Journal of Population Economics (2023) 36:935–971
https://doi.org/10.1007/s00148-021-00841-6

ORIGINAL PAPER

Skipping the doctor: evidence from a case with extended self-certification of paid sick leave

Bruno Ferman[1] · **Gaute Torsvik**[2] · **Kjell Vaage**[3]

Received: 8 September 2019 / Accepted: 15 March 2021 / Published online: 13 July 2021

Abstract

This paper examines the impact of a policy reform in a municipality in Norway that extended to workers the right to self-certify sickness absence from work. After the reform, workers were no longer obliged to obtain a certificate from a physician to receive sickness benefits. They could call in sick directly to their line leader and had to engage in a counselling program organized by the employer. To estimate the effect of this reform, we contrast the change in sickness absence among employees who were granted the extended right to self-certify absence with absence among employees who had to obtain a physician's certificate to be entitled to sickness benefits. We use both a standard difference-in-differences method and the synthetic control method to estimate the effect of the reform. We can rule out large positive effects on absence after the reform, with strong evidence that the policy change actually resulted in a reduction in absence for female workers.

Keywords Welfare transfers · Sickness absence · Moral hazard · Gatekeeping

JEL Classification H55 · I13 · C23 · J24

Responsible editor: Shuaizhang Feng

✉ Kjell Vaage
kjell.vaage@uib.no

Bruno Ferman
bruno.ferman@fgv.br

Gaute Torsvik
gaute.torsvik@econ.uio.no

[1] Sao Paulo School of Economics-FGV, Sao Paulo, Brazil

[2] Department of Economics, University of Oslo, Oslo, Norway

[3] Department of Economics, University of Bergen, Bergen, Norway

1 Introduction

Paid sick leave is an important insurance, allowing workers to smooth consumption over transitory negative health shocks. However, sickness benefit programs can, like any other insurance, be misused; employees who are fit to attend work may call in sick or may request benefits for longer periods than their health status calls for (Henrekson and Persson 2004; Hesselius et al. 2013; Dionne and St-Michel 1991). To reduce moral hazard, both welfare states and private insurers may require a medical certificate from a physician to verify sickness (OECD 2010). Physicians are used as gatekeepers to prevent illegitimate claims of paid sick leave. This is a costly practice, and it is not clear how well it works.

This paper looks at the effect of a reform that removed the physician as a sickness certifier. In 2008, one municipality in Norway allowed all workers employed in the municipal sector to self-declare health-related absence for a whole year, which is the maximum entitlement period for temporary sickness benefits in Norway. The municipal workers were free to self-report sickness absence, but had to report regularly about their health and work capacity to their line leader. In other municipalities, and in the reform municipality prior to the change, the rule was that workers could self-declare periods of absence shorter than 9 days. For longer periods, they needed a medical certificate to obtain benefits. Normally, physicians also play a role in dialogue meetings (counselling) between the employer and the worker who is on sick leave. The reform naturally relegated the physician also from that scene. Instead, the employer took a more direct and active role in the counselling of sick-listed workers.

If workers' demand for sickness absence were unaffected by the reform, removing the requirement of a medical certificate would increase absenteeism. By how much depends on how strict physicians are as gatekeepers. However, the reform may reduce workers' demand for absence. Workers' demand for sickness absence depends on their health, on the replacement rate of the benefits and on the utility difference between staying home and attending work. Allowing workers to self-certify absence will probably not directly affect the health of the workers, nor will it change the replacement rate. The reform can, however, influence the non-pecuniary costs of calling in sick.

The reform that we study here was announced (by the municipality) as the "Trust Project." Recent research in behavioral economics has shown that many individuals want to return (reciprocate) the treatment they receive from others. For example, employers who show distrust towards their employees by imposing excessive control mechanisms may induce misbehavior (Ellingsen and Johannesson 2008; Falk and Kosfeld 2006). According to this logic, allowing self-certification of absence can foster greater loyalty and motivation among employees, which in turn may lower the demand for paid sick leave. In addition, frequent meetings with the employer (the line leader) while sick, without having the physician as a counselor, may also reduce the intrinsic utility (increase the hassle) of asking for a sick leave.

Our data contains registered sickness absence for all workers in Norway during 2001 to 2014. To assess the impact of the reform on sickness absence, we compare absence among municipal workers in the reform municipality (Mandal) before and after the reform with the change in absence among municipal employees working in all other Norwegian municipalities. That is, we apply difference-in-differences (DD) logic, with Mandal as the treated unit and the other municipalities as controls, to assess the effect of the reform.

We also consider the same model for employees who work in the private sector or for the central government as a placebo exercise. This helps us rule out the possibility of other contemporary shocks that may have affected Mandal. In addition, we use the synthetic control method to construct a control that better resembles the treated unit (Abadie and Gardeazabal 2003; Abadie et al. 2010).

We can rule out large positive effects on absence in Mandal in the post-reform period, with strong evidence that the policy change actually resulted in a reduction in absence for female workers. One of the stylized facts regarding sickness absence is that in almost every country, women tend to have higher absence levels than men (Mastekaasa and Melsom 2014). It is interesting that extending self-certification of sick leave reduces the gender gap in absence. Since the female share of the labor force is around 80% in Mandal's municipal sector (which is in line with the rest of the municipal sector), the effect of the reform is far stronger than if the same change in behavior had occurred among the male workers. We also find that the number of spells decreased, while the average length of the remaining spells increased. This suggests that the reform had a larger impact in preventing relatively shorter absence spells. Our main finding stands in stark contrast to the result from a Swedish experiment where a random sample of workers was granted 1 week of extra self-certification of sickness absence (Hesselius et al. 2009). We discuss potential reasons for these opposing results at the end of the paper.

Although the main ingredient of the Mandal reform is that municipal workers are granted the right to self-certify absence, it also contains other elements, such as a stronger involvement in the sickness absence counselling by the employer. We do not have data to disentangle the importance of the different elements of the reform, so our results should be seen as the aggregate effects of all these elements.

The main contribution of this paper is to use quasi-experimental data to estimate the effect of a policy that, at its core, excludes a costly procedure for constraining moral hazard in paid sick leave. Back-of-the-envelope calculations suggest that Norwegian primary physicians spend as much as 10–15% of their working time on sickness absence certification.[1] Our finding that sickness certification can be taken out of the hands of the physicians without a subsequent rise in sickness absence should be of considerably policy relevance.

[1] In 2015, 3.8 million sickness absence certificates were issued in Norway. There are around 2500 primary physicians and they issue the bulk of sickness certificates. Assume that 3 million of the certificates are issued by primary physicians. A conservative estimate is that engaging in dialogues and doing the paperwork associated with issuing a certificate takes 10 min. This means that on average a primary physician spends 200 h/y on the administrative work associated with sickness certificates, which adds up to over 10% of a normal work year.

The paper unfolds as follows. "Section 2" provides a brief introduction to the sickness insurance system in Norway and a description of the reform. "Section 3" presents a conceptual framework for analyzing the relation between medical health certificates, gatekeeping and sickness absence, and a discussion of how our paper relates to relevant literature. "Section 4" describes our data, while Section 5 presents our empirical setup. The results are reported in Section 6, along with some suggestions to their channels in Section 7. Section 8 discusses our findings and concludes the paper.

2 Institutional setting and the policy reform

2.1 Sickness benefits and sickness absence in Norway

Sickness insurance is mandatory in Norway and covers all workers employed for more than 4 weeks. The wage compensation ratio is 100% from day one for a maximum period of 1 year.[2] The employer pays sickness benefits for the first 16 days; thereafter, the benefits are financed by the National Insurance Administration (NAV) for a maximum of 50 weeks. Municipal workers do not need a medical certificate for sickness spells lasting less than 9 days. Periods of 9 days or longer require a medical certificate, usually from a general practitioner, and for more than 8 weeks an expanded certificate is required.

The level of sickness absence is high in Norway, around 7% of contracted work hours are lost because of sickness absence (certified by a medical doctor). Around 80% of the total absence days comes from periods remunerated by NAV (lasting more than 16 days), which we define as long-term absence in the present paper. The public expenditures associated with sickness absence are in the order of 2.5% of GDP. Individuals who obtain long-term absence certificates have a high risk of never returning to ordinary work and be transferred to permanent benefits (Markussen et al. 2012).

Short-term absence in Norway is remarkably stable over time and across individual characteristics. Long-term absence, on the other hand, varies substantially over the business cycle (Askildsen et al. 2005) and across gender, age, education, and occupation (Mastekaasa 2015). The majority of sick leave certificates from doctors classify the health issues as diffuse and subjective health problems; mental disorders and muscle-skeletal symptoms accounted for about 60% of the cases in 2012. As the term "diffuse diagnoses" suggests, these are cases that cannot be objectively verified by the physician and it is difficult to prescribe evidence-based treatment. Diffuse diagnoses dominate in the long-term spells. Diagnoses that are easily verifiable, e.g., cancer and cardiovascular diseases, play only a limited role. Cardiovascular diseases, for example, accounts for only

[2] For some groups of employees there is an earning ceiling of approximately NOK 600,000/EUR 56,000 per year (2020), but most workers (all in the public sector and the majority in the private sector) obtain 100% replacement of their salaries.

5% of the absence days. Short-term absence also contains diagnoses that are difficult to verify (chronic pain, etc.). But for shorter spells, uncomplicated and observable diagnoses (airways infections, etc.) makes up a larger share than for the long-term spells.

The difference between short- and long-term absence suggests that it is long-term sickness absence that will be most influenced by the sick-listed's own judgments; in particular when it comes to length of spells (Mastekaasa 2015). Hence, moral hazard is likely to be most relevant for long-term absence.

