CESifo WORKING PAPERS

11209
2024

July 2024

Healthy at Work? Evidence from a Social Experimental Evaluation of a Firm-Based Wellness Program

Marianne Simonsen, Lars Skipper

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11209*

Healthy at Work? Evidence from a Social Experimental Evaluation of a Firm-Based Wellness Program

Abstract

We employ a large social experiment combined with register-based data allowing for up to 12-year follow-up to evaluate a long-lasting employer-sponsored health and well-being program. We show that employees at treated worksites receive fewer consultations from their primary care physician and purchase fewer prescription drugs. These effects persist up to seven years after randomization, though with some fade-out. We find no effects on overall hospitalizations, neither in the short or longer run, and the program was not successful in improving labor-related outcomes such as absence and turnover. Finally, we show some evidence of spillovers within the family.

JEL-Codes: I120, I180.

Keywords: worksite health program, health outcomes, labor outcomes, social experiment.

Marianne Simonsen
Department of Economics and Business Economics, Aarhus University / Denmark
msimonsen@econ.au.dk

Lars Skipper
Department of Economics and Business Economics, Aarhus University / Denmark
lskipper@econ.au.dk

2024

The trial has been registered with the AEA RCT Registry as “Healthy at Work” with RCT ID AEARCTR-0005710. The opinions expressed in this paper are those of the authors and neither those of Forebyggelsesfonden nor those of the team behind “Rask-i-Job”. We appreciate valuable comments and feedback from Gordon Dahl, Nabanita Datta Gupta, Tor Eriksson, Corina Mommaerts, Lars Nielsen, and Jeffrey A. Smith. We also acknowledge comments from seminar participants at Aarhus University, the University of Copenhagen, Pompeu Fabra University, the European Commission’s Joint Research Centre at Ispra, CESifo, and University of Wisconsin-Madison as well as from participants at the 4th Family and Education workshop and the 2018 ESPE conference. The usual disclaimer applies.

1. Introduction

This paper evaluates the *Healthy-at-Work* initiative, a 15-month comprehensive worksite health promotion program, designed to encourage healthier behaviors among employees. The program comprised health screenings, short courses promoting healthier living, and most significantly, the provision for employees to exercise for up to two hours weekly during working hours.

We examine the effects of this intervention through a randomized controlled trial (RCT), randomizing at the work-unit level. This allows for potential spillover effects within work-units, while more accurately reflecting a real-world employer's approach to implementation. Thus, our study is a pragmatic (or effectiveness) trial. Unlike an explanatory trial, we are not primarily interested in the intrinsic benefits of exercise, but rather in whether a comprehensive health promotion program, featuring physical exercise, can have meaningful impact when implemented in real-world, large-scale work settings.

Our study focuses on a significant and uniquely susceptible population – approximately 7,500 healthcare workers in elderly care in Denmark. This group, all covered by universal healthcare, frequently grapples with issues of burnout and illness. Their work environment is demanding, and the toll it takes on physical and emotional health is considerable.[1] The ubiquity of such challenges among these workers underscores a high potential for meaningful health improvements through initiatives such as the *Healthy-at-Work* initiative. The significant stressors they face make them an especially relevant population for exploring the impact of comprehensive health promotion programs. By focusing on this population, our study investigates the program's effects in an environment where significant improvements are not just desirable, but crucially needed.

While our study, initiated in 2008, was not pre-registered—reflecting the norm in economics at the time—we focus on outcomes directly linked to the program's goals. These include health utilization and labor market indicators such as absenteeism, turnover, hours worked, and earnings. We also align our outcome domains with those of the paper most similar to ours (Song and Baicker, 2019) and control for multiple hypothesis testing to limit false discoveries (Westfall and Young, 1993; Jones et al., 2019). To further refine our analysis, we use machine

[1] E.g., Harrad and Sulla, 2018; U.S. Bureau of Labor Statistics, 2019.

learning techniques to explore heterogeneity across different subsamples (Chernozhukov et al., 2020).

We find significant effects on health utilization. Individuals employed at worksites that received the *Healthy-at-Work* intervention had fewer consultations with their primary care physicians, a trend that persisted for up to seven years following the program's initiation. These impacts were even more pronounced for employees in poorer health prior to the intervention. Simultaneously, we find reductions in prescription drug purchases over the same period but find no overall indications of effects on hospitalizations in the long run. While we do observe a short-term reduction in primary care utilization, potentially attributable to the program's health screenings, this cannot explain the long-term reductions in health care utilization.

However, the program did not significantly improve absenteeism and turnover rates, the primary managerial goals associated with the intervention. Moreover, a cost-benefit analysis reveals that the financial gains from *Healthy-at-Work* fell far short of offsetting its costs, marking a notable deficit in its financial viability.

Finally, we observed that our intervention induced social spillovers affecting spouses' short-term health seeking behaviors, which underscores that our health behaviors are influenced by those of our close connections and highlights the importance of considering these influences when evaluating health promotion programs. As far as we are aware, only Fletcher and Marksteiner (2017) have previously leveraged experimental designs to answer a similar question.

Our research is particularly pertinent given the backdrop of increasing rates of noncommunicable, chronic diseases worldwide. Cardiovascular disease, cancer, and diabetes collectively account for 71% of global deaths, a significant proportion of which occur prematurely, before old age (WHO, 2018). Sedentary lifestyles, including those engendered by many modern workplaces, play a crucial role in the escalation of these chronic conditions, alongside other modifiable behaviors such as tobacco use, excessive alcohol consumption, and unhealthy diets. Worksite health promotion programs, like *Healthy-at-Work*, have emerged as potential tools to combat these challenges.[2] By promoting healthier behaviors and lifestyle changes within the workplace, these programs aim to curb the risk factors associated with

[2] For evidence of this wide interest, we refer the reader to Dishman et al. (1998), Aldana and Pronk (2001), Baicker et al. (2010), and Rongen et al. (2013).

chronic diseases, thereby enhancing employee health and well-being. Exercise, a key component of many of these programs, would potentially help with reducing the risk of cardiovascular diseases, diabetes, and osteoporosis, help control weight, and promote psychological well-being. However, individuals face many barriers to activity and healthy living, including lack of time, feelings of embarrassment, inability to participate, or simply lack of enjoyment, as pointed out by Charness and Gneezy (2009). Well-designed health promotion programs can mitigate these obstacles by creating supportive physical and social environments for health improvement. By integrating health promotion into the organizational structure, these programs have the potential to address the problems of sedentary behavior and poor health practices. Given that adults spend approximately half of their waking hours at worksites, these venues offer a unique opportunity for promoting and encouraging healthy living and physical activity. They provide an avenue for reaching large numbers of individuals and delivering behavioral interventions in a setting that encourages shared experiences, mutual support, and sustained changes. The anticipated benefits of such initiatives extend beyond individual health outcomes, with expectations of lower absenteeism, increased productivity, and controlled health care spending. However, despite the intuitive appeal and growing interest in worksite health promotion programs, empirical research examining their effectiveness and scalability in real-world settings has been limited to two important, recent papers, that both show disappointingly absence of benefits; Jones, Molitor and Reif (2019) and Song and Baicker (2019).

Our set-up is most like that of Song and Baicker (2019) who employ a social experiment that randomized at the work-unit level to study the effects of a workplace wellness program delivered to employees of a large US warehouse retail company. Their program comprised of eight modules, implemented over the course of 18 months. The modules focused on nutrition, physical activity, and stress reduction, with the average participant completing 1.3 modules. They show that the program resulted in significantly greater rates of some positive self-reported health behaviors but find no significant differences in clinical measures of health, health care spending and utilization, and employment outcomes after 18 months. Compared to our study population that primarily consisted of women, theirs was much more balanced in terms of sex (46% female across treatment and primary control worksites), about five years younger on average, and with clearly different job categories (retail vs. elderly care). Jones et al. (2019) use a social experiment with randomization at the individual level to study health promotion in combination with financial incentives tied to participation in activities, for employees at

University of Illinois at Urbana-Champaign. Their program consisted of an annual on-site biometric health screening and health risk assessment as well as a variety of wellness activities. Participation in the wellness activities was contingent upon completing the health screening and risk assessment. The most popular wellness activities were "HealthTrails", where participants virtually travel along famous trails; Tai Chi for Relaxation; Recess for Adults; and Stress Management. 56% of their participants completed both the health screening and the risk assessment and 27% (22%) completed the first (second) year activities. Their population were again different from ours in terms of the sex composition (57% female) and distinctly different in terms of job categories (20% were university faculty members and 44% were academic staff). Jones et al (2019) do not find significant causal effects of their two-year long program on total medical expenditures, health behaviors, employee productivity, or self-reported health status during the first 30 months after randomization. Relative to these papers, our short-term findings are more positive than those of the previous papers. Our work also contributes a long-run analysis based on complete register data.

The paper proceeds as follows: Section 2 presents the details of *Healthy-at-work* while Section 3 discusses available data and characterizes healthcare workers in terms of type of work, socio-economic background, and absence behavior. Section 4 shows results from our empirical analysis and Section 5 presents some cost-benefit considerations. Finally, Section 6 concludes.

2. Healthy at Work

2.1 The intervention and our experimental design

The *Healthy-at-Work* program was initiated by 110 employers across 11 municipalities that together constitute 10% of the Danish population. A total of 7,660 workers serving the elderly were enrolled in the project via their work-units. Work-units were groups of healthcare workers who cooperatively provided care within a specific context. This could mean serving a particular wing of a skilled nursing facility or jointly caring for a group of home-living elderly within a defined residential area.

