i \quad (\text{eq. 2}),
\]

where \(\text{Return before week 13}_i\) is a dummy for returning before week 13, \(\text{ DM 13 letter } 9_i\) equals 1 if the individuals are randomly assigned to a mandatory meeting in week 13 with a notification letter in week 9, and \(\text{ Letter } 9_i\) is a dummy for being randomly assigned to a notification of a voluntary meeting in week 9. In this specification, the ones receiving the early letters are compared to everyone not having received the letter yet. This allows for a direct test of the threat effect as well as the reversed threat effect. More specifically, if \(\beta\) is positive, there is a threat effect whereby calling people in to a mandatory meeting induces them to end their spells faster without a meeting having taken place. Similarly, if \(\gamma\) is negative, there is a reversed threat effect whereby telling people that a meeting is voluntary for them reduces the probability that they end their sickness spell in the following weeks.

The same specification can also be run with the dependent variable being \(\text{Return before week 19}_i\) and including dummies for \(\text{ DM 19 letter } 15_i\), \(\text{ DM 26 letter } 15_i\), and \(\text{ Letter } 15_i\). The coefficient for the early letter with meeting in week 13 will then not be interpretable as a threat effect as people will have started to have meetings at this point. Rather, the
coefficient can be interpreted as a combination of the effect of being assigned to a meeting (threat effect) and a short run early meeting effect. We also run the same regression with the outcome variable being return before week 22. The excluded group still consists of people not having received any letter but as meetings have started also for the DM 19 letter 15 group, only the DM 26 letter 15 coefficient can be interpreted as a threat effect. The coefficient for DM 19 letter 15 can now be interpreted as a combination of a letter effect and a short run effect of having a meeting in week 19. The coefficient for DM 13 letter 9 can now be interpreted as a total effect of early letter and a medium term effect of meetings (as 9 weeks have passed since the intended dates of the meetings). We will investigate heterogeneity in threat effects by interacting the treatment variables with baseline covariates.

5.2 Duration analysis

We will conduct duration analyzes to investigate transition rates in the weeks the letters are received and in the weeks leading up to the meeting. These analyzes will allow us to test even finer predictions about the different types of letters. These methods will also shed light on the separate effects of the conducted meetings and the effects of the letters.

In the duration analysis 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. The 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.

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. While we have no way of checking when the letters are sent, we can and will investigate when the meetings are held.\footnote{In the accompanying draft we present some descriptive statistics on meeting dates for our different treatments.}

We estimate 5 different models:

1) 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 variables 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 a meeting is voluntary (the control group) we interact the variable with a dummy for the treatment group.

2) 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.

3) In the third model, we separate between three types of letters and investigate 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.

4) We then 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 mandatory meeting (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.

5) Finally we want to investigate whether being called in to meetings have the largest effect early or late in the spell. 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).

We will also estimate these models for different sub-groups.

5.3 Uncovering heterogeneity using machine learning techniques

We also use machine learning techniques to automate the search for heterogenous treatment effects. There are many different types of machine learning algorithms and we have decided to use classification and regression trees (R package causalTree, (Athey and Imbens, 2016)); and random forests (R package grf, (Wager and Athey, 2017)). As this field is moving rapidly, however, it is possible that there will be other techniques that are relevant
for us once we start analyzing the data.

Both techniques start by splitting the sample in two partitions. We will split our analysis sample (i.e. early test sample + late sample) in two random subsets of equal size. We stratify the random draw using groups in which individuals have constant treatment probabilities and we set the seed=1987. We name the two partitions $S1$ (for the machine learning training sample) and $S2$ (for the machine learning test sample).

Our main analysis will use $Total$ as outcome variable, and $DM$ as treatment. We will also test different treatments, and in particular heterogeneity in the threat and reversed threat effects, and whenever this includes different samples we will draw $S1$ and $S2$ again (e.g. if we investigate heterogeneity in the effects for those treated with early versus late letters in the group receiving mandatory meetings). We will include as covariates all the variables mentioned in the paragraph Other variables in section 4, and also a dummy for each separate diagnosis. Continuous covariates are standardized.