A noticeable pattern in the sickness absence in most developed countries is that women have considerably and persistently higher absence rates than men (Avdic and Johansson 2017). As for Norway, the average number of sick days per year is now 60–70% higher for female than for male workers. Pregnancy and other biological differences explain only some of the gap, and do not offer explanations to the *increasing* gender difference in absence over the last 50 years. Before the 1980s, the rate of sickness absence was more or less equal for men and women. To explain the increasing gender gap, most research therefore look to the women's advent of the labor market and the corresponding change from single to dual earner families. Empirical evidence from Norway is rather inconclusive, however. Based on EU Labor Force Surveys from 1983 to 2011, Mastekaasa (2014) finds no support for increasing representation in the workforce of mothers of small children. Based on administrative register data, the "double burden" hypothesis (women have the main responsibility for the household production also when they work in labor market) is tested but rejected in Cools et al. (2017).[3] Furthermore, Mastekaasa (2014) finds no support for occupational segregation (women work in high absence occupations) as explanation of the increasing difference, while he finds some support for changing composition of the female labor force (more women with health problems or with lower job motivation).

Finally, the observed gap in sickness absence may be explained by gender differences in health-related behavior, preferences and norms. Women may be more concerned about health and/or be less devoted to their job and career and therefore have lower threshold for reporting sick. Alternatively, they may be more susceptible for influence from local absence culture. Even with the rich Norwegian register data, hypotheses like these are notoriously hard to test. An attempt is found in Hauge et al. (2015), who combine survey and register data from the city of Oslo. They do find gender differences in relevant attitudes, norms and preferences, but not of a size that manage to explain the huge differences in sickness absence.

Summing up, previous research has given several explanations to aspects and elements which *do not explain* the persistent and even increasing gender

[3] On the other hand, Angelov and Johansson (2020) find, based on Swedish register data, a substantial effect on parenthood on the within-couple gender gap in sickness absence for couples who become parents.

gap in the sickness absence. Knowledge about the actual causes is still lacking, however. Nevertheless, it is important to determine whether men and women respond differently to the Mandal reform. Approximately 80% of the employees in the municipal sector are women, and the gender gap in the sickness absence is as large here as in other sectors. Hence, from the view of the municipality as an employer, the degree of success depends critically on female response to the reform.

2.2 The reform: extended self-certification of sickness absence in Mandal

In 2014, there were 428 municipalities in Norway. They are all responsible for producing the same services: compulsory education (until the 10th grade), outpatient health services, senior citizen services, and maintenance of the road infrastructure within a municipality. In 2012, 23% of the total workforce was employed by municipalities. The vast majority (about 75%) of the municipal workers are women. Although they all serve the same functions, municipalities vary widely in size. The smallest has fewer than 300 inhabitants and the largest more than 600,000. In 2012, Mandal—the reform municipality—had 15,000 inhabitants and 1200 employees (around 900 workers in full time positions), which is slightly above the average municipality size in Norway.

Historically, the level of sickness absence for municipal employees in Mandal has been around the average for the sector in Norway. During the last decade, several municipalities—and firms more generally—have experimented with various local reforms to reduce sickness absence. This is also the case for the municipality of Mandal; at the end of 2003, it launched the so-called "presence project" to reduce sickness absence among municipal workers. From this project grew an initiative directed to the Ministry of Labour, requesting permission to "bypass" the physician as a sickness absence certifier. The suggestion was to allow municipal employees to self-certify their sickness absence for the entire benefit period (1 year).

Regarding the reform we consider here, the municipal administration predicted that the employees would respond positively to extended trust and counselling in relation to sickness absence. The administrative leadership in Mandal worked out a detailed plan for how lower level leaders should follow-up workers who self-certified sickness absence. The idea was that a stronger involvement from the employees' line managers would substitute for the physician's involvement and advice. For shorter spells, leaders were instructed to call the absentees (after 3 days and after 8 days). For longer periods, the leaders were instructed to initiate a number of different meetings for individual counselling and follow-up plans, and to also regularly contact the absentee, and send cards, and flowers etc. A hierarchical system of email-based action reminders among the leaders guaranteed that the follow-up plan was implemented.

The application of the system with extended self-certification of sickness absence was approved by the Ministry in June 2007. The "Trust Project" (with

a handshake as the official logo) was officially launched on July 1, 2007. With this, Mandal became the only municipality—and firm—in Norway with a permission to operate with a sickness insurance scheme that made the medical certificate from a physician optional for the full length of the sickness period. After some months of piloting a web-based system of self-reported absence was in place in May 2008. At the end of 2008, almost 90% of all sickness absence was self-reported.

3 Demand for health-related work absence

Sickness absence insurance allows workers to be absent from work and receive benefits in periods when their health temporarily drops under the level that is required for them to perform their work tasks. Health is a multidimensional and complex entity, but for our purpose here it can be represented by a scalar h, with higher h indicating better health. The implicit sickness absence insurance agreement is that if h drops below h^0, the worker is unable to do his or her job and it is legitimate to call in sick.

Workers request (demand) for sickness absence depends on several factors: (i) the health condition of the worker, (ii) the replacement rate of the sickness benefit scheme, (iii) the costs associated with obtaining a permit to be absent and receive benefits, and (iv) the non-pecuniary utility difference between being sick absent at home and being at work.

In this context, moral hazard is the potential problem that workers, who are fit to work ($h > h^0$), demand sick absence. The standard way to constrain this problem is to have a system where workers must get a medical certificate from a physician in order to get paid sick leave. The idea is that doctors will screen those who request sick leave and deny a medical certificate to those who are healthy enough to do their job. This is optimistic. It is often difficult to diagnose a patient, and the health status that separates legitimate sickness absence from illegitimate absence is open for interpretation. In addition, it is not obvious that doctors will act as gatekeepers to welfare benefits. Many physicians consider themselves as their patients' advocate; requests for sick leave certificates might then be difficult to deny (Svärdsudd and Englund 2000; Carlsen et al. 2020; Markussen et al. 2013). Their own economic interests may also weaken the role of general physicians (GPs) as gatekeepers, as they may lose patients if they decline requests for a sickness absence certificate.

We consider what might happen to overall sickness absence when a system with doctor certification is replaced by a system where workers can self-certify their absences. In this case, workers would instead have to enter into a counselling relationship with their line leader at the workplace.

If the reform leaves the demand for absence unchanged, there will be an increase in absenteeism after self-certification is introduced. By how much depends on the magnitude of the moral hazard problem, and by how lenient physicians are as gatekeepers. There are, however, good reasons to expect that the reform will change the demand for absence. Self-certification means that

workers can skip the trip to the doctor and avoid the costs associated with obtaining a sick leave certificate. This change would increase demand for sick leave. Several other aspects of the self-certification reform, however, may reduce the demand for sick leave.

An employer-initiated reform that allows workers to self-declare health issues that reduce their work capacity signifies both generosity and trust. Indeed, the reform in Mandal was branded as "The Trust Project." Workers may therefore feel more guilt if they call in sick, or they may have a higher intrinsic utility of attending work, after having been granted permission to self-declare sickness absence.[4] Another reason why the reform may reduce demand for sickness absence is that the arrangement implies frequent and direct consultations with the employer. In these meetings, the physician is no longer the mediator between the employer and the employee in dialogue meetings where the absentee, the absence certifier and the employer discuss adjustment that could be carried out at the workplace to make full or partial work resumption possible. With no certifier, there is only a direct dialogue between the absentee and the employer. Direct counselling and activation, not having the medical doctor as the patients' advocate, may also reduce the intrinsic utility of staying home with a sick leave.

A third reason for lower absence is lost legitimacy. Workers with diffuse symptoms may themselves be uncertain whether they are unfit for work or not. Some of these workers will probably feel that a medical certificate relieves them from some of the guilt and remorse that comes with calling in sick, given their health status. In a regime with self-certification of absence, the legitimacy of the medical certificate disappears and this may lower demand for absence.

Lower demand for absence will affect both the extensive and the intensive margin, both the number of spells that are realized and the length of those that are realized. It is likely that a negative shift in demand will reduce the length of the sick absence spells that are realized. Lower demand will also relegate spells for which the net utility of demanding sick absence leave was just above zero. If these marginal spells are among the shorter spells, the effect on the extensive margin implies that we should observe longer average spells when demand for sick absence declines.

Theoretically it is, therefore, ambiguous how a reform granting workers the right to self-certify sickness absence will affect the absence level. For a given demand for sickness absence, skipping the physician as a sickness certifier will increase absence, but, as we have argued, the demand for sick leave may fall because of extended self-certification and employer involvement. It is also unclear what we should expect with respect to the length of the spells.

[4] There is now an extensive theoretical and experimental literature on reciprocal motivation and how generosity and trust affect behavior. Some prominent examples are Ellingsen and Johannesson (2008); Bénabou and Tirole (2006); Falk and Kosfeld (2006).

4 Data

Our unit of observation is yearly sickness absence at the municipality level, broken down by sector (municipal employees and others (private and central government)), by gender, and by four age intervals, $[16-39]$, $[40-49]$, $[50-59]$, and $[60-69]$. We have data from 2001 until 2014, implying 7 years of observations before and 7 years of observations after the reform.

Sickness absence rate (percent) is defined as the number of aggregated sick days (for the respective sector and gender) divided by the corresponding sum of contractual workdays. The latter is defined as the number of days a person has agreed to work for his employer in a given period, adjusted for fraction of employment, weekends and public holidays.[5] (This is the common way of reporting sickness absence by Statistics Norway.) Every sick leave is counted separately and included in the aggregate measures whether they are multiple leaves from the same individual or not.

Our analysis is based on absence spells that are longer than 16 days. We use long spells primarily because of data reliability. The data are obtained from the National Insurance Administration (NAV). Employers (municipalities in our case) pay for absence spells that are shorter than 17 days. The government takes over the financial responsibility after 16 days. To be reimbursed for absence benefits that extend 16 days, employers must report all the absence spells that last longer than this to NAV. Hence, it is in the economic self-interest of the employer to report long-term sickness absence to NAV. For short-term periods only absence certified by a medical doctor is reported to NAV. Hence, if we were to use data on short-term spells that are reported to NAV, absence in Mandal would drop mechanically, simply because only physician certified absence is recorded in the NAV data.