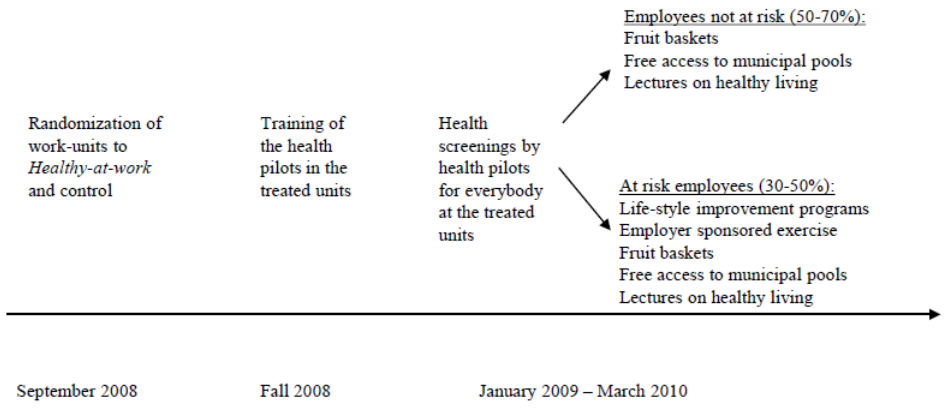
Randomization, conducted in September 2008, was applied at the level of the 314 participating work-units. Of these, 46% (equating to 144 work-units, each with an average of 28 workers) were allocated to the treatment group, with the remaining units functioning as controls. It is

noteworthy that *Healthy-at-Work* was generally inaccessible to employees in the control work-units. The cost of the program amounted to approximately DKK 8,500 (around €1,100) per employee at the treated work-units.

The program was devised with a dual managerial goal: to reduce employee absenteeism rates and retain workers through a comprehensive health promotion effort. The initiative aimed to help individuals develop and maintain a healthier lifestyle. We present a detailed timeline of the project and its components in Figure 1.

FIGURE 1

TIMELINE OF *HEALTHY-AT-WORK* AND ITS COMPONENTS



In the fall of 2008, the *Healthy-at-Work* program launched a recruitment campaign within the treatment work-units. The objective was to enlist local coordinators, or 'health pilots', for the program. Key to the intervention design was entrusting employees without managerial responsibilities with the support, implementation, and local adaptation of the health initiatives. By doing so, the program's architects aimed to boost participation rates among coworkers.

Project managers from the *Healthy-at-Work* team collaborated with local unit managers to select employees who demonstrated high motivation and enthusiasm to take on the health pilot role. The result was a corps of 352 health pilots, as larger work-units were allowed more than one coordinator.

These health pilots underwent training in two three-day, off-site seminars conducted by university-trained physiologists during fall and winter 2008. The seminars covered a range of lifestyle improvement programs, including weight control, smoking cessation, stress

management, nutrition education, and blood pressure measurement. Additionally, health pilots were introduced to various cardio-focused exercises.

After the seminars, the health pilots were allocated up to seven hours per week for duties related to the implementation and administration of *Healthy-at-Work*. These responsibilities included coordinating activities with local management and receiving supervision from the project managers. This strategy was rooted in the belief that direct involvement in the project's operational aspects would increase coworkers' participation rates.

Upon returning from the seminars, the health pilots commenced their first task: conducting a comprehensive physical and psychological health screening for their coworkers. Offered to all staff within the treatment work-units, the screening incorporated both physical and psychological elements.

The physical assessment included measures such as body mass index calculation, blood pressure evaluation, glucose level measurement, and fitness rating. Alongside these, the psychological aspect involved an interview with the health pilot and a self-administered online stress-assessment survey.

Based on the results of these assessments, employees presenting concerning indicators were classified as ‘high-risk’. These included a body mass index of 30 or more, a blood pressure above 140 (systolic) and 90 (diastolic), a glycaemia index at or above 5.8 mmol/L, and a fitness rating below 30. Additionally, employees who self-reported their health as poor or very poor, or those who frequently experienced stress based on the online survey, were also classified as high-risk. Notably, an employee needed to fail only one of these indicators to be categorized as high-risk.

These high-risk employees were then offered tailored lifestyle improvement programs, including individual level coaching and group activities within the high-risk group. A key feature of the Healthy-at-Work intervention was employer-sponsored exercise during working hours, offered for up to two hours per week for around 15 months. This element underscored the initiative's commitment to fostering long-term, healthy lifestyle changes among the workforce: some initiatives may initially alter behaviors and appear beneficial but not subsequently change habits when the novelty of the program wears off (e.g., Royer et al., 2015)

Furthermore, irrespective of their health status, all employees within the treatment work-units were granted additional amenities as part of the intervention. These included complimentary fruit at their workplace, free access to municipal pools, and invitations to lectures promoting healthy living. These measures aimed to create a healthier working environment and promote general wellbeing for everyone involved, not just the identified high-risk individuals. [3]

The *Healthy-at-Work* team acted as consultants, assisting the participating worksites in establishing and sustaining the program. However, the onus was on local management, in collaboration with the health pilots, to foster local leadership and oversee the selection, promotion, and scheduling of employees for the lifestyle improvement programs. An explicit aim of the study was to permit variation across sites in the handling of program content and facilities. This approach was designed to enhance the external validity of the results, catering to the diversity of real-world settings. While all sites adhered to the fundamental strategy of offering a blend of lifestyle improvement programs and exercise classes, they were permitted, even encouraged, to adapt the specifics of implementation to their unique circumstances.

In broader terms, our research represents an effectiveness or pragmatic study. We aimed not merely to test the efficacy of the intervention in an ideal or controlled environment, but rather to investigate its impact in real-world settings, with all their inherent complexities and variability.

As evaluators, it is vital to clarify our role in this study. We had no access to the health capital measures collected by the health pilots, nor were we privy to the details of individual eligibility. Furthermore, individual-level information regarding the classification of high-risk groups was intentionally not gathered by us. This approach was motivated by several factors.

Primarily, we aimed to maintain a minimal level of involvement to avoid influencing the treated units. Perceived monitoring could potentially act as an alternative form of treatment, one not available to the control group. In addition, keeping health data confidential between the employee and health pilot was deemed likely to encourage honesty in self-assessed health outcome measures.

[3] It is possible, of course, that employees in the control work-units could have taken advantage of the freely available fruit baskets at the workplace but we consider this component a minor part of the intervention.

Secondly, we could not replicate the same data collection process at the control worksites without introducing elements of the treatment components. Doing so would transform the control units from a 'business-as-usual' scenario to one influenced by the health screening, thereby altering the counterfactual, and hence the comparison point in our experiment. Any data collected would only have enabled a before-after assessment within participating units, a comparison we assessed as not valuable enough to justify the potential complications and the disruption of our experimental design. Instead, we chose to classify the high-risk group in our empirical analyses based on administrative data, a more non-intrusive method.

In our pursuit of understanding the intervention's consequences at the work-unit level—where colleagues interact—we again refrained from gathering individual-level information about program participation. Our understanding of the program's scope and uptake is primarily derived from the implementation study. It suggests that 30-50% of employees at treated work-units participated in the lifestyle improvement programs and employer-sponsored exercise (Andersen and Lauritzen, 2010).

Cross-referencing these findings with budgetary information (see appendix Table A1) suggests a similar participation rate in employer-sponsored exercise of around 30%. The total budget allocated €3.5 million to wage costs, excluding program management, equating to approximately €900 per employee in the treated work-units. At an average hourly wage of €23, this amount financed 15 months of employer-sponsored exercise for 30% of the employees at the treated work-units.

The implementation study also provides insights into program content, with 66% of the health pilots reporting they conducted individual health coaching. Other activities included Nordic pole walking (24%), fitness activities (43%), ball games (26%), and other types of physical exercise (68%).

The program was well received, believed not only to improve physical health but also workplace dynamics. Focus groups among health pilots and interviews with project managers indicated that *Healthy-at-Work* fostered better communication and collaboration among coworkers (Andersen and Lauritzen, 2010).

2.2 Conceptual framework: key behavioral changes and outcomes from Healthy-at-work

Our *Healthy-at-Work* program was an intricate intervention, targeting all employees in randomized work-units while dedicating extensive resources to high-risk individuals. This section will outline our projected theory of change and discuss the intervention's impact over different calendar time periods.

Health-Related Behaviors and Outcomes

The High-Risk Group

In the design of Healthy-at-Work, the high-risk group was targeted with the most comprehensive treatment. The early health screenings were intended to enhance health awareness and potentially reveal previously undetected illnesses. This might initially lead to increased health care uptake, especially for primary care consultations, which serve as gatekeepers to more specialized care. Yet, the health screenings could also substitute for some primary care visits. This dual possibility means that the net short-term impact on primary care uptake, particularly in the first year after randomization, is empirically uncertain.

Healthy-at-work also provided the high-risk group with lifestyle improvement programs and employer-sponsored exercise, in addition to lighter interventions such as fruit baskets, access to public pools, and lectures. The intent was that this comprehensive approach would foster healthier behaviors, increase social interaction among colleagues, and subsequently enhance physical and mental health. Accordingly, we anticipated seeing reductions in primary care usage and associated decreases in prescription drug purchase - the second study outcome - in the medium term, i.e., during the second year after randomization when the program is still ongoing.

Furthermore, if the program successfully promoted healthier behaviors, we could potentially see positive effects on more severe health outcomes such as hospital visits - the third study outcome. As health outcomes can respond in complex and sometimes immediate ways to changes in behavior and awareness, we remain open to observing such impacts at any point during our evaluation period, although some effects may be more likely to appear in the longer term.

It is important to remember that our program concluded within two years of the initial randomization. Therefore, any continued effects on health care utilization beyond this point could not be driven by direct treatment effects and would suggest sustained behavioral changes.

However, we also acknowledged the possibility of fade-out effects, where the impact of the program might diminish over time unless the healthier behaviors learned during the program were maintained.

The Low-Risk Group

The low-risk group's experience in *Healthy-at-Work* started with health screenings like those of the high-risk group. By design, this group had a lower probability of illness detection, yet it is plausible that some health issues were brought to light. These screenings could also substitute for some primary care visits.

After screenings, the low-risk group engaged in less intensive but still potentially impactful activities: receiving fruit baskets, gaining free access to public pools, and participating in health-focused lectures. The knowledge that their high-risk colleagues were undertaking more intensive health improvements could also influence their behavior.

Our hypothesis was that even these lighter-touch treatments could promote a culture of health awareness, which might translate into healthier behaviors. If this were to occur, we could expect to see a reduction in primary care use, prescription drug usage, and potentially even hospital visits. However, we anticipated that any such effects would likely be smaller and fade more quickly than in the high-risk group, given the less intensive nature of their intervention.

Work-Related Behaviors and Outcomes

Additionally, we turn our attention to work-related behaviors and outcomes. While these can be viewed as secondary in relation to the direct health-related outcomes of the intervention, they are of crucial importance from the perspectives of societal cost-benefit analyses and the employers themselves.