One of the advantages of machine learning techniques is that they are able to handle a very large number of covariates, even more covariates than the number of observations. For that reason we will also include in the set of covariates any other observables that we will be able to collect (for example by merging our data with register data) and that we deem potentially interesting. All analyses will, however, also include the covariate set explicitly specified here.

5.3.1 Honest causal tree

We will estimate conditional average treatment effects (CATEs) using the honest causal tree by Athey and Imbens (2016), with the R package causalTree. We will proceed as follows:

1) We grow the tree using only $S1$ with the R function causalTree with options split.Rule = "CT", split.Honest = TRUE, cv.option = "CT", cv.Honest = TRUE. We take into account that we have different treatment probabilities by passing to the routine the ex-ante
probability of being treated in different strata. We want to grow a tree with a manageable number of leaves, so we set the minimum number of treated and control observations in each leaf that is split (minsize) to be approximately equal to 5% of the sample size in S1. We select the number of cross validations equal to 100 (xval).\footnote{As recommended by the authors we also use the option split.Bucket = TRUE with parameters bucketNum = minsize*3, bucketMax = 20.}

2) We prune the tree (cut back leaves) by selecting the complexity parameter corresponding to the minimum normalized cross validation error (xerror). Note that each leaf of the pruned tree is defined by the same values of given covariates (for example women born after 1978.) The routine calculates the treatment effects in each terminal leaf. We name this tree dishonest tree.

3) To make the procedure what Athey and Imbens (2016) call honest we use one sample for growing the tree and another sample for testing the predictions. That is, we classify each observation in S2 as belonging to one of the terminal leaves defined by the tree grown at point 2). Using only S2, we estimate the average treatment effect in each leaf, which is called the conditional average treatment effect (CATE, where conditional refers to conditional on being in a given leaf). We name this tree honest tree.

4) Using only S2, we regress (OLS) the outcome on dummies for each terminal leaf, their interactions with the treatment and block fixed effects (without constant nor treatment included). We use robust standard errors. This regression recovers the same CATEs as in the previous point, but with standard errors.

5) We consider evidence of treatment heterogeneity if we find statistically significant treatment effects in one or more leaves in S2, and if in these leaves the CATE has the same sign in both the honest and in the dishonest tree.
5.3.2 Honest causal forest

We will estimate conditional average treatment effects (CATEs) using an honest causal forest by (Wager and Athey, 2017) with the package \texttt{grf}. We proceed as follows:

1) Using the R function \texttt{causal forest} we grow the forest (ensemble of trees) using only $S_1$. Each tree uses only a random fraction (50%) of $S_1$ ($\texttt{sample.fraction} = 1/2$). This sub-sample will be further cut in half at random to implement honest estimation (grow tree in one half, and estimate CATE in the other half). The estimated CATEs for each leaf and each tree are stored. We set the following parameters:

- Number of trees: at least 25000 (and if feasible 50000).
- Number of covariates used in each tree: half the total number of available covariates.
- Penalty parameter $\lambda=0$, and \texttt{downweight.penalty=FALSE}
- Number of trees for each subsample (to construct confidence intervals): 50 if computation time allows (based on simulation it should take around 30 minutes). Otherwise we reduce it until feasible (minimum is 2).
- \texttt{precompute.nuisance = FALSE}
- $\texttt{min.node.size} = 4$
- $\texttt{alpha} = 0.05$
- $\texttt{seed} = 1987$
- $\texttt{sample.fraction} = 1/2$

2) Using the R function \texttt{predict} we predict the CATEs in $S_2$ based on the estimates from point 1). We report a histogram, summary statistics and a variance plot of these CATEs.
We also report the fraction of observations that have a CATE significantly different from 0 at the 10% level.