We do not consider our focus on long-term spells to be a serious limitation of our study. As explained above, but perhaps contrary to the conventional wisdom, moral hazard problems appear to be especially relevant for long-term absence. In addition, since we are concerned with the number of working days that are lost because of illness, long-term spells dominate. Among municipal employees, around 80% of the contracted workdays that are lost because of sickness absence come from periods that extend beyond 16 days. Furthermore, all municipal workers in Mandal, and in other municipalities in Norway, could already self-certify spells shorter than 9 days before 2007. Hence, to the extent that Mandal reform had an effect on short-term spells, it could only apply for spells of a duration in the interval of 9–16 days.

We present descriptive statistics from Mandal and other municipalities in Appendix Table 3.

[5] It means that, say, lower absence rate figures can come from a drop in the use of sick leave, but also from a relative increase in the denominator. For ease of transparency, we include contractual workdays separately in Table 1 and Appendix Table 3.

5 Empirical setup

We rely on a difference-in-differences (DD) design to estimate the effects of the reform, where we contrast measures of absence between Mandal and other municipalities before and after the reform. More specifically, we construct a balanced municipality-year panel dataset with all municipalities in Norway, and estimate the following equation[6]:

$$y_{it} = \alpha_i + \beta_t + \delta DD_{it} + \varepsilon_{it}, \tag{1}$$

where y_{it} is an outcome variable for municipality i in year t, α_i are municipality fixed effect, and β_t are year fixed effects. The variable DD_{it} is an indicator variable equal to one for the treated municipality Mandal in the post-reform period, and 0 otherwise. Note that the indicator variables for the post-reform period and for the treated municipality are absorbed by the fixed effects. In our main specifications, the outcome variable is the percentage of working days lost because of health-related absence for municipal workers. We also consider the effect of the reform on the length of absence spells and the use of graded sickness absence (where the workers are partly at work).

The identifying assumption in this DD framework is that the average absence among all municipal workers in Norway (except Mandal) and in the treated municipality (Mandal) would have followed parallel trends in the absence of the reform. As we discuss in details in "Section 6.1.1," such assumption seems plausible when we consider 2004–2007 as pre-treatment periods. Therefore, we focus on estimation of Eq. (1) using data from 2004 to 2014.

As a robustness check, we also estimate the same model for employees who work in the private sector or for the central government. Since those workers were not directly affected by the reform, we should not expect to find significant effects for them if the identification assumption is valid. Therefore, considering the effects on those workers is informative about potential contemporaneous Mandal shocks to sickness absence that could invalidate our main results. We also consider an alternative DD specification in which we compare municipal and non-municipal workers within Mandal.

Finally, we also estimate the effect of the reform with a more tailored-fit control group of municipalities, using the synthetic control method, SCM, developed by Abadie and Gardeazabal (2003) and Abadie et al. (2010). The essence of this method is to use the pre-reform period to construct a synthetic control unit ("Synthetic Mandal")—a convex combination of potential control municipalities—that resembles the treated unit along the dimensions that are important predictors for sickness absence in the post-reform period. Intuitively, the idea of the method is to construct a comparison unit that is affected by potential common shocks in the same way as the treated unit, relaxing the assumption on parallel trends (see Abadie et al. 2010).

[6] We restrict to municipalities in which data is available for all periods, so that we can use the inference method proposed by Ferman and Pinto (2021). We discard around 4% of the municipalities with incomplete data. Point estimates remain similar if we include all municipalities.

6 Results

6.1 Difference–in–differences analysis

6.1.1 Validation of the DD assumptions

We start presenting a graphical evidence to check the validity of the parallel trends assumption. Figure 1 plots long-term sickness absence in Mandal and in the control municipalities, using yearly data. Since absence levels are in general higher for women than for men, and since around 80% of the workers in the municipal sector are women, we also provide a separate plot for female workers. Figure 1 depicts yearly averages and the vertical line indicates the time of the reform.

If we look first at the pre-reform development in absence rates, there is a visible drop in absence between 2003 and 2004, both in Mandal and in the average of the other municipalities. This drop came in the wake of a major nationwide reform in the absence certification regulation that was implemented in July 2004. The 2004 reform and its effects on absence are discussed in Markussen et al. (2012). The 2004 drop seems larger for Mandal than for the other municipalities. This suggests that the effects of the reform were potentially heterogeneous across municipalities, with stronger effects for Mandal. From 2004 to 2007, however, the difference between Mandal and the control municipalities remain stable, suggesting that the parallel trends assumption is reasonable if we do not include pre-2004 data. Considering the information after 2007, this graphical evidence suggests that this gap between Mandal and the control municipalities widens just after the reform. Such effect seems particularly large for female workers.

In order to provide more evidence on the validity of the DD assumptions, Table 1 tabulates the before and after mean values for some key variables. Comparing levels before the reform (2004–2007) and after the reform (2008–2014), Table 1 shows that Mandal moves roughly in tandem with the average of all other municipalities on all variables, with one important exception; sickness absence. There is a large drop in absence in Mandal, especially for women, while there is no such drop in the average long-term absence for all other municipalities. We can also see that the length of the spells increases by 14% in Mandal while there is a small drop in other municipalities. While the number of female employees and the number of female contracted workdays increased slightly more in Mandal when compared to other municipalities, these differential changes are not statistically significant. The differential change in unemployment is also not statistically significant.[7]

[7] The p-values are 0.522 for female employment, 0.624 for female workdays, and 0.553 for unemployment. We calculate the p-values using the inference method proposed by Ferman and Pinto (2019), which we present in more details in "Section 6.1.2" and in the Appendix C.

Fig. 1 Sickness absence for municipal workers before and after the reform. Figure 1 compares the municipal sector in Mandal to all other municipal sectors in Norway, for each year in the time interval of our analysis (2001–2014). Sickness absence (% of workdays) is the fraction of workdays lost as a percentage of the contracted workdays, counting only periods > 16 days. All municipal workers in the upper panel; only female municipal workers in the lower panel

Table 1 Descriptive statistics, before (2004–2007) and after (2008–2014) the reform. Mandal municipality (treated) and other municipalities (control)

	Mandal			Other municipalities ($N=440$)		
	Before	After	% Change	Before	After	% Change
Population	14,030	14,904	6.2	10,775	11,474	6.5
Unemployment	2.0	1.9	-5.0	2.2	1.8	-8.1
Employees	1126	1282	13.9	870	971	11.6
Contracted workdays	201,261	221,747	10.2	160,908	179,334	11.5
Sick absence (% of workdays)	6.62	5.72	-13.6	7.32	7.37	0.6
Avg. length of sick spells (days)	52.2	59.6	14.1	52.3	51.2	-2.1
Number of sickness spells	282	227	-19.6	263	294	11.6
Graded sickness spells	26.0	30.0	15.4	23.7	29.0	22.4
Female employees	885	1044	18.0	682	762	11.7
Female contracted workdays	146,727	172,031	17.2	122,456	137,218	12.1
Female sick abs (% of workdays)	7.48	6.21	-17.0	8.07	8.14	0.9
Female avg. length of spells	52.9	57.1	7.9	49.1	49.2	0.2
Female number of sick spells	233	194	-16.8	225	253	12.6
Female graded spells	29.3	34.5	17.7	27.5	33.6	22.2

Table 1 compares the municipal sector in Mandal to all other municipal sectors inNorway before and after the reform. The numbers are averaged over the respectivetime intervals (2004–2007 and 2008–2014). Sickness absence (% of workdays) is the fraction of workdays lost as a percentage of the contracted workdays, countingonly periods > 16 days. The average length of absence is found by dividing the numberof absence days by sickness spells during a year. The percentage of graded spellsis the fraction of all spells during which the worker is partly at work and partly onsick leave. Female employees refer to female workers in the municipal sector.

6.1.2 Regression results

To quantify the effects depicted in Fig. 1, we estimate Eq. (1). From the discussion in "Section 6.1.1," we consider 2004–2007 as the pre-treatment periods, and 2008–2014 as post-treatment periods. In Appendix B, we formalize the conditions in which the DD estimator is valid once we exclude the years before 2004.

From column 1 of Table 2, we estimate a 13% drop in absence relative to the baseline. Note that cluster robust standard errors at the municipality level are not reliable when we have only one treated cluster (e.g., Conley and Taber (2011)). Indeed, the inference assessment proposed by Ferman (2019a) indicates that we should expect over-rejections at the order of 91% for a 5%-level test if we rely on cluster robust standard errors at the municipality level. Therefore, in order to evaluate whether such effect is statistically different from zero, we estimate standard errors, and calculate p-values and confidence intervals, using the method proposed by Ferman and Pinto (2019). This method is an extension of the inference method proposed by Conley and Taber (2011), and is suited for settings with only one treated municipality when there is heteroskedasticity generated from variation in municipality sizes. This method allows for unrestricted serial

Table 2 Effects of the reform on absence rates for municipal and non-municipal workers

	Municipal workers			Non-municipal workers		
	All	Male	Female	All	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)
Mandal*Mandal 2008 Reform	−0.949	−0.043	−1.348	0.187	0.159	0.194
Standard error	(0.631)	(0.685)	(0.629)	(0.395)	(0.482)	(0.527)
p-value	[0.125]	[0.939]	[0.026]	[0.626]	[0.726]	[0.752]
90% confidence interval	[−2.00, 0.06]	[−1.20, 1.06]	[−2.47, −0.40]	[-0.49,0.79]	[−0.60, 0.83]	[−0.73, 1.05]
Observations	4653	4653	4653	4653	4653	4653
Baseline	7.317	4.822	8.068	5.680	5.066	6.827

correlation in the errors (a problem well documented by Bertrand et al. (2004) for DD designs), and, since we have only one treated municipality, it also even allows for some kinds of spatial correlation (see Ferman (2020)). We present in the Appendix C more details on the implementation of this inference method.