The health improvements and increased social interactions fostered by *Healthy-at-Work* could translate into lower employee absenteeism (our fourth study outcome), increased hours worked (the fifth study outcome), and a stronger tendency for employees to stay with the same employer (the sixth study outcome).

Furthermore, we also consider the potential impact on earnings (the seventh study outcome). On the one hand, improved health could boost productivity and working hours, thereby leading to higher earnings. On the other hand, the health promotion program could be perceived as a

valued fringe benefit, which might compensate for lower financial remuneration. Determining which of these scenarios dominates is ultimately an empirical question.

Exploratory Dimensions

Beyond health risk-based heterogeneity, we will explore potential spillover effects on spouses. If health awareness and behaviors transmit to partners, we might see mirrored effects in the partner population, although likely smaller and with quicker fade-out.

Lastly, we consider potential heterogeneity in the effects of the program. As this study was conducted before the practice of pre-registration of experimental protocols became common in economics, we did not have any pre-registered hypotheses for any aspects of this analysis. Consequently, some of our exploration of heterogeneity lacks the stringent structure of pre-committed analyses. To tackle this, we resort to machine learning techniques below following Chernozhukov et al. (2020). Their methods allow us to systematically incorporate the full set of baseline characteristics and pre-randomization measures of all outcomes, providing a structured approach to learning about heterogeneity in the absence of pre-specified hypotheses.

3. Data: characterizing health care workers

We use individual-level, register-based data from Statistics Denmark to investigate health care use and expenditures, employment variables, and socio-economic backgrounds. This data is meticulously linked using the central personal register number, a unique identifier for all Danish residents. We were able to successfully link 98% (7,541 out of 7,660) of the individuals in the experimental data to the administrative data. Furthermore, the experimental data connects individuals to their respective work-units around the time of randomization (September 2008). We track outcomes for these individuals over time, from the short run, immediately post-randomization, to the very long run, a decade after the intervention concludes. Our data allows for a 12-year follow-up on primary care physician consultations and work-related outcomes (available until 2020), prescription drug purchases can be traced until 11 years post-randomization (2019), and hospitalizations for 10 years (2018).

Table 1 presents the descriptive statistics for the estimation sample prior to randomization, split by treatment status. The statistics reveal that our sample employees were predominantly female (96%), with an average age of 43 years, and the majority being of Danish origin (93%). Most held a degree in health and welfare services and had, on average, one child.

For a perspective on external validity, we compare our intervention group with the broader population of Danish employees (Table 1, columns 6-7), and with workers in elderly care both within and outside our project municipalities (Table A2). Our project participants were notably more likely to be female, have children at home, and hold a health and welfare-related education. However, when compared to other workers in elderly care within and outside the project municipalities, these demographic differences are smaller.

TABLE 1

CHARACTERISTICS OF THE SAMPLE JUST PRIOR TO RANDOMIZATION

	Comparisons: Within-experiment					*Comparisons: To employed population*	
	(1) Mean (std. dev.) treatment group *N=3,735*	(2) Mean (std. dev.) control group *N=3,806*	(3) Std. diff. treatment vs. control group	(4) P-value, diff. treatment vs. control group	(5) Adj. P-value diff. treatment vs. control group	(6) Mean (std. dev.) Empl. population *N=2.65 million*	(7) Std. diff. intervention group vs. employed population
Demographic characteristics							
Age	43.2	42.9	0.024	0.307	1.000	40.049	0.234
	(12.0)	(12.2)				(13.8)	
Female, %	96%	96%	0.016	0.476	1.000	48%	1.258
Danish, %	93%	93%	-0.001	0.962	1.000	92%	0.027
Married, %	56%	57%	-0.013	0.565	1.000	50%	0.137
Number of children	0.982	0.974	0.006	0.778	1.000	0.997	-0.012
	(1.12)	(1.12)				(2.04)	
Years of schooling	12.9	12.7	0.101	<0.000	0.851	13.6	-0.327
	(2.0)	(2.2)				(2.7)	
Field of education (ISCED-based), %							
... Health and welfare	62%	61%	0.016	0.391	1.000	13%	1.099
... Business, administration, and law	10%	11%	-0.010	0.894	1.000	17%	-0.480
... Services	4%	3%	0.272	0.029	0.981	3%	-0.031
... Engineering, manufacturing, and construction	3%	3%	-0.081	0.569	1.000	18%	-1.191
Health care utilization, one year prior							
Any PCP visits	93%	93%	0.012	0.601	0.825	83%	0.290
PCP visits	8.197	8.135	0.008	0.727	0.824	5.500	0.370
	(8.0)	(7.5)				(6.3)	
Any prescription drug purchase	80%	82%	-0.045	0.051	0.123	66%	0.342
Pharmaceutical months supply	10.652	10.735	-0.005	0.832	0.856	7.585	0.191
	(17.4)	(16.3)				(15.7)	
Any hospital contact, psychiatric diagnosis	1%	1%	0.025	0.274	0.660	1%	0.025
Days hospitalized, psychiatric diagnosis	0.15	0.165	-0.004	0.857	0.851	0.094	0.019
	(3.3)	(4.0)				(3.0)	
Any hospital contact, somatic diagnosis	37%	36%	0.027	0.244	0.660	30%	0.128
Days hospitalized, somatic diagnosis	1.772	1.576	0.036	0.113	0.415	1.448	0.039
	(6.4)	(4.2)				(6.4)	

TABLE 1 CTD.

CHARACTERISTICS OF THE SAMPLE JUST PRIOR TO RANDOMIZATION

	Comparisons: Within-experiment					*Comparisons: To employed population*	
	(1) Mean (std. dev.) treatment group *N=3,735*	(2) Mean (std. dev.) control group *N=3,806*	(3) Std. diff. treatment vs. control group	(4) P-value, diff. treatment vs. control group	(5) Adj. P-value diff. treatment vs. control group	(6) Mean (std. dev.) Empl. population *N=2.65 million*	(7) Std. diff. intervention group vs. employed population
Employment outcomes, six months prior							
Any absenteeism, Jan-June 2008	81%	81%	-0.01	0.664	0.991	67%	0.326
% of scheduled hours missed, Jan-June 2008	6.85	6.97	-0.010	0.814	0.991	4.48	0.226
	(11.7)	(11.7)				(9.6)	
Any hours worked, Jan-June 2008	99%	99%	<0.000	0.995	0.994	98%	0.074
Hours worked, Jan-June 2008	785	772	0.023	0.294	0.956	809	-0.012
	(677)	(433)				(2,332)	
% working for same employer, Jan 2008	84%	82%	0.059	0.011	0.573	78%	0.124
Earnings Jan-June 2008, DKK	119,379	122,142	-0.057	0.014	0.826	153,566	-0.342
	(46,722)	(50,223)				(120,106)	
F-statistic, no strata indicators (p-value)	1.34 (0.134)						
F-statistic, with strata indicators (p-value)	1.17 (0.248)						

Note: This table compares the treatment group to the control group and the intervention group to the population of Danish employees. Outcomes related to health care utilization measured during the 12 months year prior to randomization; employment outcomes measured during the six months prior to randomization. *N* represents the number of observations, except in the case of absence. The number of observations for measures of absenteeism is based on 3,542 (95%) treated, 3,581 (94%) controls, and about 1.3 million (50%) in the background population. P-values stem from conventional t-tests; adjusted p-values are family-wise p-values that adjust for the number of variables in each family (in cursive) and are estimated using 1,000 bootstraps. F-statistic stems from regression of individual level pre-randomization outcomes and characteristics on an indicator for being in the treatment group with standard errors clustered at the work-unit level, versions with and without strata indicators.

In the lower part of Table 1, we present the pre-treatment characteristics of our participants, specifically their health seeking behavior and employment variables. Our participants had on average eight contacts with their primary care physicians in the calendar year leading up to randomization. About 80% of participants purchased prescription drugs during the calendar year prior to randomization. These include drugs commonly associated with conditions stemming from sedentary and unhealthy lifestyles, such as cardiovascular medications, psychotropics, anti-diabetics, and drugs for acid-related disorders, amongst others.[4] Finally, one-third had at least one hospital contact associated with a somatic diagnosis. As for employment, sickness absence was notably high, with workers missing on average 7% of scheduled hours, or about 1.4 days per month assuming a 20-day full-time work schedule. In comparison to the overall population of employees, our project participants used more health care services, had higher rates of absenteeism, and lower earnings. The differences between our intervention group and other workers in elderly care were much smaller, both when compared to employees in eldercare outside of and within the project municipalities, with the only exception that our population was somewhat more absent but also worked more hours.

Finally, Table 1 also shows standardized differences (column 3) along with results from conventional t-tests of differences in means (column 4). Additionally, family-wise p-values are provided to control for multiple hypothesis testing (column 5). Though some differences in means are statistically significant under conventional hypothesis testing, these differences do not hold up when adjusting for multiple hypothesis testing.

4. Results

In this section, we outline our statistical approach used to analyze the effects of the Healthy-at-work program, detailing the preferred regression model, considerations for treatment assignment, and methods for characterizing high-risk individuals.

Our main strategy regresses individual-level outcomes on an indicator for work-unit access to Healthy-at-work, using the following specification:

[4] The prescription drug data used in this study is not exhaustive due to confidentiality restrictions but still cover 82% of all pharmaceutical claims in 2008. The four most common excluded drug categories are ophthalmologicals (24%), nasal preparations (15%), antihistamines (12%), and cough and cold preparations (8%). The percentages in parentheses represent the share of the excluded categories in the 18% missing data.

$$(1) \qquad Y_i = \alpha + \beta D_i + \boldsymbol{\gamma X_i} + \varepsilon_i,$$

where Y is the outcome of interest, D is a treatment indicator, and X is a vector of control variables. The control variables include lagged values of all outcomes, strata (municipality) indicators, and specific individual-level background characteristics.[5] We interpret β as the effect of treatment provision at the work-unit level.

We cluster our standard errors at the work-unit level to account for the fact that randomization was performed, and treatment offered at this level. To allow for correlation of unobserved factors among workers employed with the same employers, we also show results that instead cluster at this level as a sensitivity analysis.