3) To assess whether the forest detected true heterogeneity, we create a dummy for observations in $S2$ that have a positive CATE, and one for those that have a negative one (obtained in point 2)). We regress the outcome on these dummies, their interactions with the treatment, and block fixed effects. If the treatment heterogeneity is not spurious, this regression will estimate a positive treatment effect in the group of observations with positive CATEs, and negative treatment effects for the other group. We will also test if the two are significantly different from each other.

4) We will do the same as in the previous point but splitting observations in $S2$ at the median of CATE instead of at CATE=0, as well as in quartiles of CATE.

5) If evidence from the previous two points suggests that the treatment heterogeneity is not spurious, we will characterize observations with positive and negative CATEs and in different quartiles of CATEs inspecting summary statistics of their covariates in these different groups.

6 Main results in the early training sample

In the companion document ”Effects of Dialogue Meetings - results from the early training sample” we present results from the specifications described here as well as some additional analyzes.

The main results from the early training sample are that the effect of mandatory as opposed to voluntary meetings seems to be negative but the effect is not statistically significant. There seems to be substantial effect heterogeneity with respect to the caseworkers’ predictions of long meetings whereby those that are predicted to be away for a long period have a much larger effect of being called in to a mandatory meeting. Interestingly, we find less evidence of heterogeneity when using the caseworkers predictions for the effects of the
meetings.

The duration analysis suggests that there is a threat effect such that individuals are more likely to end the sick leave spell in the week they get the letter and in the week before the meeting. There also seems to be a reversed threat effect whereby individuals called in to a voluntary meeting are less likely to end their sick-leave spell in the week they get the letter. We also observe this reversed threat effect when running a regression of returning before week 13 on receiving a letter about a voluntary meeting in week 9.

The causal tree analysis suggests that there is heterogeneity with respect to the caseworker predictions of long spells but in particular for individuals with a higher share of full-time absence (as opposed to part time absentees). The results from the causal forest are inconclusive at this point but we expect this method to be more powerful once we receive more data.

7 Hypotheses to test in the test sample

When we get outcome data and covariates for the early test sample we will run all the analyzes for this sample that we ran in creating the companion document ”Effects of Dialogue Meetings - results from the early training sample”. The machine learning algorithms will use the early training sample as $S_1$ and the early test sample as $S_2$.

We will then compare the results and draw more reliable conclusions about the different effects. The conclusions we draw will be based on the full picture of the results but if we find the same results as those described in Section 6 with respect to effect heterogeneity by caseworker predictions of long spells, threat effects and reversed threat effects, we will suggest some policy changes.

The proposed changes in policy will be made in close cooperation with NAV and they have the final word on any change. We expect two relatively easily implementable changes to the design already. If the result of a reversed threat effect of sending out letters of voluntary
meetings hold up, a suggested policy change is to send these types of letters out later (if they have to be sent at all). If the results that the meetings have larger effects for people with long predicted spells hold up, these individuals can be targeted to have a higher probability of mandatory meetings.

Irrespective of the results in the early test sample we are also having discussions about other types of policy changes. We are discussing changes to the letters to see if the nudging aspect can be improved. We are also discussing ways to test different types of meetings by conducting some targeted experiments. For instance, do meetings work better when the physician attends or not? We are currently surveying the caseworkers and conducting interviews with physicians in order to derive more focused hypotheses here.

If a change is implemented we will register an additional plan detailing how that change will be tested and we will document how the change affects our analysis sample.

8 Hypotheses to test in the analysis sample

When we decide that the current phase of the experiment is over we will have NAV add data on outcomes and covariates for the full analysis sample. On that sample we will conduct all the analyzes specified in this plan. Other analyzes not specified here nor in later registrations will be seen as more explorative. As there are many hypotheses and tests in this pre-registration we here outline what we view as our main hypotheses.

Hypothesis 1: Mandatory meetings will cause a shorter total absence spell than voluntary meetings. Support of this hypothesis would be a statistically significant negative coefficient for $DM_i$ in the OLS regression of $Total$, controlling for the specified baseline covariates and block fixed effects (equation 1).