When we consider the full sample of workers, the p value of a test that the effect of the reform is zero is equal to 0.125. While we are not able to reject the null hypothesis in this case at standard significance levels, note that we can rule out large positive effects of the reform, providing evidence that the reform did not lead to a large increase in absences. The upper bound of our 90% confidence interval would imply an increase of less than 1% in the absence rate relative to the baseline.

We consider the estimated effects separately for men and women in columns 2 and 3 of Table 2. We find a small and non-significant effect for men, but a large and significant reduction in absence for women (p-value = 0.026). This effect for women is also economically meaningful, representing a drop of 17% in absence. Our data do not allow an investigation into the causal mechanisms behind the observed gender differences. Still, it is interesting that our results suggest that the policy reduced the gender gap in absence rates. Moreover, since the female share of the labor force is around 80% in Mandal municipality (which is in line with the rest of the municipal sector), it is particularly important to estimate the effects on female workers to assess the impact of such policy. All results remain virtually the same if we include unemployment rate as a covariate in the DD regression, as presented in Appendix Table 4.

In columns 4 to 6 of Table 2, we consider a placebo exercise, where we estimate the effects on workers who live in Mandal, but are employed in other sectors. Since these workers were not directly affected by the reform, we should not expect to find significant effects. Both unconditionally and when we separate by gender, we find non-significant and economically small estimated effects in these placebo regressions (the p-values are always greater than 0.60). We present in Appendix Figure 4 the trajectories of absence rates for non-municipal workers in Mandal and in other municipalities. We also present in Appendix Table 5 and Appendix Figure 5 the results from a DD model comparing municipal and non-municipal workers in Mandal, before and after the reform. All results are remarkably similar to the findings presented in columns 1 to 3 in Table 2. Overall, all these results reiterate that the effects we estimate in columns 1 to 3 are not capturing shocks specific to Mandal other than the 2008 reform, giving us confidence that our main results are not driven by municipality level unobserved variables that are not included in the DD model.

We also consider in Appendix Table 6 the changes in the gap between Mandal and other municipalities after the 2004 national reform, contrasting data from 2001 to 2003 with data from 2004 to 2007. Consistently with the graphical inspection from Fig. 1, we find a reduction in absence, which could indicate that Mandal was differentially affected by the reform, although these effects are not statistically significant at standard levels. Interestingly, we find that the point estimates on the effects of the 2004 national reform are very similar for workers in the municipal sector and for workers in other sectors, which is consistent with the fact that the national

reform affected both types of workers alike. This makes us even more confident that, if we had relevant time-varying unobservables that differentially affected Mandal, then the placebo exercise presented in columns 4 to 6 in Table 2 would have captured that. While the effects we estimate for the 2004 national reform are not statistically significant, the point estimates are relatively large, so we still consider that the DD estimates are more reliable when we exclude the 2001–2003 data from the main analysis of the Mandal reform. Including 2001–2003 data in the main DD analysis would lead to stronger negative estimated effects of the Mandal 2008 reform, which would be statistically significant even when we consider the full sample.

We also test whether the pre-trends from 2004 to 2007 where different between Mandal and other municipalities by estimating a DD model with those periods, and including a placebo dummy equal to one for Mandal after 2005. The point estimates are very small, and the p-values are very large, ranging from 0.55 to 0.77, providing further evidence in favor of our identification assumption that Mandal and other municipalities would have followed parallel trends from 2004 to 2014 in the absence of the reform.

Finally, a potential concern is that other municipalities may have implemented other reforms in the same period to reduce absence rates. If this is the case, and if these reforms were effective in reducing absence rates, then this would bias our DD estimator in the direction of finding an increase in absence rates in Mandal. In this case, our DD estimator would provide an upper bound on the effects of Mandal's reform, making our evidence even more convincing that the reform did not imply in large increases in absence rates. This same rationale is valid for the SC analysis we present next.

6.2 Synthetic control

This section assesses the effects of the reform using the synthetic control method (SCM), which was developed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) for settings with aggregate data in which a single unit is treated. The SCM applies information from the pre-reform periods to construct a synthetic Mandal, namely a convex combination of the control municipalities that best resembles the trajectory of the treated unit—Mandal—prior to the reform. Following Ferman et al. (2020), in order to avoid specification searching in the choice of predictor variables, we use the outcome of all pre-treatment periods as predictors. In this case, we choose weights for the control municipalities, $\mathbf{w}=(w_1,\ w_2,..w_N)$, that minimize the root mean squared prediction error (RMSPE) between the weighted control unit and the treatment unit over the pre-treatment period for (a weighted) combination of sickness-absence predictors and pre-treatment levels of the outcome variable (sickness absence). These weights are restricted to be non-negative and sum one. Following Ferman and Pinto (2021), we demean the data using the pre-treatment periods before estimating the SC weights, to adjust for possible bias due to discrepancies in levels between the treated and the synthetic municipality in the pre-treatment periods.

Contrary to the analysis in 'Section 6.1," where we considered only 2004–2007 as pre-treatment periods, we consider 2001–2007 as pre-treatment periods. The reason is that the SC estimator is well suited to take into account

changes in parallel trends as the one considered after the 2004 national reform. More specifically, we believe the trends between Mandal and the average of the other municipalities are not parallel because Mandal was differentially affected by the 2004 national reform. What the SC estimator aims to do in this case is to consider a weighted average of the control municipalities that was affected by the 2004 national reform in the same way as Mandal. We present this idea more formally in Appendix B. While the number of pre-treatment periods in our setting is not very large, we should expect the SC estimator to have a lower bias relative to DD if there were other common shocks after 2008 that differentially affected Mandal.[8] We also consider the SC estimator considering only 2004–2007 as pre-treatment periods.

We present in Fig. 2 a comparison between Mandal and the synthetic Mandal. This figure shows a very good fit in the pre-treatment periods, followed by a large drop in absence after the reform, particularly when we consider female workers. When we consider the average effects across all post-treatment periods using the SC estimator, we find a point estimate of −0.961 for the full sample, which is remarkably similar to our DD estimate when we consider 2004–2007 as pre-treatment periods (−0.949).[9] The estimates are also very similar when we consider the effects for women (−1.620 using SC vs. −1.348 using DD) and for men (−0.294 using SC vs. −0.043 using DD). In all cases, the SC point estimate is within the 90% confidence interval of the DD estimator. Overall, these results suggest that the parallel trends assumption we consider in the DD analysis is reasonable once we restrict the sample to 2004–2014. In contrast, if we considered the DD estimator using all periods (2001–2014), then the DD estimates would be larger than the SC estimates (in absolute values), raising concerns about the validity of the DD assumptions. We present in Appendix Table 7 the weights assigned to the municipality the received the largest weights, for each of these SC estimates.

To examine the uncertainty of the synthetic control estimates, Abadie et al. (2010) suggest comparing the effect of the real reform with placebo reforms in all the control units (compare the synthetic x and x, where x is a municipality that did not extend self-certification of sickness absence). We present in Appendix Figure 6 the differences between Mandal and synthetic Mandal, compared to the differences between the placebos and their synthetic controls. In order to take into account that the pre-treatment fit might vary depending on the placebo municipality, they construct a test statistic that is given by the post–pre ratio of the RMSPE. If this test statistic assumes an extreme value for the treated municipality, then this would indicate that we should reject the null that the reform had no effect. We present in Appendix Figure 7 the distribution of post–pre RMSPE ratios. We find a p-value of 0.028 when we do not restrict by gender, and a

[8] See Abadie et al. (2010), Botosaru and Ferman (2019), Ferman (2020), and Ferman and Pinto (2021) for a discussion on the properties of the SC estimator when potential outcomes follow a linear factor model.

[9] If we do not demean the data, as suggested by Ferman and Pinto (2021), then the SC estimate would be equal to −1.106, which is similar to what we find when we demean.

Fig. 2 Mandal and synthetic Mandal. Figure 2 compares the observed sickness absences in Mandal to the sickness absences in a synthetic Mandal for each year in the time interval of our analysis (2001–2014). Sickness absence (% of workdays) is the fraction of workdays lost as a percentage of the contracted workdays, counting only periods > 16 days. All municipal workers in the upper panel; only female municipal workers in the lower panel

p-value of 0.052 when we consider the results for women. While these results suggest that the reform had statistically significant effects on absence rates, we consider such *p*-values with caution. Since the number of municipalities is much larger than the number of pre-treatment periods, the pre-treatment RMSPE is very close to zero for many municipalities (including Mandal), which might distort the distribution of the test statistics.[10] In any case, it is reassuring that we find extreme pre-treatment RMSPE values relative to the distribution of placebos, especially once we combine with the information that both the SC and the DD estimators are remarkably similar. Since the SC estimator attempt to construct a comparison municipality that follows a similar business cycle as the treated municipality, this gives us additional confidence that the results we find are not driven by variations in the business cycle.

As a robustness check, we re-estimate the SC model considering only 2004–2007 as pre-treatment periods. We present these results in Appendix Figure 8. The estimated average effects are again very similar to the DD estimates and to the SC estimates using all pre-treatment periods (−0.949 for all workers, and −1.244 for female workers). As a final robustness check, we also re-estimate the SC model, but considering only 2001–2006 as pre-treatment periods. In this case, the weights are *not* chosen to minimize the distance between Mandal and synthetic Mandal in 2007. Still, since the treatment only started in 2008, we should expect that the synthetic Mandal would capture the trajectory of Mandal in 2007 if the SC method is working well. We present these results in Appendix Figure 9. The synthetic Mandal reconstructs the outcome for Mandal in 2007 remarkably well, even though this year was not used in the estimation of the weights. The estimated treatment effect in this exercise, considering the average between 2008 and 2014, is −1.037 for all workers and −1.597 for female workers, again very similar to our findings based on a series of other different DD and SC specifications.