As outlined above, our intervention spanned from the initial randomization of work-units into treatment and control status in September 2008 until March 2010. Given this lengthy timeframe, real-world employment dynamics inevitably introduced movements between treated and control units, as well as employees leaving and joining participating employers. Recognizing these intricacies, we fix the treatment status based on the initial randomization month, effectively avoiding any bias stemming from post-randomization movements. However, this decision also leads to a potential dilution of our treatment. Some workers in treated units may not receive full exposure to *Healthy-at-work* and others initially assigned to control units may cross over into treated work-units, benefiting from the intervention. This mixture of exposure levels means our estimate of β should be interpreted as a conservative representation of the effect of the availability of a workplace offer such as *Healthy-at-work*. Nevertheless, it aligns with the real-world application of the *Healthy-at-work* program, ensuring our findings are both robust and reflective of practical implementation.

In our analysis, we examine an extensive set of outcomes across various domains, such as take-up of physician services, prescription drug purchases, hospitalizations, and labor market outcomes. With such a breadth of testing, the risk of Type I errors, or incorrectly rejecting one or more true null hypotheses, becomes a serious concern. To mitigate this risk and control for the family-wise error rate, we follow the method of Jones et al. (2019), employing family-wise adjusted p-values based on

[5] The individual-level background characteristics include an indicator for being male, an indicator for being of Danish origin, age indicators, an indicator for being married, number of children, years of schooling, and indicators for the type of schooling (within health care, clerical, food and nutrition, or craftsmanship). We present robustness analyses omitting the controls for lagged outcomes and individual level background characteristics in appendix Table A3.

1,000 bootstraps. We define four families of hypotheses based on the mentioned outcome domains and calculate the family-wise adjusted p-values within a given time period. This approach ensures that different time periods are treated as separate events or "families". Furthermore, we present results where all health-related outcomes are grouped into the same family.

Our primary analysis focuses on the work-unit level, evaluating the *Healthy-at-work* program as it would naturally be implemented in the real world. This approach considers the effects on all workers within the treated units, capturing both those who actively participated in the program and any potential spill-over effects that might influence non-participating coworkers. However, it could also be of interest to understand the specific impacts on those who actively engaged with the program, which requires a different level of analysis.

One straightforward way to estimate the impact on those who actively participated in the program is to reweigh our primary estimates by the fraction of participants in each work-unit; doing this would more than double the size of our estimates. This method, however, assumes away any spill-over effects within the work-units and may provide a conservative estimate. Two alternative approaches allow us to explore different aspects of our intervention. The first approach focuses on identifying the high-risk population, those offered the opportunity to exercise during working hours. Since we did not collect any biometric markers, we instead classify workers as high-risk based on their frequency of primary care visits prior to the intervention. This proxy method allows us to capture those more likely to participate in the program. The second approach leverages machine learning to explore treatment effect heterogeneity using the full set of baseline characteristics and pre-randomization measures following Chernozhukov et al. (2020). Together, these strategies allow us to delve deeper into the multifaceted impacts of the intervention, complementing our main analysis and offering additional insights, although we do want to note that our study is not strongly powered to detect such heterogeneity.

4.1 Health-related outcomes

Primary Care Physician Consultations

We start by analyzing outcomes concerning health behaviors that are closely linked to program content. Table 2, Panel A shows the results for any consultation with the primary care physician within a given period, while Panel B shows the results for the number of consultations per year. We

first present the results for the entire sample followed by the results for the high-risk and the low-risk groups and maintain this structure in the proceeding estimation tables. Treated individuals receive fewer services from their primary care physicians in the short run, namely the first year after randomization (0.5 visits). The size of the estimate is non-negligible; relative to the control group mean, number of contacts decrease by about 6 percent. Effects persist into the second year (.4 visits; 5 percent reduction relative to the control mean) in which the *Healthy-at-work* intervention was phased-out – and even remain substantial in size 3-7 years after randomization (0.4 visits; 4 percent reduction relative to control mean), i.e., long after the program has ended. Estimated effects are still negative but small and imprecise in the very long run, 8-12 years after randomization. Effects on the number of consultations also remain significant on a 5% significance level once we account for multiple hypothesis testing within the family of outcomes that cover primary care physician consultations during the first seven years after randomization.

TABLE 2

EFFECTS OF "HEALTHY-AT-WORK" ON CONSULTATIONS WITH PRIMARY CARE PHYSICIANS

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-12
Panel A. Any PCP consultation				
All *(N = 7,541)*	-0.006	0.001	0.001	0.001
	[0.006]	[0.006]	[0.002]	[0.004]
Adjusted *P* value	(0.346)	(0.855)	(0.501)	(0.772)
Mean outcome, control group	0.927	0.917	0.990	0.974
High-risk group *(N = 3,604)*	-0.004	-0.002	0.001	0.005
	[0.004]	[0.005]	[0.002]	[0.004]
Adjusted *P* value	(0.324)	(0.665)	(0.590)	(0.528)
Mean outcome, control group	0.983	0.977	0.995	0.978
Low-risk group *(N = 3,937)*	-0.010	0.003	0.001	-0.003
	[0.011]	[0.010]	[0.003]	[0.005]
Adjusted *P* value	(0.575)	(0.772)	(0.987)	(0.795)
Mean outcome, control group	0.875	0.861	0.986	0.970
Panel B. # PCP consultations per year				
All (*N = 7,541)*	**-0.502**	**-0.433**	**-0.365**	-0.142
	[0.132]	[0.149]	[0.124]	[0.144]
Adjusted *P* value	(0.000)	(0.008)	(0.014)	(0.566)
Mean outcome, control group	8.55	8.73	8.38	8.68
High-risk group *(N = 3,604)*	**-0.944**	**-1.041**	**-0.746**	-0.235
	[0.227]	[0.264]	[0.213]	[0.242]
Adjusted *P* value	(0.000)	(0.001)	(0.001)	(0.528)
Mean outcome, control group	12.13	12.25	11.42	11.22
Low-risk group *(N = 3,937)*	-0.064	0.118	-0.017	-0.083
	[0.156]	[0.177]	[0.138]	[0.166]
Adjusted *P* value	(0.687)	(0.737)	(0.987)	(0.795)
Mean outcome, control group	12.13	12.25	11.42	11.22

Note: This table shows estimates from regressions of outcomes on an indicator for being in the treatment group following the specification in Equation (1). Standard errors, reported in brackets, are clustered at the work-unit level. Italic indicates significance at the 10% level; bold indicates significance at 5% level. Family-wise p-values, reported in parentheses, adjust for the number of outcome variables in each family and are estimated using 1,000 bootstraps. Randomization occurred in September 2008. Full population consists of individuals employed at relevant work-units in the month just prior to randomization; the high-risk group consists of individuals from the full population with above median number of consultations with the primary care physician in the six months leading up to randomization, while the number of consultations in low-risk group are at or below the median.

Effects are seen on the intensive but not the extensive margin, however. This has two important implications: first, individuals do not discontinue their interaction with their primary care physician because of program participation and 2) the effect on number of consultations must be driven by individuals who interact *more* with their primary care physician (and who are then presumably in worse health), our proxy for being in the high-risk group. In line with this, we find even stronger effects on the intensive margin for individuals with elevated pre-randomization levels of primary care physician consultations, or the high-risk group. These results also survive the correction for multiple hypothesis testing.

Our baseline model clusters standard errors at the work-unit level but in most cases, this matters little compared to not clustering at all. As shown in Table A3 moreover, results are robust to clustering at the employer level as well as to omitting individual level characteristics, strata indicators, and pre-randomization outcomes.[6]

Prescription Drug Purchases

Table 3 investigates effects on prescription drug purchases. We consider indicators for any purchase as well as months' supply of drugs targeted either hypertension, cholesterol-lowering drugs, diabetes drugs, severe pain relief, or depression. We find reductions in the long run, 3-7 years after randomization (.76 months; 5 percent reduction relative to the control group mean), and for the high-risk group even already in the medium run, namely the second year after randomization (1 month; 6 percent of the control group mean) and 3-7 years after (1.5 months; 7 percent of the control group mean). All these estimates are significant at least at the 10% level when correcting for multiple hypothesis testing. Like the results for physician consultations, both effects survive in the long run but fade out in the very long run.[7]

[6] All subsequent conclusions are robust to these specification changes as well and results available upon request.
[7] Granular analysis (not shown) indicates short-run effects on purchases of drugs targeted hypertension, depression, and diabetes and more persistent effects on drugs targeted pain relief. These estimates do not survive correction for multiple hypothesis testing, however, and should be interpreted with caution.

TABLE 3

EFFECTS OF "HEALTHY-AT-WORK" ON PURCHASES OF PRESCRIPTION DRUGS

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-11
Panel A. Any prescription drug purchase				
All *(N = 7,541)*	-0.012	0.007	-0.002	-0.005
	[0.008]	[0.008]	[0.004]	[0.005]
Adjusted *P* value	(0.289)	(0.453)	(0.647)	(0.321)
Mean outcome, control group	0.826	0.819	0.968	0.947
High-risk group *(N = 3,604)*	-0.012	-0.012	0.004	0.001
	[0.008]	[0.009]	[0.003]	[0.005]
Adjusted *P* value	(0.232)	(0.192)	(0.161)	(0.817)
Mean outcome, control group	0.930	0.926	0.989	0.969
Low-risk group *(N = 3,937)*	-0.012	*0.025*	-0.009	-0.012
	[0.013]	[0.012]	[0.007]	[0.008]
Adjusted *P* value	(0.564)	(0.105)	(0.399)	(0.279)
Mean outcome, control group	0.728	0.926	0.989	0.969
Panel B. Months' supply per year				
All *(N = 7,541)*	-0.148	-0.311	**-0.759**	*-0.781*
	[0.186]	[0.266]	[0.357]	[0.482]
Adjusted *P* value	(0.690)	(0.346)	(0.041)	(0.197)
Mean outcome, control group	11.88	12.83	15.27	19.04
High-risk group *(N = 3,604)*	-0.242	**-1.025**	**-1.497**	-1.303
	[0.312]	[0.433]	[0.574]	[0.814]
Adjusted *P* value	(0.454)	(0.051)	(0.036)	(0.238)
Mean outcome, control group	16.84	18.27	21.04	25.10
Low-risk group *(N = 3,937)*	-0.079	0.248	-0.183	-0.432
	[0.218]	[0.344]	[0.399]	[0.530]
Adjusted *P* value	(0.706)	(0.488)	(0.681)	(0.439)
Mean outcome, control group	7.21	7.71	9.82	13.33

Notes: This table shows coefficient estimates regressions of outcomes on an indicator for being in the treatment group following the specification in Equation (1). Prescriptions include drugs targeted hypertension (ATC-category C9), cholesterol-lowering drugs (C10), diabetes drugs (A10), severe pain relief (N2), and depression (N05A, N05B, N05C, and N06A). Standard errors, reported in brackets, are clustered at the work-unit level. Italic indicates significance at the 10% level; bold indicates significance at 5% level. Family-wise p-values, reported in parentheses, adjust for the number of outcome variables in each family and are estimated using 1,000 bootstraps. Randomization occurred in September 2008. Full population consists of individuals employed at relevant work-units in the month just prior to randomization; the high-risk group consists of individuals from the full population with above median number of consultations with the primary care physician in the six months leading up to randomization, while the number of consultations in low-risk group are at or below the median.