Hypothesis 2: There is a threat effect of letters calling in to mandatory meetings. This can be tested in different ways. As we saw in the duration analysis individuals randomly assigned to mandatory meetings may be more likely to end their spell in both the letter
week and in the week before meeting. We will, however, use equation 2 and and view a statistically significant positive coefficient for DM 13 letter 9\textsubscript{i} as supporting the hypothesis.

**Hypothesis 3:** There is a reversed threat effect of voluntary letters. Also this hypothesis can be tested in different ways and we saw indications in the duration analysis of a reversed threat effect by people being less likely to end their spell in the week the letter arrives if the letter called in to a voluntary meeting. Again, however, we will use use equation 2 and and view a statistically significant negative coefficient for Letter\textsubscript{9\textsubscript{i}} as supporting the hypothesis.

The remaining hypotheses in this plan can be seen as secondary hypotheses. With respect to the heterogeneity in treatment effects we expect that the machine learning techniques will give a clear and consistent answer. These techniques are in some sense self validating as they are based on sample splits. We will nonetheless always include the covariates specified here.

### 9 Power analysis

We have calculated minimum detectable effects (MDE) under different assumptions. On February 9 2018 we have a total of 11,196 individuals entered into the system. We have already used 1,871 for the early training sample. A lower bound for the analysis number of observations in the analysis sample is therefore 9,325 before dropping any individual. In the early training sample we had to drop 13 percent of the observations since they had returned from their sick-leave spell already at the date of the draft. If we assume a 15 percent drop rate in the analysis sample and round downwards we will have a sample of 7,900.

The minimum detectable effect for a sample of 7,900 with a significance level of 0.05 is 0.063. That is, we would be able to find effects as small as 0.063 standard deviations of the dependent variable. As seen in Table 1 the standard deviation for Total is 90 days. Hence, we would be able to identify effects as small as 5.67 total days. We further know that the R-squared in a regression including DM and the baseline covariates is 0.22. The corresponding MDE if we take this into account is 0.056 (which corresponds to 5.04 total
As we are testing multiple hypotheses we will also adjust the p-values. We have already anticipated the adjustment and therefore we have restricted the number of main hypotheses to test to be only three. To account for having three different hypotheses we follow the recommendations of Fink, McConnell, and Vollmer (2014) and use a method developed by Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) to minimize the false non-discovery rate (see also Almeida (2012) and Finseraas and Kotsadam (2017) for pre-analysis plans with the same decision rules for correction of p-values). The main advantage of the method is that it is limiting the risk of false discoveries while only adjusting the critical values based on other true hypotheses. The false discovery rate method developed by Benjamini and Hochberg (1995) implies that the m p-values of the i hypotheses are ordered from low to high and that the critical value of the p-value is then $p(i) = \alpha \cdot \frac{i}{m}$. To take a concrete example, with 3 hypotheses and a significance level ($\alpha$) of 0.05, the critical p-value would be 0.017 for the one with the lowest p-value (0.05* 1/3, which is the same as a Bonferroni correction). For the second and third hypotheses, the critical p-values would be 0.033 (0.05*2/3) and 0.05 (0.05*3/3). With a p-value of 0.017 our calculated MDE is 0.074 (0.064 accounting for an R-squared on 0.22). We expect the sample size to be larger but we conclude that even with the lowest bound of the sample size, the experiment is well powered to detect small effects.

10 Archive

The pre-analysis plan is archived before the early test sample is matched with any outcome variables. We archive it at the registry for randomized controlled trials in economics held by The American Economic Association: https://www.socialscienceregistry.org/ on February 12 2018. We will receive the matched early test sample on February 13. The experiment is still running and it is unclear when the analysis sample will be finished. We
will test the results from the early training sample in the early test sample and we may file a separate analysis plan with more registered decisions taken after this point.
References