7 Channels

To better understand why extended self-certification and employer involvement lead to a reduction in absence, we examine the fraction of spells that are graded (partial sickness absence) and also the number and length of the absence spells.

A worker with a graded absence certificate has moderate health problems and some work capacity left and should, accordingly, spend some time at work. Markussen et al. (2012) study a nationwide Norwegian reform in 2004 that, among other points, encouraged the substitution of graded for non-graded sick leave certificates. They argue that the reform led to shorter spells of sickness absence which in turn reduced absence levels, with graded sickness insurance workers utilize their remaining work capacity and this leads to a faster recovery and to a reduction in sickness benefits claims. Normally, it

[10] The best way to draw inference in the SC method is still under study. See, for example, Ferman and Pinto (2017), Firpo and Possebom (2018), and Hahn and Shi (2017) for discussions on the placebo test proposed in the original SC papers. Chernozhukov et al. (2019a, b) propose alternative inference methods for the SC estimator, but their methods rely on a large number of periods, which is not a good approximation to our setting. See also Abadie (2020) for an extensive review of inference methods for the SC estimator.

is the physician, together with the employer and the worker, who decides the grading of absence spells. After the reform in Mandal, the physician did not take part in this decision and the employer (the line leader) and worker together decided the grading of the absence. Could it be that the Mandal reform increased the use of grading, which then reduced overall absence, as found in Markussen et al. (2012).

Figure 3 plots both the fraction of graded spells and the average length of spells. In comparing with control municipalities, there is no evidence that graded sickness absence is more frequently used in Mandal. There is a general trend toward more grading of absence, but Mandal basically follows the trend.

Turning to the number and length of the absence spells, Table 1 uncovers that the number of absence spells in Mandal dropped almost 20% after the reform, while it increased in other municipalities. If we divide the number of spells per year by the number of municipality employees to obtain the fraction of spells per employee, there is a drop from 0.25 in the years before the reform to 0.18 in the post-reform period. This amounts to a decline of 28% in absence spells per worker. When comparing the same periods in the other municipalities, the fraction of spells per employee is the same before and after the reform (0.30). We conclude that the reform apparently lowered sickness absence at the extensive margin.

According to Fig. 3(b), the remaining spells have become longer than in the average of the other municipalities in the post-reform period. When we use length of spell as the outcome variable and estimate Eq. (1) on yearly data, we obtain a DD estimate of 8.5 days with a standard error of 4.7 days. The p-value using the inference method proposed by Ferman and Pinto (2019) equals 0.059. Measured against a pre-reform base average length of 53 days (both in Mandal and in the average of all other municipalities), this amounts to an increase of almost 17% in the length of the spells in Mandal in the reform period. When we use data from the private and central government sectors (workers who are not affected by the Mandal reform), the (placebo) DD estimate is very close to zero (0.5 days).

Combining these results, we find evidence that the self-certification reform in Mandal relegated shorter, marginal spells (of those that lasted more than 16 days). This reconciles well with the conceptual framework presented earlier, according to which the reform led to a negative shift in demand, implying that workers with minor health issues did not demand sick leave.[11]

8 Discussion and conclusion

We find that allowing workers to self-declare absence—allowing them to skip the doctor certificate—did not lead to increased sickness absence in Mandal. On the contrary, for female workers, the reform actually resulted in a significant drop in absence. We believe that the DD estimator we use captures the effect of the reform. The pre-reform

[11] It is not possible to disentangle whether the effects on average length of spells comes from a change in the composition of the spells after the reform, or from an intensive margin effect of the reform on the length of the spells. Since we find evidence that the reform reduced demand for sick leave, we interpret this increase in average length of spells from a change in the composition of spells.

Fig. 3 Graded sickness absence and the length of spells. Figure 3 displays graded sickness absence as a percentage of all spells (upper panel) and mean length of absence spells (lower panel); Mandal relative to an average of all other municipalities in Norway, for each year in the time interval of our analysis (2001–2014). Graded sickness absence is a partial absence used when the worker has some work capacity left and spend some time at work

trend in the treated municipality is moving in parallel with the average of all other municipalities. There is a sharp drop in absence just around the time of the reform in the treated municipality. If there were a contemporaneous Mandal shock to absence at this time, we should expect to see a similar drop in absence for workers in Mandal who are not affected by the reform. That did not happen.

We explain the effect as a decline in the demand for sickness absence. For a given demand for absence, skipping the doctor as an absence certifier, as a gatekeeper, would increase the level of absence. By how much depends on the magnitude of moral hazard, that is, how many workers who are healthy enough to work claim absence benefits, and how rigorous physicians are as gatekeepers. However, as we explain in detail in a "Section 3," there are several features of this reform that may induce workers to lower the demand for absence. Our results show that the decline in demand dominates the direct effect of removing the physician as a gatekeeper.

There is little prior empirical research on the effects of medical certificates. One exception is the assessment of a Swedish experiment with extended self-certification of work absence, by Hesselius et al. (2013) and Hesselius et al. (2009). A random sample of workers in two different municipalities could self-certify one extra week—two instead of one—of sickness absence. In comparison with workers who did not obtain the extra week, the treatment group increased their absence; on average, absenteeism increased by 0.8 days per year, from 11.8 to 12.6 days. In the Swedish case, the gatekeeper effect dominated the demand effect.

There are several differences between Swedish experiment and the self-certification reform we study. First, the Mandal reform differs in the level of discretion and trust it grants the employees. In Sweden, the workers' received one extra week, while in the present case they can self-certify for the whole entitlement period (1 year). The Mandal reform was branded (with a handshake logo) as the "Trust Project" and appealed openly to workers' responsibility and reciprocity. In addition, and maybe just as important, the reform in Mandal also implied a stronger employer involvement in the counselling of the sick workers, which was not the case in Sweden. This intense counselling may have increased the hazzle and costs of being absent. Olsen and Jentoft (2012) report from focus group interviews with leaders and employees that participated in the Mandal project. Some of those who were interviewed reported that the meticulous registration of absence and the frequent meetings between the line leaders and the sick absent employees felt intrusive and added costs to being absent (page 104).

For policy, our finding is a significant result. Using medical personnel to certify absence is costly for the doctors, the patients, and the insurer (which reimburses the medical doctors). Our analysis indicates that sickness certification can be taken out of the hands of the physicians without a subsequent rise in sickness absence. In fact, extended self-certification of sickness absence in Mandal appears to be a win–win reform, with less absence and fewer resources needing to be used on certification. However, note that extended self-certification of absence implies extended employer involvement, which is likely to use administrative resources in the municipality.

A natural question at this point concerns the extent to which our findings have external validity. First, since the reform involved stronger involvement from the municipality administration in addition to self-certification, we cannot guarantee that the effects would be the same if we did not have this involvement. Still, since these follow-ups from the employees'

line managers are much less costly than physician certification, our results indicate that it is possible to turn down the requirement of physician certification without implying large increases in absence rates. Moreover, even if we identify a reform effect in Mandal, there could be specific attributes of Mandal that made skipping the medical doctor as a sickness certifier especially effective. It is reassuring that Mandal appears to be very much an "average" municipality if we look at the pre-reform data (on sickness absence or other variables such as, age, gender composition, and unemployment). Another possible source of singularity of Mandal might be that the key persons who initiated the Trust Project in Mandal are just as important as the reform itself. Again, it is reassuring that a team of leaders of the Trust Project were present also in the years before the extended self-certification was introduced. Another relevant point is that physicians in other countries may be stricter gatekeepers than the physicians in Norway, and hence if the same reform was introduced in another country the direct effect of skipping the doctor as an absence certifier—thereby pulling towards more absence—may dominate the decline in demand for absence. Although we cannot address this issue, our findings should encourage more sickness absence insurers to experiment with extended self-certification of sickness absence.

Appendix A: Appendix Tables and Figures

Table 3 Descriptive statistics, 2001–2014

	Mandal		Other municipal ($N=422$)	
	Municip. sector (treatment)	Non-municip	Municip. sector (control)	Non-municip
Population	14,347 (669)		11,073 (33,100)	
Unemployment	2.1 (0.6)		2.1 (1.0)	
Employees	1196 (100)	3602 (260)	910 (2298)	3840 (18,264)
Contracted workdays	209,005 (15,524)	717,551 (57,814)	167,811 (440,687)	805,927 (3,905,003)
Sick absence (% of workdays)	6.58 (1.20)	6.67 (1.06)	7.55 (1.62)	5.90 (1.46)
Av. length of sick spell (days)	57 (6)	70 (6)	52 (10)	66 (10)
Graded sickness spells	27.7 (5.0)	26.2 (6.0)	25.2 (9.5)	23.6 (8.0)
Female employees	959 (97)	1444 (80)	717 (1680)	1585 (7769)
Female contracted workdays	158,119 (16,199)	250,153 (19,365)	128,633 (318,527)	307,571 (1,578,729)
Female sick absence (%)	7.26 (1.36)	7.82 (1.09)	8.35 (1.82)	7.18 (1.96)
Female av. length of spell	56 (5)	62 (5)	49 (8)	58 (12)
Female graded spells (%)	31.5 (5.8)	29.6 (6.8)	29.3 (8.8)	27.3 (10.2)

Table 3 compare Mandal municipality with all other municipalities in Norway. The numbers are averaged over the time interval of our analysis [2001, 2014]. Standard errors in parenthesis. Employees in "Non-municipal" sectors include all workers in the private sector and those working for the central government. Sickness absence (% of contracted workdays) is the fraction of workdays lost as a percentage of the contracted workdays, counting only periods > 16 days. The average length of absence is found by dividing the number of absence days by sickness spells during a year. The percentage of graded spells is the fraction of all spells during which the worker is partly at work and partly on sick leave