Hospitalization Outcomes

We also analyze hospitalization outcomes in Table 4. Overall, we see very little in terms of this type of outcome level when correcting for multiple hypothesis testing; we observe some very early declines in hospitalizations associated with psychiatric diagnoses, of which the bulk are due to stress and depression, but effects die out in the second year after randomization. We also detect early decreases in hospitalizations due to somatic disease for the high-risk group and corresponding increases for the low-high group but both effects subsequently disappear.

Connecting Health Utilization to Health Outcomes

Our findings have revealed consistent reductions in consultations with primary care physicians and prescription drug purchases, both in the short and long run. While these reductions indicate changes in healthcare utilization, they do not necessarily translate directly into health improvements. Hospitalizations, albeit a crude measure, provide a more direct insight into health outcomes. Our analysis in this regard shows limited effects, with some early declines in psychiatric-related hospitalizations and transient changes in somatic disease hospitalizations for different risk groups.

TABLE 4

EFFECTS OF "HEALTHY-AT-WORK" ON DIAGNOSES

ASSOCIATED WITH HOSPITALIZATIONS

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-10
Panel A. Any hospital contact, psychiatric diagnosis				
All *(N = 7,541)*	**-0.007**	0.0002	-0.005	-0.002
	[0.003]	[0.003]	[0.005]	[0.005]
Adjusted *P* value	(0.057)	(0.952)	(0.749)	(0.932)
Mean outcome, control group	0.017	0.014	0.056	0.041
High-risk group *(N = 3,604)*	**-0.013**	-0.003	-0.013	-0.001
	[0.005]	[0.004]	[0.009]	[0.007]
Adjusted *P* value	(0.036)	(0.766)	(0.504)	(0.984)
Mean outcome, control group	0.029	0.023	0.083	0.055
Low-risk group *(N = 3,937)*	-0.0003	0.003	0.003	-0.002
	[0.002]	[0.002]	[0.005]	[0.006]
Adjusted *P* value	(0.909)	(0.620)	(0.791)	(0.955)
Mean outcome, control group	0.006	0.006	0.030	0.028
Panel B. Number of days admitted to hospital, psychiatric diagnosis				
All *(N = 7,541)*	0.002	-0.172	-0.044	**-0.142**
	[0.153]	[0.176]	[0.098]	[0.063]
Adjusted *P* value	(0.995)	(0.788)	(0.749)	(0.113)
Mean outcome, control group	0.360	0.471	0.451	0.338
High-risk group *(N = 3,604)*	-0.070	-0.395	-0.166	**-0.278**
	[0.301]	[0.361]	[0.175]	[0.122]
Adjusted *P* value	(0.827)	(0.713)	(0.747)	(0.119)
Mean outcome, control group	0.617	0.880	0.688	0.504
Low-risk group *(N = 3,937)*	0.056	0.075	0.066	-0.026
	[0.092]	[0.080]	[0.104]	[0.054]
Adjusted *P* value	(0.907)	(0.631)	(0.791)	(0.955)
Mean outcome, control group	0.118	0.086	0.228	0.181

TABLE 4 CTD.

EFFECTS OF "HEALTHY-AT-WORK" ON DIAGNOSES ASSOCIATED WITH HOSPITALIZATIONS

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-10
Panel C. Any hospital contact, somatic diagnosis				
All *(N = 7,541)*	-0.001	0.011	0.012	-0.0005
	[0.010]	[0.012]	[0.009]	[0.010]
Adjusted *P* value	(0.995)	(0.788)	(0.526)	(0.952)
Mean outcome, control group	0.408	0.428	0.838	0.797
High-risk group *(N = 3,604)*	**-0.037**	0.005	0.004	0.002
	[0.015]	[0.017]	[0.011]	[0.012]
Adjusted *P* value	(0.052)	(0.793)	(0.876)	(0.984)
Mean outcome, control group	0.522	0.513	0.894	0.831
Low-risk group *(N = 3,937)*	**0.035**	0.018	0.02	-0.003
	[0.014]	[0.016]	[0.013]	[0.014]
Adjusted *P* value	(0.077)	(0.620)	(0.301)	(0.955)
Mean outcome, control group	0.300	0.348	0.785	0.766
Panel D. Number of days admitted to hospital, somatic diagnosis				
All *(N = 7,541)*	-0.222	-0.076	0.073	-0.116
	[0.172]	[0.132]	[0.100]	[0.101]
Adjusted *P* value	(0.516)	(0.833)	(0.749)	(0.604)
Mean outcome, control group	2.55	2.25	2.57	2.30
High-risk group *(N = 3,604)*	-0.515	-0.188	-0.078	-0.144
	[0.317]	[0.203]	[0.158]	[0.151]
Adjusted *P* value	(0.221)	(0.729)	(0.876)	(0.747)
Mean outcome, control group	3.57	2.81	3.24	2.70
Low-risk group *(N = 3,937)*	0.069	0.034	*0.204*	-0.120
	[0.172]	[0.165]	[0.116]	[0.121]
Adjusted *P* value	(0.909)	(0.850)	(0.283)	(0.803)
Mean outcome, control group	1.80	1.73	1.94	1.93

Notes: This table shows coefficient estimates from regressions of outcomes on an indicator for being in the treatment group following the specification in Equation (1). Standard errors, reported in brackets, are clustered at the work-unit level. Italic indicates significance at the 10% level; bold indicates significance at 5% level. Family-wise p-values, reported in parentheses, adjust for the number of outcome variables in each family and are estimated using 1,000 bootstraps. Randomization occurred in September 2008. Full population consists of individuals employed at relevant work-units in the month just prior to randomization; the high-risk group consists of individuals from the full population with above median number of consultations with the primary care physician in the six months leading up to randomization, while the number of consultations in low-risk group are at or below the median.

4.2. Work-related outcomes

Absence from Work

Our analysis begins with an examination of the effects of *Healthy-at-work* on work absence, see Table 5, Panels A and B. Surprisingly, given the health use findings, estimated effects on both any absence and yearly absence rate are small and statistically insignificant across all time horizons. This apparent inconsistency can be partially explained by the nature of absence in healthcare work, as an anonymous survey (FOA, 2010) suggests that short-term absences are mainly driven by infections (68%), musculoskeletal system disease (13%), and stress (4%). Our intervention's impact would likely relate more to the latter and less common types of absence.

Hours Worked, Employer Continuity, and Earnings

We further investigated our intervention's effects on other work-related outcomes, including hours worked, the propensity to continue working with the same employer, and earnings. The results, presented in Table 5, Panels C-F, reveal no significant impact on any of these outcomes. While a few estimates (mostly related to hours worked) show statistical significance at conventional levels, they are small in magnitude and do not withstand correction for multiple hypothesis testing. Importantly, the lack of effects on earnings is not due to perfect predictability by factors like education and tenure. Our extended Mincer-type earnings regressions show an R^2 ranging from 0.35 to 0.58 depending on the time horizon, leaving substantial room for the intervention to affect wage outcomes.

TABLE 5

EFFECTS OF "HEALTHY-AT-WORK" ON LABOR OUTCOMES

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-12
Panel A. Any absence | hours > 0				
All	-0.006	0.015	-0.005	0.004
	[0.009]	[0.010]	[0.006]	[0.007]
Adjusted *P* value	(0.878)	(0.452)	(0.738)	(0.675)
Mean outcome, control group	0.834	0.813	0.939	0.930
N	7,043	6,585	6,532	5,307
High-risk group	-0.001	0.016	0.002	0.006
	[0.012]	[0.012]	[0.009]	[0.008]
Adjusted *P* value	(0.983)	(0.715)	(0.934)	(0.473)
Mean outcome, control group	0.871	0.848	0.947	0.941
N	3,286	3,023	2,999	2,371
Low-risk group	-0.008	0.013	-0.011	0.001
	[0.012]	[0.014]	[0.009]	[0.010]
Adjusted *P* value	(0.970)	(0.616)	(0.604)	(0.992)
Mean outcome, control group	0.802	0.783	0.932	0.920
N	3,757	3,562	3,533	2,936
Panel B. Yearly absence rate | hours > 0				
All	-0.193	0.392	0.313	0.708
	[0.355]	[0.391]	[0.361]	[0.436]
Adjusted *P* value	(0.908)	(0.792)	(0.884)	(0.560)
Mean outcome, control group	8.39	7.81	8.67	9.14
N	7,043	6,585	6,532	5,307
High-risk group	-0.595	-0.020	0.441	0.780
	[0.633]	[0.614]	[0.665]	[0.732]
Adjusted *P* value	(0.816)	(0.965)	(0.738)	(0.473)
Mean outcome, control group	11.02	9.87	10.86	10.69
N	3,286	3,023	2,999	2,371
Low-risk group	0.225	*0.821*	0.229	0.656
	[0.362]	[0.472]	[0.328]	[0.508]
Adjusted *P* value	(0.976)	(0.390)	(0.937)	(0.892)
Mean outcome, control group	6.02	6.00	6.76	7.85
N	3,757	3,562	3,533	2,936

TABLE 5 CTD.

EFFECTS OF "HEALTHY-AT-WORK" ON LABOR OUTCOMES

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-12
Panel C. Any hours				
All *(N = 7,541)*	0.001	*0.009*	0.003	*0.019*
	[0.003]	[0.005]	[0.006]	[0.010]
Adjusted *P* value	(0.918)	(0.319)	(0.884)	(0.259)
Mean outcome, control group	0.988	0.940	0.931	0.783
High-risk group *(N = 3,604)*	0.001	0.007	0.009	**0.034**
	[0.004]	[0.008]	[0.009]	[0.014]
Adjusted *P* value	(0.983)	(0.806)	(0.700)	(0.128)
Mean outcome, control group	0.985	0.925	0.915	0.780
Low-risk group *(N = 3,937)*	0.002	*0.010*	-0.003	0.003
	[0.003]	[0.006]	[0.007]	[0.011]
Adjusted *P* value	(0.973)	(0.390)	(0.937)	(0.992)
Mean outcome, control group	0.992	0.954	0.946	0.806
Panel D. Hours worked per year				
All *(N = 7,541)*	1.1	7.7	8.5	18.6
	[10.6]	[12.9]	[14.9]	[17.7]
Adjusted *P* value	(0.908)	(0.792)	(0.884)	(0.578)
Mean outcome, control group	1,513	1,420	1,255	1,072
High-risk group *(N = 3,604)*	-6.8	0.3	29.3	*41.8*
	[14.9]	[19.6]	[23.7]	[24.7]
Adjusted *P* value	(0.983)	(0.994)	(0.700)	(0.313)
Mean outcome, control group	1,475	1,362	1,166	992
Low-risk group *(N = 3,937)*	4.8	11.6	-14.2	-5.9
	[12.0]	[16.2]	[17.5]	[19.3]
Adjusted *P* value	(0.976)	(0.804)	(0.874)	(0.992)
Mean outcome, control group	1,549	1,475	1,338	1,147

TABLE 5 CTD.

EFFECTS OF "HEALTHY-AT-WORK" ON LABOR OUTCOMES

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-12
Panel E. Continue employment with same employer				
All *(N = 7,541)*	0.013	0.017	0.019	0.018
	[0.011]	[0.015]	[0.015]	[0.013]
Adjusted *P* value	(0.756)	(0.651)	(0.738)	(0.560)
Mean outcome, control group	0.803	0.701	0.481	0.346
High-risk group *(N = 3,604)*	0.022	0.026	0.026	*0.032*
	[0.015]	[0.021]	[0.021]	[0.017]
Adjusted *P* value	(0.583)	(0.708)	(0.700)	(0.283)
Mean outcome, control group	0.767	0.661	0.428	0.302
Low-risk group *(N = 3,937)*	0.005	0.008	0.009	0.005
	[0.013]	[0.016]	[0.017]	[0.016]
Adjusted *P* value	(0.976)	(0.868)	(0.937)	(0.992)
Mean outcome, control group	0.836	0.756	0.530	0.386
Panel F. Earnings per year				
All *(N = 7,541)*	370	608	940	1808
	[1,728]	[1,933]	[2,244]	[3,006]
Adjusted *P* value	(0.967)	(0.792)	(0.884)	(0.675)
Mean outcome, control group	251,065	242,707	210,534	192,712
High-risk group *(N = 3,604)*	173	274	4,905	6,507
	[2,538]	[3,237]	[3,524]	[4,437]
Adjusted *P* value	(0.983)	(0.994)	(0.612)	(0.394)
Mean outcome, control group	240,679	229,017	192,935	175,528
Low-risk group *(N = 3,937)*	1,004	776	-2,972	-3,025
	[1,932]	[2,600]	[2,925]	[3,410]
Adjusted *P* value	(0.976)	(0.868)	(0.795)	(0.892)
Mean outcome, control group	260,857	225,614	227,127	208,913

Notes: This table shows coefficient estimates from regressions of regressions of outcomes on an indicator for being in the treatment group following the specification in Equation (1). Standard errors, reported in brackets, are clustered at the work-unit level. Italic indicates significance at the 10% level; bold indicates significance at 5% level. None of the estimates are significant when relying on family-wise p-values that adjust for the number of outcome variables in each family. Randomization occurred in September 2008. Full population consists of individuals employed at relevant work-units in the month just prior to randomization; the high-risk group consists of individuals from the full population with above median number of consultations with the primary care physician in the six months leading up to randomization, while the number of consultations in low-risk group are at or below the median.

4.3 Key sensitivity analyses

We conduct a set of sensitivity analyses to assess the robustness of our results, particularly focusing on consultations with primary care physicians and purchases of prescription drugs, where we detected significant effects.

P-value Corrections

Our analyses considered varying corrections for multiple hypothesis testing, including treating physician services and prescription drug purchases as separate or combined families (Appendix Table A4). While p-values generally increase when more outcomes are allowed within a particular family, most conclusions remain unchanged, indicating robust findings. Only one conclusion shifts from significant at the 5%-level to non-significant when adopting the most conservative specification.

Heterogeneity Considerations

Using algorithmic model selection, we explored systematic treatment effect heterogeneity, investigating whether different indicators of prior health risks yielded similar conclusions. Our approach, following Chernozhukov et al. (2020), primarily focused on consultations with primary care physicians, where *Healthy-at-work* had the most substantial effects. While some indications of effect heterogeneity emerged, differences were not statistically significant. Importantly, our conclusions regarding the high- and low-risk groups[8] appear robust to using alternative health proxies. For a more detailed explanation of these analyses, including the methodology and results, we refer the reader to Appendix B.

4.4 Spousal responses

We now turn our attention to the area of spousal health use responses to comprehensive health programs, an aspect that, to the best of our knowledge, has not been previously explored in the literature. While earlier research has examined positive partner reactions to health interventions related to drinking and smoking (Fletcher and Marksteiner, 2017), our study presents the first evidence of social spillovers from a multifaceted health program like *Healthy-at-work*. This examination is not only an intriguing addition to existing research but also carries policy implications, such as the potential improvements in cost-effectiveness when spousal gains are considered (Fletcher

[8] I.e., the most and least affected quintiles of the proxy predictor.

and Marksteiner, 2017), and the influence of partners' health shocks on individual health behaviors (Fadlon and Nielsen, 2019).

Table 6 presents the effects on spousal health use outcomes, using a specification akin to Equation (1). It is important to recognize that 96% of the Healthy-at-work main estimation sample is female, rendering the spousal population predominantly male.

TABLE 6

EFFECTS OF "HEALTHY-AT-WORK" ON SPOUSAL OUTCOMES

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-12
Panel A. Any PCP consultation				
All *(N = 5,430)*	**-0.023**	-0.0002	-0.001	-0.0001
	[0.010]	[0.010]	[0.007]	[0.007]
Adjusted *P* value	(0.040)	(0.985)	(0.855)	(1.000)
Mean outcome, control group	0.772	0.754	0.915	0.901
Panel B. # PCP consultations per year				
All *(N = 5,430)*	-0.045	-0.164	-0.063	0.007
	[0.123]	[0.149]	[0.132]	[0.169]
Adjusted *P* value	(0.741)	(0.466)	(0.855)	(0.997)
Mean outcome, control group	5.08	5.45	5.42	6.1
Panel C. Any prescription drug purchase				
All *(N = 5,430)*	*-0.019*	*-0.020*	-0.007	0.002
	[0.011]	[0.012]	[0.008]	[0.010]
Adjusted *P* value	(0.163)	(0.198)	(0.657)	(0.872)
Mean outcome, control group	0.619	0.623	0.863	0.836
Panel D. Months' supply per year				
All *(N = 5,430)*	-0.119	-0.275	-0.269	0.382
	[0.238]	[0.349]	[0.482]	[0.741]
Adjusted *P* value	(0.637)	(0.436)	(0.657)	(0.850)
Mean outcome, control group	11.0	12.5	15.6	20.3

Notes: This table shows coefficient estimates from regressions of outcomes on an indicator for being in the treatment group following the specification in Equation (1). Population consists of individuals either married to or cohabiting with individuals employed at relevant work-units in the month just prior to randomization. Standard errors, reported in brackets, are clustered at the work-unit level. Italic indicates significance at the 10% level; bold indicates significance at 5% level. Family-wise p-values, reported in parentheses, adjust for the number of outcome variables in each family and are estimated using 1,000 bootstraps. Randomization occurred in September 2008.

Intriguingly, treated spouses reduce the incidence of consultations with their primary care physician by 2.3 percentage points after randomization, relative to a mean of 77% in the control population. This estimate is statistically significant even when accounting for multiple hypothesis testing**.** The effects on prescription drug purchases are significant at the 10%-level in the initial two years after randomization. Representing a reduction of 2 percentage points compared to a mean of 62%, these findings are noteworthy in the short run but lose significance under multiple hypothesis testing. Moreover, the estimated effects in both areas diminish and lose significance two years post-randomization and remain insignificant in the longer term. Thus, *Healthy-at-work* seems to create temporary spillover effects on health behaviors. As indicated earlier, the primary effects' larger magnitude and more prolonged persistence compared to spillover effects is not unexpected.

5. Cost-Benefit Analysis of *Healthy-at-work* intervention

Having explored the effects of Healthy-at-work on various health outcomes, we now turn to an essential question: Do the benefits of participation outweigh the costs? To answer this, we measure the net social return by subtracting the (discounted) costs of *Healthy-at-work* from the discounted stream of benefits, following the evaluation literature's dominant approach (see e.g., Heckman et al., 1999).

Savings from Health Care Use

The first component of the net social benefits consists of the present discounted value of the estimated saved expenditures from the reduction in primary care physician consultations. We calculate this using individual level register-based information on the fee-for-services to primary care physicians for all consultations. Similarly, we include the saved expenditures on prescription drug purchases, relying on individual register-based data for the associated costs within the categories analyzed.

Costs and Other Considerations

From the calculated savings, we deduct the per participant expenditures from *Healthy-at-work*, as detailed in Section 2.1. We also present versions accounting for savings related to spouses' health use and show calculations using annual discount rates of both 3% and 6%. It is essential to note that our

calculation does not include potential utility gains beyond purchase price reductions, changes in other types of drug usage, or differences in leisure time value due to program participation.