Table 4 Effects of the reform on absence rates for municipal and non-municipal workers

	Municipal workers			Non-municipal workers		
	All	Male	Female	All	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)
Mandal*Mandal 2008 Reform	-0.925	-0.024	-1.325	0.190	0.162	0.203
Standard error	(0.634)	(0.688)	(0.622)	(0.399)	(0.485)	(0.537)
p value	[0.132]	[0.969]	[0.028]	[0.619]	[0.726]	[0.738]
90% confidence interval	[-1.99,0.04]	[-1.23,1.07]	[-2.45,-0.36]	[-0.50,0.80]	[-0.61,0.83]	[-0.72,1.08]
Observations	4653	4653	4653	4653	4653	4653
Baseline	7.317	4.822	8.068	5.680	5.066	6.827

This table replicates Table 2, but including unemployment rate as a control variable in the DD regression

Table 5 DD estimates using non-municipal workers in Mandal as control group

	All	Male	Female
	(1)	(2)	(3)
Panel A: exclude years before national reform			
(Munic worker)*Mandal 2008 Reform	−1.086	−0.311	−1.725
Standard error	(0.553)	(0.871)	(0.308)
p-value	[0.053]	[0.688]	[0.025]
90% confidence interval	[−1.93, −0.07]	[−1.82, 0.95]	[−2.60, −1.10]
Panel B: include years before national reform			
(Munic worker)*Mandal 2008 Reform	−1.022	−0.207	−1.598
Standard error	(0.623)	(0.975)	(0.415)
p-value	[0.100]	[0.829]	[0.016]
90% confidence interval	[−1.90, 0.09]	[−1.72, 1.39]	[−2.56, −0.79]
Panel C: pre-trends (data from 2001 to 2007)			
(Munic worker)*Post 2004 dummy	0.151	0.241	0.295
Standard error	(0.774)	(1.195)	(0.888)
p-value	[0.796]	[0.863]	[0.709]
90% confidence interval	[−0.97, 1.44]	[−1.20, 2.41]	[−1.16, 1.64]

The outcome variable is the fraction of workdays lost as a percentage of the contracted workdays, counting only periods > 16 days. We estimate a DD regression using non-municipal workers in Mandal as control group. The DD variable is equal to one for Mandal after 2007. Year and municipality dummies are included in the regressions, but not reported in the table. In Panel A, we use data from 2004 to 2014, while in Panel B we use data from 2001 to 2014. In Panel C, we use data from 2001 to 2007, and consider the differential effects of the national reform for municipal and non-municipal workers. Standard errors, *p*-values, and confidence intervals are calculated using the method similar to the one proposed by Ferman and Pinto (2019), where we estimate the same regression for each control municipality, and then adjust for heteroskedasticity as a function of the number of municipal and non-municipal workers.

Table 6 Differential effects of the 2004 national reform

	Municipal workers			Non-municipal workers		
	All	Male	Female	All	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)
Mandal* National 2004 reform	−1.014	−1.167	−0.834	−1.045	−0.983	−1.195
Standard error	(0.690)	(0.948)	(0.725)	(0.671)	(0.697)	(1.009)
p -value	[0.137]	[0.187]	[0.234]	[0.104]	[0.139]	[0.158]
90% confidence interval	[−2.18, 0.11]	[−2.60, 0.36]	[−2.12, 0.31]	[−2.27, −0.15]	[−2.12, 0.03]	[−2.83, 0.23]
Observations	2961	2961	2961	2961	2961	2961
Baseline	8.253	5.427	9.185	6.707	6.072	7.857

The outcome variable is the fraction of workdays lost as a percentage of the contracted workdays, counting only periods > 16 days. We estimate Eq. (1) using 2001–2007 data where the DD variable is equal to one for Mandal after 2003. Year and municipality dummies are included in the regressions, but not reported in the table. Columns 1 to 3 consider absence for municipal workers, while columns 4 to 6 consider absence for workers from other sectors. Standard errors, p-values, and confidence intervals are calculated using the method proposed by Ferman and Pinto (2019). The baseline estimate is absence in the control municipalities in the pre-reform period

Table 7 Distribution of SC weights

All workers			Female workers		
Municiplity		Weights	Municipality		Weights
1852	Tjeldsund	0.100	2015	Frogn	0.090
1943	Kvænangen	0.057	1943	Kvænangen	0.048
1920	Lavangen	0.047	1828	Senja	0.045
2024	Berlevåg	0.042	1853	Evenes	0.013
1845	Sørfold	0.038	1832	Hemnes	0.011
1832	Hemnes	0.019	1850	Narvik	0.009
1853	Evenes	0.016	1632	Oppdal	0.007
1815	Vega	0.010	1441	Stad	0.006
1828	Nesna	0.008	1750	Nærøysund	0.006
1913	Tjeldsund	0.007	1815	Vega	0.006
2022	Lebesby	0.005	1852	Hamarøy	0.006
1740	Namsskogan	0.004	1740	Namsskogan	0.005
1755	Inderøy	0.004	1913	Tjeldsund	0.005
1812	Sømna	0.004	1920	Lavangen	0.005
1849	Steigen	0.004	118	Aremark	0.004
1854	Lødingen	0.004	429	Åmot	0.004
1856	Røst	0.004	1032	Lyngdal	0.004
1859	Flakstad	0.004	1145	Bokn	0.004
1929	Senja	0.004	1755	Inderøy	0.004
118	Aremark	0.003	1859	Flakstad	0.004

This table presents the distribution of SC weights for the 20 municipalities that received the largest weights. The SC weights are estimated using 2001–2007 as pre-treatment periods, and uses all pre-treatment outcome values as predictors.

Fig. 4 Absence for non-municipal workers before and after the reform. Figure 4 compares the non-municipal workers in Mandal to non-municipal workers in all other municipalities in Norway, for each year in the time interval of our analysis [2001, 2014]. Sickness absence (% of workdays) is the fraction of workdays lost as a percentage of the contracted workdays, counting only periods > 16 days. Importantly, the differences in trends before and after the reform are not statistically different from zero, as presented in columns 4 to 6 from Table 2. All municipal workers in the upper panel; only female municipal workers in the lower panel

Fig. 5 Absence for municipal and non-municipal workers in Mandal before and after the reform. Figure 5 compares the municipal and non-municipal workers in Mandal, for each year in the time interval of our analysis [2001, 2014]. Sickness absence (% of workdays) is the fraction of workdays lost as a percentage of the contracted workdays, counting only periods > 16 days. The *p*-values of tests that the differences in trends before and after the reform are statistically different from zero are presented in Appendix Table 4. All municipal workers in the upper panel; only female municipal workers in the lower panel

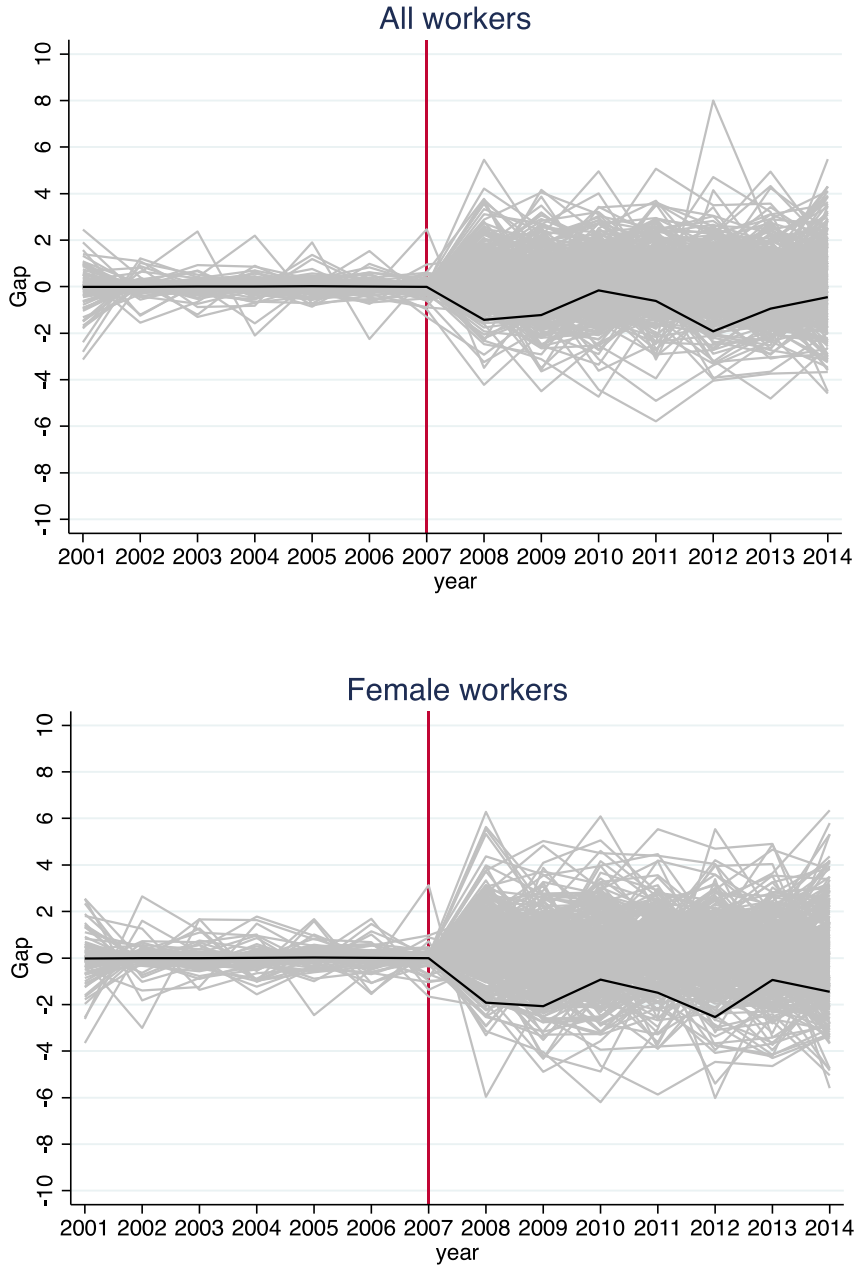

Fig. 6 Differences between Mandal and synthetic Mandal, compared to placebos. The black series present the differences between Mandal and the synthetic Mandal when we estimate the SC weights using 2001–2007 as pre-treatment periods. The gray lines present the results from placebo estimates using each of the control municipalities as the treated. Note that, in addition to having a very large number of control municipalities, there are many municipalities much smaller than Mandal, which explains why we have some placebos with worse pre-treatment fit and with larger gaps in the post-treatment periods