TABLE 7

NET BENEFITS OF *HEALTHY-AT-WORK,*

PER EMPLOYEE AT PARTICIPATING WORK-UNITS

	NPV, 3% discount rate Estimate, DKK (Std. error)	NPV, 6% discount rate Estimate, DKK (Std. error)
A. Focal individuals		
Expenditures, contacts to primary care physicians	-299	-262
	(124)	(105)
Expenditures, prescription drug purchases	-622	-484
	(480)	(400)
Total expenditures, own health use	-921	-746
	(532)	(443)
Net benefit	-7,579	-7,754
	(532)	(443)
B. Focal individuals and spouses		
Total expenditures, own and spouse health use	-1,192	-1,008
	(856)	(722)
Net benefit	-7,308	-7,492
	(856)	(722)

Notes: NPV indicates net present value. All values are stated in DKK deflated to 2008 using the GDP deflator. Expenditures associated with contacts to primary care physicians are based on individual level fee-for-services starting after randomization and ending seven years later. Expenditures associated with prescription drug purchases are based on actual, induvial level prices in the same, seven-year period.

Results

The results, shown in Table 7, indicate that the net present value of saved expenditures associated with personal health care use is around DKK 900 (€120) with a 3% discount rate, increasing to about DKK 1,200 (€160) when including spouses' health care use. However, these savings only cover 14% (12% with a 6% discount rate) of the program's expenditures, resulting in clear negative net benefits. This contrasts sharply with returns on investment reported for the most rigorous studies included in Baicker, Cutler, and Song (2010), which range between 2.7 and 3.3. Note though, that less than half of the studies included in Baicker et al. (2010) had access to pre-intervention data and relied on randomization or used a matched comparison group in their analysis.

In conclusion, while the *Healthy-at-work* program has shown some effects on health outcomes, the cost-benefit analysis reveals that the intervention's economic return is limited, representing a key consideration for future policy development.

6. Conclusion

This paper presents the evaluation of a comprehensive employer-sponsored health and well-being program, '*Healthy-at-work'*, conducted through a social experiment involving over 7,500 healthcare workers, more than 100 employers, and over 300 work-units in Denmark. The experiment ran for almost two years, from randomization at the work-unit level in September 2008 to the end of the program in March 2010. The components of the program included health screenings with continued individual health coaching, physical activities with coworkers during work hours, and shorter courses on promoting healthy living, targeting key employees. We successfully linked experimental protocols with administrative data for 98% of the participants (7,541 out of 7,660 workers), thereby overcoming a common challenge of attrition often faced in similar studies. The current study assesses a range of outcomes, from consultations with primary care physicians to absenteeism and turnover, with a focus on both contemporaneous and long-run effects.

We find several important outcomes from the 'Healthy-at-work' program. Participants randomized into the program experience a reduction in consultations with their primary care physicians, especially among those in worse health prior to randomization. The effects are not merely offset by the program's health screenings and physical activities but reflect genuine reductions in health care utilization and prescription drug purchases long after the program's conclusion. Interestingly, the current study also uncovers temporary social spillovers to spouses' health-seeking behaviors. However, we discover no long-run effects on hospital admissions and no evidence of success in the primary managerial goals of the intervention, such as reducing absenteeism and turnover or affecting hourly wages or hours worked.

Despite these positive aspects of the program, our cost-benefit considerations yield clear and disappointing insights into the economic viability of 'Healthy-at-work'. In fact, the gains associated with the program do not exceed the costs of providing it. Put differently, the net present value of expenditures associated with health care use and savings represents only a small fraction of the expenditures necessary to operate the program, yielding negative net benefits.

Compared to the small existing literature on programs that operate at scale in real-world settings, our findings present a more optimistic view, indicating reductions in healthcare utilization with broader implications for both the work environment and general quality of life. These variations in results can be attributed to differences in program content and delivery, diverse characteristics of program recipients, and data availability. Unique to our study is the focus on healthcare workers, who likely possess greater health-related knowledge. Thus, our program's main function seems to be removing barriers to healthy lifestyles rather than merely providing information, potentially making our estimates a conservative indication of what might be achieved in other sectors. This uniqueness, however, may also make our study population more responsive to the intervention due to their healthcare training and higher susceptibility to absenteeism and burnout, possibly magnifying the program’s impact.

Literature

Andersen, H. L. & H. B. Lauritzen (2010): "Implementation and effects of Healthy-at-Work" (Implementering og effekter af Rask-i-job), AKF working paper.

Aldana, S. G. & N. P. Pronk (2001): "Health Promotion Programs, Modifiable Health Risks, and Employee Absenteeism", *Journal of Occupational and Environmental Medicine* 43(1), p. 36-46.

Baicker, K., D. Cutler, & Z. Song (2010): "Workplace wellness programs can generate savings", *Health Affairs* 29(2), 304-11.

Charness, G., & U. Gneezy (2009): "Incentives to Exercise", *Econometrica* 77(3), p. 909-931.

Chernozhukov, V., M. Demirer, E. Duflo, & I. Fernández-Val (2020): "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." Working paper 24678, National Bureau of Economic Research.

Dishman, R.K., B. Oldenburg, H. O'Neal, & R. J. Shephard (1998): "Worksite Physical Activity Interventions", *American Journal of Preventive Medicine* 15(4), p. 344-61.

Fadlon, I. & T. H. Nielsen (2019): "Family Health Behaviors", *American Economic Review* 109(9), p. 3162-91.

Fletcher, J. & R. Marksteiner (2017): "Causal Spousal Health Spillover Effects and Implications for Program Evaluation", *American Economic Journal: Economic Policy* 9(4), p. 144–166.

Harrad, R. & Sulla, F. (2018): "Factors associated with and impact of burnout in nursing and residential home care workers for the elderly", *Acta biomedica: Atenei Parmensis* 89(7-S), p. 60–69.

Heckman, J., LaLonde, R., & Smith, J., 1999. The economics and econometrics of active labor market programs. In: Ashenfelter, A., Card, D. (Eds.), Handbook of Labor Economics, vol. 3. Elsevier Science, Amsterdam, pp. 1865–2097.

Jones, D., D. Molitor, & J. Reif (2019): "What do workplace wellness programs do? Evidence from the Illinois Workplace Wellness Study", *Quarterly Journal of Economics* 134 (4), p. 1747-1791.

Rongen, A., S. J. W. Robroek, F. J van Lenthe, & A. Burdorf (2013): "Workplace Health Promotion: A Meta-Analysis of Effectiveness", *American Journal of Preventive Medicine* 44(4), p. 406-415.

Royer, H., M. Stehr & J. Sydnor (2015) "Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company", *American Economic Journal: Applied Economics* 7(3), p. 51–84

Song, Z. & C. Baicker (2019): "Effect of a WorkplaceWellness Program on Employee Health and Economic Outcomes. A Randomized Clinical Trial", *JAMA* 321(15), p. 1491-1501.

U.S. Bureau of Labor Statistics (2019): "2018 survey of Occupational Injuries & Illnesses", Charts Package.

Westfall, Peter H. & S. Stanley Young (1993): "Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment", Hoboken, NJ: John Wiley & Sons.

WHO (2018): "Noncommunicable diseases", Fact sheet, accessed November 8, 2019. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Appendix A. Figures and tables

TABLE A1

BUDGET FOR *HEALTHY-AT-WORK*, €

Budget posts	
Wage costs, project management	80,066
Other wage costs	3,510,400
Administration	20,000
Courses	42,057
Dissemination	15,429
Evaluation	80,000
Accounting	11,429
Consultants	131,867
Medical treatment	335,776
Other costs	24,000
Sum	4,251,023

Note: This table shows budget numbers that stem from the original application to the Prevention Fund

TABLE A2

EXTERNAL VALIDITY, SAMPLE JUST PRIOR TO RANDOMIZATION

	Within-experiment		*Comparisons: To eldercare workers in participating municipalities*		*Comparisons: To eldercare workers in non-participating municipalities*	
	Mean, treatment group *N=3,735*	Mean, control group *N=3,806*	Other eldercare workers *N=7,945*	Std. diff. intervention group vs. other eldercare workers	Other eldercare workers *N=107,292*	Std. diff. intervention group vs. other eldercare workers
Demographic characteristics						
Age	43.2	42.9	41.599	0.149	42.335	0.092
	(12.0)	(12.2)	(13.2)		(13.0)	
Female, %	96%	96%	0.923	0.180	92%	0.202
Danish, %	93%	93%	0.928	0.008	91%	0.076
Married, %	56%	57%	0.511	0.120	52%	0.104
Number of children	0.982	0.974	0.911	0.050	0.915	0.040
	(1.12)	(1.12)	(1.21)		(1.53)	
Years of schooling	12.9	12.7	12.9	-0.052	12.9	-0.030
	(2.0)	(2.2)	(2.1)		(2.1)	
Field of education (ISCED-based), %						
... Health and welfare	62%	61%	50%	0.223	52%	0.177
... Business, administration, and law	10%	11%	10%	0.082	10%	0.029
... Services	4%	3%	4%	-0.255	3%	-0.041
... Engineering, manufacturing, and construction	3%	3%	3%	-0.258	3%	-0.019
Health care utilization, one year prior						
Any PCP visits	93%	93%	92%	0.031	91%	0.061
PCP visits	8.197	8.135	8.147	0.009	7.806	0.054
	(8.0)	(7.5)	(8.0)		(7.6)	
Any prescription drug purchase	80%	82%	79%	0.055	79%	0.057
Pharmaceutical months supply	10.7	10.7	10.8	0.011	10.5	0.025
	(17.4)	(16.3)	(16.9)		(17.4)	

TABLE A2 CTD.