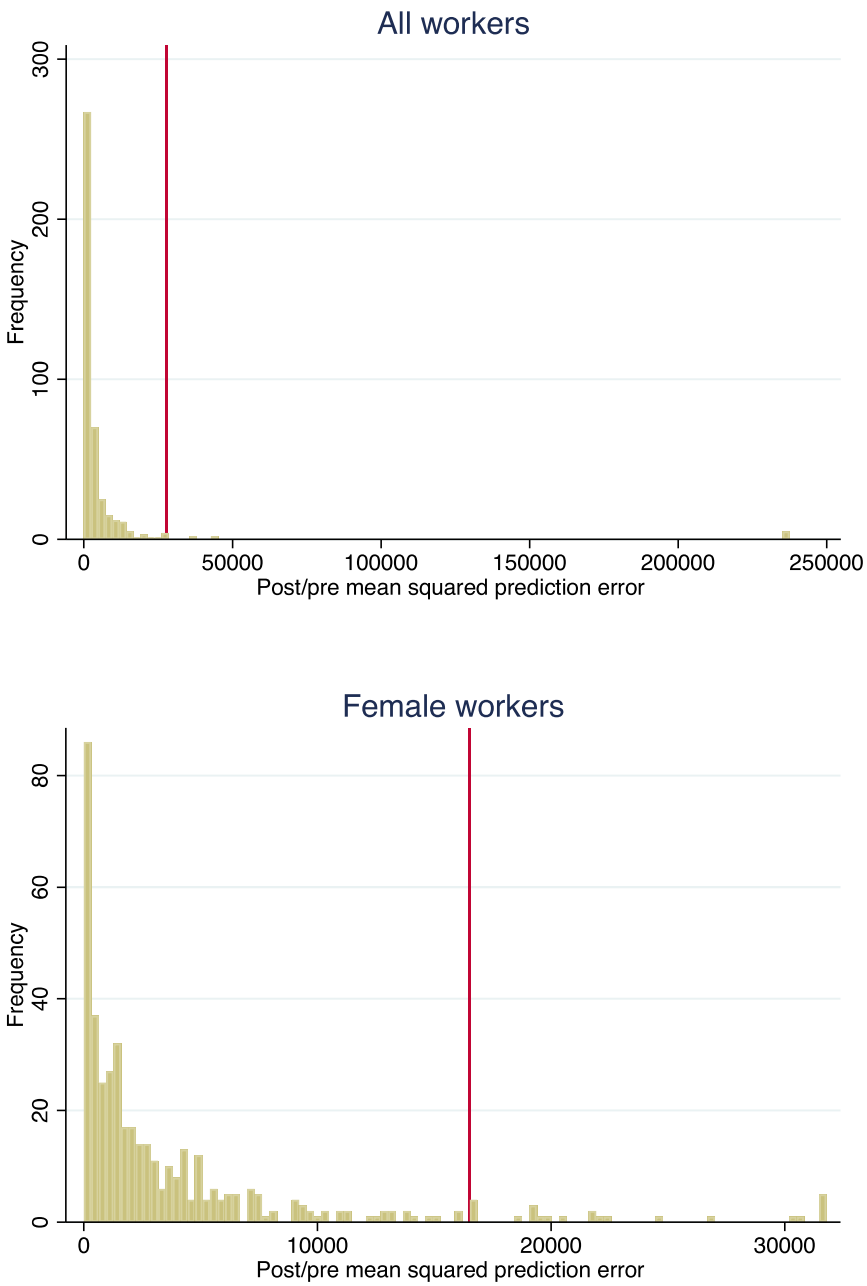

Fig. 7 Distribution of post/pre mean squared prediction error. These figures present the distribution of post/pre mean squared prediction error when we estimate the effects for Mandal and for the placebo municipalities. The red line represents the statistic for Mandal. When we consider all workers, there are only 2.8% of the placebos with a statistic larger than the one for Mandal. When we consider female workers, there are only 5.2% of the placebos with a statistic larger than the one for Mandal

Fig. 8 Synthetic control results using only 2004–2007 as pre-treatment periods. These figures compare Mandal and the synthetic Mandal when we estimate the SC weights using only 2004–2007 as pre-treatment periods. The estimated treatment effect in this exercise is −0.949 for all workers and −1.244 for female workers. In both cases, the estimates are close to the DD estimates, and to the SC estimates using 2001–2007 as pre-treatment periods

Fig. 9 Synthetic control results using only 2001–2006 as pre-treatment periods. These figures compare Mandal and the synthetic Mandal when we estimate the SC weights using only 2001–2006 as pre-treatment periods. In both cases, the synthetic Mandal reconstructs the pre-treatment 2007 remarkably well, even though 2007 was not used in the estimation of the weights. The estimated treatment effect in this exercise is − 1.063 for all workers and − 1.628 for female workers

Appendix B: formalizing the potential heterogeneous effects of the 2004 national reform

Consider a model

$$Y_{it} = ad_{it} + \gamma_t + \theta_i + \lambda_t \mu_i + \epsilon_{it},$$

where Y_{it} is the outcomes of municipality i at time t, d_{it} is the treatment dummy (= 1 for Mandal after 2007, and zero otherwise), and θ_i and γ_t are municipality and time fixed effects. The term $\lambda_t \mu_i$ represents the effects of the 2004 national reform, which we allow to have differential effects for each municipality. The unobserved parameter μ_i reflects how municipality i was affected by the national reform. We assume that $\lambda_t = 0$ for the periods before the reform, and $\lambda_t = \lambda$ for the periods after the reform. This is arguably a reasonable assumption, given the evidence from Fig. 1 that Mandal and the other municipalities follow parallel trends after the national reform (and before 2008). We assume that the idiosyncratic shocks ϵ_{it} have mean zero for all i and t, which is standard in the DID and SC literatures.

Given this model, the DID estimator is unbiased if we restrict the analysis to the periods after the national reform. In this case, the municipality fixed effects would capture the terms $\theta_i + \lambda_t \mu_i = \theta_i + \lambda \mu_i$, which are constant across time for each municipality in this time frame. Therefore, the fact that the national reform may have had different effects for each municipality does not pose any problem for the DID estimator. If, however, we consider periods before the reform, then the DID estimator would be biased. In this case, we would have that $\lambda_t \mu_i$ would not be constant across time (because it is zero before the national reform), so the municipality fixed effects would not correctly control for that. This is why we focus on the DID estimator with the post-national reform periods in Section 6.1.

In contrast, the SC estimator would remain unbiased if the SC weights are such that the weighted average of the controls using those weights recover μ_i. As shown by Abadie et al (2010) and Botosaru and Ferman (2019), this will happen when we have a good pre-treatment fit for many pre-treatment periods. As Ferman (2019b) show, this will also happen when weights are diluted among many controls in a setting with many pre-treatment periods and many controls, even when the pre-treatment fit is imperfect. Finally, as shown by Ferman and Pinto (2021), even when these conditions do not hold, a demeaned version of the SC estimator will generally improve in terms of variance and bias relative to the DID estimator. For these reasons, we consider the demeaned SC estimator using all pre-treatment periods. It is also reassuring that the results remain virtually the same if we consider only the pre-treatment periods after the national reform.

Finally, note that the same rationale above would be valid if we have other time-varying unobserved confounders of the form $\tilde{\lambda}_t \tilde{\mu}_i$, even if they are related to events that we do not observe. For example, we can think that $\tilde{\lambda}_t$ reflects some aggregate shocks, while $\tilde{\mu}_i$ reflects how such aggregate shocks affect municipality i. If those aggregate shocks affect all municipalities in the same way (that is, $\tilde{\mu}_i = \tilde{\mu}$), then the year fixed effects would take that into account in the DID estimator, and such aggregate shocks would not generate bias. In contrast, if $\tilde{\mu}_i$ varies across i, then the DID estimator may be biased, while the SC estimator would generally correct for that (or at least ameliorate it). The fact that we find

remarkably similar estimates when we consider the DID and the SC estimators provides further evidence that the possibility of time-varying confounders $\tilde{\lambda}_t\tilde{\mu}_i$ do not represent first order concerns in our setting.

Appendix C: Inference method

We present in more details the inference method considered in Section 6.1.2 for the DD estimator, which is based on Ferman and Pinto (2019). The main idea is similar to the inference method proposed by Conley and Taber (2011), but allowing for heteroskedasticity. More specifically, we allow for heteroskedasticity that arises from the fact that different municipalities have different population sizes, and therefore municipality × time aggregates will have lower variance when a municipality has a larger population.

Let T be the total number of periods. Treatment starts after period t^*, and let municipality $i = 1$ be the treated one. Define the linear combination of the errors of municipality i given by $W_i = \frac{1}{T-t^*}\sum_{t=t^*+1}^{T}\varepsilon_{it} - \frac{1}{t^*}\sum_{t=1}^{t^*}\varepsilon_{it}$. As Conley and Taber (2011) show, if we have a single treated municipality and the number of municipalities goes to infinity, then the DD estimator converges in probability to $\delta + W_1$. Therefore, the estimator is unbiased for δ, but there is an uncertainty due to the linear combination of the errors of the treated municipality, W_1.

The idea from Conley and Taber (2011) to tackle uncertainty in this setting is the following: if W_i has the same distribution for all i, then we can estimate the distribution of W_1 by considering the empirical distribution of the residuals of the control $\widehat{W}_i$ for $i > 1$. They show that, under this homoskedasticity assumption, this strategy asymptotically works when the number of control municipalities goes to infinity. A potential problem with this approach noticed by Ferman and Pinto (2019), however, is that such homoskedasticity assumption will generally not hold if we observe municipality × time aggregates, because municipalities with larger populations would tend to have lower variances. Therefore, we would tend to have over-rejection when the treated municipality is relatively small, and over-rejection when it is relatively large.

Let M_{it} be the number of individual observations used to calculate the outcomes of municipality i at time t (in our setting, this would be given by the population of municipality i at time t), and $M_i = (M_{i1}, ..., M_{iT})$. Then, Ferman and Pinto (2019) show that, under a wide range of possible assumptions on the intra-municipality correlation,

$$\mathrm{Var}\left[\mathrm{W_i}\middle|\mathrm{M_i}\right] = \mathrm{A} + \mathrm{B}\left(\frac{1}{(\mathrm{T}-\mathrm{t}^*)^2}\sum_{\mathrm{t}=\mathrm{t}^*+1}^{\mathrm{T}}\frac{1}{\mathrm{M_{it}}} + \frac{1}{(\mathrm{t}^*)}\sum_{\mathrm{t}=1}^{\mathrm{t}^*}\frac{1}{\mathrm{M_{it}}}\right)$$

where A and B are constants, and $h(M_i) \equiv \frac{1}{(T-t^*)^2}\sum_{t=t^*+1}^{T}\frac{1}{M_{it}} + \frac{1}{(t^*)^2}\sum_{t=1}^{t^*}\frac{1}{M_{it}}$.

The idea then is to estimate the parameters A and B, which implies that we can have an estimator for the variance of W conditional M. For the specification considered in column 1 of Table 2, we estimate $\widehat{A} = 0.321$ and $\widehat{B} = 231.67$. Evaluating this function at M_1 implies that the variance of W_1 is equal to 0.398. Since,

asymptotically, the distribution of the DD estimator depends only on W_1, our estimate for the standard error of the DD estimator, which is presented in Table 2, is given by $\sqrt{0.398} = 0.631$.

In our application, the estimated variance of W_i conditional on M_i ranges from 0.323 to 2.473 suggesting a relevant level of heteroskedasticity coming from variation in population sizes. Interestingly, the median variance of W_i is 0.528, which is larger than the variance of W_1. Therefore, if we followed Conley and Taber (2011) approach, then we should expect their inference method to be too conservative. The reason is that there are many municipalities with smaller population relative to Mandal. Therefore, we would recover a more disperse distribution for W_1 than we should if we do not take that into account.

Assuming further that W_i has the same distribution for all i up to a scale parameter, then we can recover the distribution of W_1 by dividing the residuals $\widehat{W_i}$ of the controls by the squared root of the estimated variance conditional on M_i (which asymptotically recovers a distribution with variance equal to one), and then multiplying by the squared root of the estimated variance conditional on M_1 (which then asymptotically recovers a distribution with the variance of the linear combination of the errors of the treated unit). Let's denote these re-scaled residuals by $\tilde{W}_i$.

With this estimated distribution of W_1, we can calculate the p value for this test, which is given by the proportion of control municipalities in which the absolute value of $\tilde{W}_i$ is greater than the absolute value of the DD estimator.[12] We can also construct confidence intervals by looking at the quantiles of $\tilde{W}_i$ among the control municipalities. In this setting with only a single treated municipality, Ferman (2020) shows that this approach is valid even when we allow for spatial correlation in the errors, provided we assume a strongly mixing condition.

Acknowledgments We gratefully acknowledge comments from three anonymous referees, editor Shuaizhang Feng, Tarjei Havnes, Edwin Leuven, Knut Roed, Torben Mideksa, and Arnstein Mykletun.

Funding Open access funding provided by University of Bergen (incl Haukeland University Hospital). Financial support from the Norwegian Labour and Welfare Administration (the FARVE program) and from the Research Council of Norway, project No 257598, is greatly appreciated.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

[12] This is the approach considered by Conley and Taber (2011) and Ferman (2020). The approach described by Ferman and Pinto (2019) is slightly different in that they propose a bootstrap resampling on the distribution of $\tilde{W}_i$ for both the treated and the control groups. These two approaches are asymptotically equivalent.

References

Abadie A(2020) Using synthetic Controls: Feasibility, Data Requirements, and methodological aspects. Journal of Economic Literature, forthcoming.

Abadie A, Gardeazabal J (2003) The economic costs of conflict: a case study of the Basque country. Am Econ Rev 93(1):113–132

Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc 105(490):493–505

Angelov N, Johansson P (2020) Lindahl, E (2020) Sick of family responsibilities? Empir Econ 58:777–814

Askildsen JE, Bratberg E, Nilsen ØA (2005) Unemployment, labor force composition and sickness absence: a panel data study. Health Econ 14(11):1087–1101

Avdic D, Johansson P (2017) Absenteeism, gender and the morbidity-mortality paradox. J Appl Econometrics 32:440–462

Bénabou R, Tirole J (2006) Incentives and prosocial behavior. Am Econ Rev 96(5):1652–1678

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust difference-in-difference estimates? Q J Econ 119(1):249–275

Botosaru I, Ferman B (2019) On the role of covariates in the synthetic control method. Econometrics J 22(2):117–130

Carlsen B, Lind JT, Nyborg K (2020) Why physicians are lousy gatekeepers: sicklisting decisions when patients have private information on symptoms. Health Econ 29:778–789

Chernozhukov V, Wuthrich K, Zhu Y (2019a) An exact and robust conformal inference method for counterfactual and synthetic controls. arXiv e-prints, arXiv:1712.09089v6

Chernozhukov V, Wuthrich K, Zhu Y (2019b) Practical and robust *t*-test based inference for synthetic control and related methods. arXiv e-prints, arXiv:1812.10820v2

Conley TG, Taber CR (2011) Inference with difference in differences with a small number of policy changes. Rev Econ Stat 93(1):113–125

Cools S, Markussen S, Strøm M (2017) Children and careers: how family size affects parents' labor market outcomes in the long run. Demography 54(5):1773–1793

Dionne G, St-Michel P (1991) Workers' compensation and moral hazard. Rev Econ Stat 73(2):236–244

Ellingsen T, Johannesson M (2008) Pride and prejudice: the human side of incentive theory. Am Econ Rev 98(3):990–1008

Falk A, Kosfeld M (2006) The hidden costs of control. Am Econ Rev 96(5):1611–1630

Ferman B (2019a) A simple way to assess inference methods. arXiv e-prints, arXiv: 1912.08772

Ferman B (2019b) On the properties of the synthetic control estimator with many periods and many controls. arXiv e-prints, arXiv: 1906.06665

Ferman B (2020) Inference in differences-in-differences with few treated units and spatial correlation. arXiv e-prints, arXiv: 2006.16997

Ferman B, and Pinto C (2017) Placebo tests for synthetic controls. MPRA Paper 78079, Germany: University Library of Munich

Ferman B, Pinto C (2019) Inference in differences-in-differences with few treated groups and heteroskedasticity. Rev Econ Stat 101(3):452–467

Ferman B, Pinto C (2021) Synthetic controls with imperfect pre-treatment fit. Quantitative Economics, forthcoming

Ferman B, Pinto C, Possebom V (2020) Cherry picking with synthetic controls. J Pol Anal Manag 39(2):510–532

Firpo S, Possebom V (2018) Synthetic control method: inference, sensitivity analysis and confidence sets. J Causal Inference, 6(2)

Hahn J, Shi R (2017) Synthetic control and inference. Econometrics 5(4):52

Hauge K, Markussen S, Raaum O, Ulvestad M (2015) Can the gender gap in sickness absence be explained by attitudes, norms and preferences? Søkelys på arbeidslivet 32(4):298–324 in Norwegian

Henrekson M, Persson M (2004) The effects on sick leave of changes in the sickness insurance system. J Labor Econ 22(1):87–113

Hesselius P, Nilsson JP, Johansson P (2009) Sick of your colleagues' absence? J Eur Econ Assoc 7(2–3):583–594

Hesselius P, Johansson P, Larsson L (2013) Monitoring sickness insurance claimants: evidence from a social experiment. Lab Econ 20:48–56

Markussen S, Mykletun A, Røed K (2012) The case for presenteeism. evidence from Norway's sickness insurance program. J Publ Econ 96(11):959–972
Markussen S, Røed K, Røgeberg O (2013) The changing of the guards: can family doctors contain worker absenteeism? J Health Econ 32(6):1230–1239
Mastekaasa A (2014) The gender gap in sickness absence: long-term trends in eight European countries. Eur J Publ Health 24(4):656–662
Mastekaasa A (2015) Social and demographic variations in short-term sickness absence. Søkelys på arbeidslivet 31:3–20 in Norwegian
Mastekaasa A, Melsom AM (2014) Occupational segregation and gender differences in sickness absence: Evidence from 17 european countries. Eur Socio Rev 30(5):582–594
OECD (2010) Sickness, disability and work: breaking the barriers; a synthesis of findings across OECD countries. OECD
Olsen T, Jentoft N (2012) Tillitsprosjektet: Innovasjon ved bruk av 365 egenmeldingsdager. Evaluering av Tillitsprosjektet i Mandal kommune. (In Norwegian)
Svärdsudd L, Englund K (2000) Sick-listing habits among general practitioners in a Swedish county. Scand J Prim Health Care 18(2):81–86