EXTERNAL VALIDITY, SAMPLE JUST PRIOR TO RANDOMIZATION

	Within-experiment		*Comparisons: To eldercare workers in participating municipalities*		*Comparisons: To eldercare workers in non-participating municipalities*	
	Mean, treatment group *N=3,735*	Mean, control group *N=3,806*	Other eldercare workers *N=7,945*	Std. diff. intervention group vs. other eldercare workers	Other eldercare workers *N=107,292*	Std. diff. intervention group vs. other eldercare workers
Any hospital contact, psychiatric diagnosis	1%	1%	1%	-0.022	1%	-0.019
Days hospitalized, psychiatric diagnosis	0.15	0.165	0.134	-0.006	0.174	-0.015
	(3.3)	(4.0)	(2.5)		4.5	
Any hospital contact, somatic diagnosis	37%	36%	35%	0.025	36%	0.012
Days hospitalized, somatic diagnosis	1.772	1.576	1.862	-0.022	1.787	-0.011
	(6.4)	(4.2)	(6.8)		5.915	
Employment outcomes, six months prior						
Any absenteeism, Jan-June 2008	81%	81%	74%	0.185	77%	0.135
% of scheduled hours missed, Jan-June 2008	6.85	6.97	5.81	0.107	6.40	0.052
	(11.7)	(11.7)	(10.6)		(11.5)	
Any hours worked, Jan-June 2008	99%	99%	98%	0.074	98%	0.074
Hours worked, Jan-June 2008	785	772	739	0.081	754	0.089
	(677)	(433)	(611)		(276)	
% working for same employer, Jan 2008	84%	82%	78%	0.180	78%	0.176
Earnings Jan-June 2008, DKK	119,379	122,142	113,802	0.168	118,994	0.066
	(46,722)	(50,223)	(55,477)		(53,531)	

Note: This table compares the intervention group to the eldercare workers in the participating municipalities and to eldercare workers in other municipalities. Outcomes related to health care utilization measured during the 12 months year prior to randomization; employment outcomes measured during the six months prior to randomization. The number of observations for measures of absenteeism is based on 3,542 (95%) treated, 3,581 (94%) controls, 6,898 (87%) eldercare workers in the participating municipalities, and 95,294 (89%) eldercare workers in the non-participating municipalities.

TABLE A3

EFFECTS OF "HEALTHY-AT-WORK" ON INTERACTIONS WITH PRIMARY CARE PHYSICIANS, ROBUSTNESS ANALYSES

	(1) Year 1	(2) Year 2	(3) Year 3-7	(4) Year 8-12
Panel A. Any PCP consultation (N = 7,541)				
Main specification	-0.006	0.001	0.001	0.001
	[0.006]	[0.006]	[0.002]	[0.004]
Adjusted *P* value	(0.346)	(0.855)	(0.501)	(0.772)
No covariates	-0.008	-0.003	0.001	0.003
	[0.006]	[0.007]	[0.001]	[0.001]
Adjusted *P* value	(0.172)	(0.719)	(0.786)	(0.937)
Clustering at employer level	-0.006	0.001	0.001	0.001
	[0.008]	[0.007]	[0.002]	[0.006]
Adjusted *P* value	(0.494)	(0.875)	(0.478)	(0.826)
Panel B. # PCP consultations per year (N = 7,541)				
Main specification	**-0.502**	**-0.433**	**-0.365**	-0.142
	[0.132]	[0.149]	[0.124]	[0.144]
Adjusted *P* value	(0.000)	(0.008)	(0.014)	(0.566)
No covariates	**-0.554**	**-0.514**	**-0.463**	-0.263
	[0.220]	[0.231]	[0.198]	[0.198]
Adjusted *P* value	(0.023)	(0.051)	(0.042)	(0.365)
Clustering at employer level	**-0.502**	**-0.433**	**-0.365**	-0.142
	[0.186]	[0.179]	[0.124]	[0.156]
Adjusted *P* value	(0.074)	(0.049)	(0.036)	(0.623)

TABLE A4

P-VALUES ASSOCIATED WITH ESTIMATES

AFTER CORRECTION FOR MULTIPLE HYPOTHESIS TESTING

	(1) No correction	(2) Correct for own group of outcomes	(3) Correct for all health related outcomes
Year 1			
Any PCP consultation	0.298	0.346	0.537
# PCP consultations	<0.000	<0.000	<0.000
Any prescription drug purchase	0.140	0.289	0.391
Months' supply per year	0.427	0.690	0.537
Year 2			
Any PCP consultation	0.850	0.855	0.855
# PCP consultations	0.004	0.008	0.014
Any prescription drug purchase	0.385	0.453	0.611
Months' supply per year	0.242	0.346	0.583
Years 3-7			
Any PCP consultation	0.492	0.501	0.733
# PCP consultations	0.004	0.014	0.022
Any prescription drug purchase	0.633	0.647	0.733
Months' supply per year	0.034	0.041	0.107
*Years 8-12**			
Any PCP consultation	0.745	0.772	0.772
# PCP consultations	0.326	0.566	0.669
Any prescription drug purchase	0.309	0.321	0.669
Months' supply per year	0.106	0.197	0.385

Notes: This table shows p-values associated with estimates from Tables 3 and 4 with varying corrections for multiple hypothesis testing. Column 1 does not account for multiple hypothesis testing; Column 2 considers primary care physician consultations and prescription drug purchases as separate families; and Column 3 considers all health-related outcomes, including hospitalizations, as one family.

* Prescription drug data are only available up to 11 years after randomization and hospitalization data that enter the calculations in column 3 are only available up to 10 years after randomization.

Appendix B. Heterogeneity considerations

This section follows the approach from Chernozhukov et al. (2020) for RCTs and use their associated R-code. The analysis is based on the same conditioning set, or moderators, as in regression analysis above. We also cluster at the work-unit level and define our stratification groups by the employer (or municipality) identities. In line with Buhl-Wiggers et al. (2022) we use 50 repeated random splits of the raw data into main and auxiliary samples.[9] The machine learning procedure uses the auxiliary sample to deliver estimates of a control group conditional mean function $B(X) = E[Y_0|X]$ and a proxy predictor $S(X) = E[Y_1 - Y_0|X]$ of the conditional average treatment effect (CATE), $s_0(X)$. Like Buhl-Wiggers et al. (2022), we present the Best Linear Predictor (BLP) of the CATE and the Sorted Group Average Treatment Effects (GATES).

Table B1 shows the coefficients from the BLP of the CATE using the main sample. These stem from an estimation of the following linear regression using predictions $\tilde{B}(X)$ and $\tilde{S}(X)$ from the first-step machine learning procedure:

$$(2)\ Y = \alpha_0 + \alpha_1\tilde{B}(X) + \beta_1(D - E(D)) + \beta_2(D - E(D))(\tilde{S}(X) - E(\tilde{S}(X))) + u$$

Based on performance tests for machine learning methods, we show results from the elastic net method. The first column reports $\hat{\beta}_1$ that corresponds to the estimate of the average treatment effect and the second column informs about $\hat{\beta}_2$, the heterogeneity loading. Note that if the heterogeneity loading is zero, it is either because effects of *Healthy-at-work* do not vary across individuals, or because effects do not vary with our available conditioning set. Reassuringly, we find that the estimates of $\hat{\beta}_1$ all both resemble our main estimates of the treatment effects from Table 2 above. The estimates of β_2 are small and statistically insignificant regardless of the period in question.

[9] Results are almost identical when using 200 splits instead.

TABLE B1

BEST LINEAR PREDICTOR OF THE CONDITIONAL TREATMENT EFFECTS OF HEALTHY-AT-WORK USING CAUSAL PROXIES, NUMBER OF CONSULTATIONS TO PRIMARY CARE PHYSICIAN

ATE (β_1)	HET (β_2)
Panel A. Year 1	
-0.42	0.20
(-0.86,0.02)	(-0.25,0.64)
[0.13]	[0.70]
Panel B. Year 2	
-0.35	0.37
(-0.80,0.11)	(-0.11,0.86)
[0.28]	[0.28]
Panel C. Year 3-7	
-0.29	0.22
(-0.66,0.08)	(-0.30,0.72)
[0.25]	[0.82]

Notes: Best linear predictors of the conditional average treatment effects are estimated using the approach from Chernozhukov et al. (2020). This table presents median values over 50 random splits of the sample; 95% confidence intervals in parentheses. We present the elastic net results based on performance tests for machine learning methods.

We continue to produce estimates of the GATES by quintiles of the proxy predictor, $S(X)$. Table B2 shows the differences in the GATES between the most and least affected quintiles. Here, the elastic net method calculates a difference of almost one visit between effects for the two groups in the first year after randomization. This is in the ballpark of the difference between the estimated effects in our high-risk and low-risk groups in the conventional analysis from above. We detect a difference across the two quintiles of 1.5 visits in the second year after randomization and of 0.6 visits 3-7 years after. Since the differences are not statistically significant, we conservatively conclude that there is some indication of heterogeneity across the quintiles.

TABLE B2

DIFFERENCES IN SORTED GROUP AVERAGE TREATMENT EFFECTS, MOST AND LEAST AFFECTED QUINTILES

20% least affected	20% most affected	Difference
Panel A. Year 1		
0.00	-0.81	0.85
(-0.71,0.73)	(-2.07,0.38)	(-0.62,2.27)
[1.00]	[0.31]	[0.51]
Panel B. Year 2		
0.20	-1.40	1.51
(-0.63,1.02)	(-2.75,-0.07)	(-0.101,3.179)
[1.00]	[0.08]	[0.14]
Panel C. Year 3-7		
-0.04	-0.59	0.60
(-0.71,0.63)	(-1.57,0.38)	(-0.61,1.79)
[1.00]	[0.47]	[0.67]

Notes: This table shows differences in sorted group average treatment effects (GATES) across the most and least affected quintiles. GATES are estimated using the approach of Chernozhukov et al. (2020). Note that we have switched the wording of most and least compared to Chernozhukov et al. (2020) because *Healthy-at-work* negatively affects this outcome. Point estimates by ML proxy quintile and joint uniform 95 percent confidence intervals are estimated based on 50 random splits of the sample.

We finally perform a classification analysis on the covariates; full tables are available upon request. Here, we compare the average characteristics of the most and least affected quintiles using two-sample t-tests. The model indicates that those with higher gains are more likely to have many contacts to their primary care physician (our indicator for belonging to the high-risk group), more likely to have any prescription drug purchase, to purchase more prescription drugs, and to be more likely to have a hospital visit associated with a psychiatric diagnosis. Thus, our conclusions from above regarding the high- and low-risk groups are robust to using alternative proxies for being in poor health. The evidence on heterogeneity by socio-economic conditions is less clear and the conclusions sometimes vary across post-randomization periods.

Literature

Buhl-Wiggers, J., J. Kerwin, J. S. Munoz, J. Smith, & R. Thornton (2022): “Some Children Left Behind: Variation in the Effects of an Educational Intervention.” Journal of Econometrics. In Press.

Chernozhukov, V., M. Demirer, E. Duflo, & I. Fernández-Val (2020): “Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India.” Working paper 24678, National Bureau of Economic Research.