Workplace wellness programs have become increasingly popular as employers have aimed to lower health care costs and improve employee health and productivity. In 2018, 82% of large firms and 53% of small employers in the United States offered a wellness program, amounting to an $8 billion industry.[1,2] This growth has been aided by public investments such as the Affordable Care Act, which included funds to promote the development of workplace wellness programs.

Workplace wellness programs tend to focus on modifiable risk factors of disease, such as nutrition, physical activity, and smoking cessation. Despite widespread adoption, causal evidence of such programs' effects on health and economic outcomes has been limited. Meta-analyses have produced varying estimates of benefits relative to costs.[3-5] Observational studies have often been limited by a lack of valid control groups, selection bias, and small samples.[6-8] Experimental studies of comprehensive wellness programs have been scarce and have produced mixed results, with most of the more rigorous studies now dated.[9,10] Other experimental studies have focused on certain components of wellness, such as smoking cessation and weight loss, using an intervention of limited duration.[11-14] A recent rigorous randomized study used individual-level rather than workplace-wide randomization, making it difficult to assess the effects of the tools used by many programs aiming to improve workplace culture or harness peer effects.[15]

Using a design that randomized the implementation of wellness programming at the worksite level, this study evaluated the effect of a multiyear workplace wellness program on health and economic outcomes over 18 months in a middle- and lower-income employee population at locations across the eastern United States.

## Key Points

**Question** What is the effect of a multicomponent workplace wellness program on health and economic outcomes?

**Findings** In this cluster randomized trial involving 32 974 employees at a large US warehouse retail company, worksites with the wellness program had an 8.3-percentage point higher rate of employees who reported engaging in regular exercise and a 13.6-percentage point higher rate of employees who reported actively managing their weight, but there were no significant differences in other self-reported health and behaviors; clinical markers of health; health care spending or utilization; or absenteeism, tenure, or job performance after 18 months.

**Meaning** Employees exposed to a workplace wellness program reported significantly greater rates of some positive health behaviors compared with those who were not exposed, but there were no significant effects on clinical measures of health, health care spending and utilization, or employment outcomes after 18 months.

## Methods

### Setting and Intervention

The research protocol was reviewed and approved by the institutional review boards at the Harvard T.H. Chan School of Public Health and Harvard Medical School. Written informed consent was obtained from all participants prior to primary data collection. All statistical analyses were prespecified in advance of making any treatment-control outcome comparisons and were publicly archived at clinicaltrials.gov and the American Economic Association Randomized Clinical Trials Registry. The protocol and analysis plan are available in Supplement 1.

A comprehensive workplace wellness program was implemented at a large warehouse retail company, BJ's Wholesale Club, which employs approximately 26 000 workers across 201 worksites along the eastern United States (eFigure 1 in Supplement 2). The wellness program comprised 8 modules implemented over 18 months, from January 2015 through June 2016 (eTable 1 in Supplement 2). Each module lasted 4 to 8 weeks and focused on key elements of health and wellness, including nutrition, physical activity, stress reduction, and prevention. Programming content was delivered by registered dietitians assigned to the treatment worksites, using both individual and team-based activities and challenges. Modules included modest incentives for participation, most commonly a $25 BJ's gift card for completing a particular module. Total potential incentives across the program averaged about $250 (details about the modules and incentives are provided in eMethods 1 in Supplement 2). The intervention was designed and implemented by an established wellness vendor, Wellness Workdays.

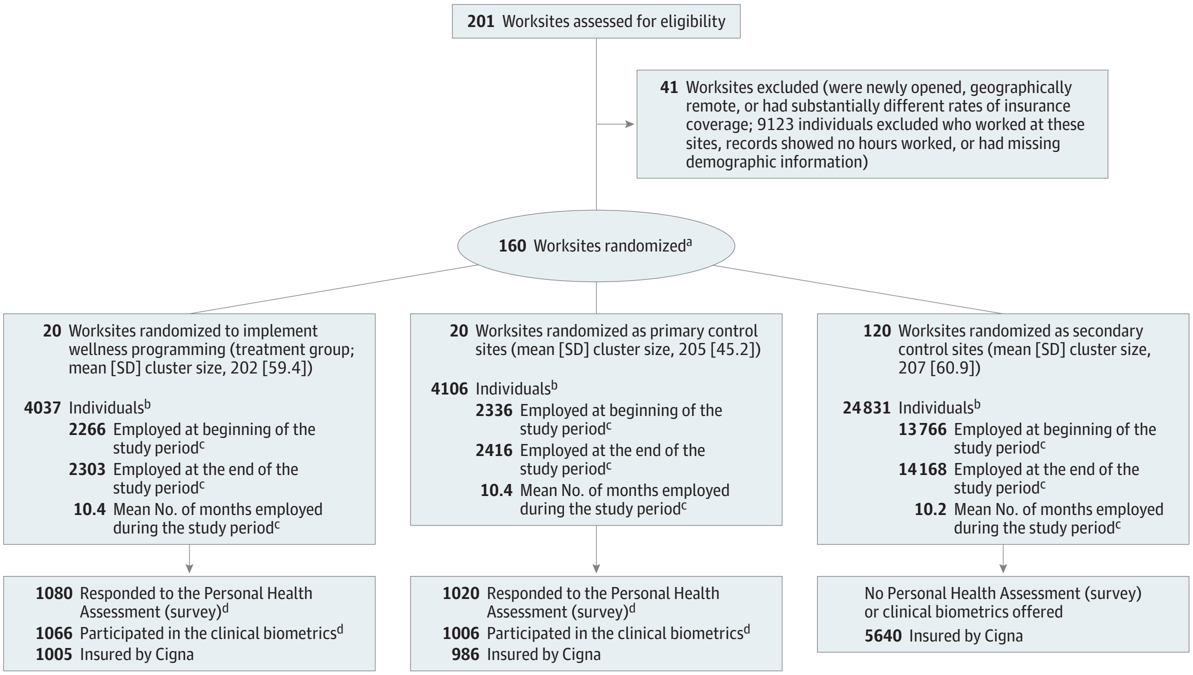
### Randomization

The wellness program was implemented in a randomly selected subset of worksites through simple randomization using a computer-generated random number (**Figure**). Worksites, rather than individuals, were randomized because wellness programs often use team-based interventions and aim to change workplace culture and environment.

Forty-one worksites were excluded because they were geographically remote or had substantially different insurance coverage, leaving 160 sites in the sample (mean of 108 employees per site). We randomly selected 20 treatment sites in which the program was available to all employees, with the remaining sites serving as controls in which there was no wellness program. Survey and clinical data were collected at the 20 treatment sites and at 20 randomly selected primary control sites. The remaining 120 sites served as secondary controls. Administrative data were collected from all 160 worksites (eFigure 2 in Supplement 2).

Individuals were assigned to treatment or control status based on their worksite at the time of randomization or initial employment, as subsequent movement between worksites could, in principle, be influenced by the wellness program. Individuals in treatment worksites were eligible, but not required, to participate in the program and could exit the program at any time.[16] All individuals employed in treatment and primary control sites at the 18-month mark were free, but not required, to complete the survey and clinical screening.

Figure. Flow of Participants in the Trial of a Workplace Wellness Program on Employee Health and Economic Outcomes


201 Worksites assessed for eligibility
41 Worksites excluded (were newly opened, geographically remote, or had substantially different rates of insurance coverage; 9123 individuals excluded who worked at these sites, records showed no hours worked, or had missing demographic information)
160 Worksites randomized[a]
20 Worksites randomized to implement wellness programming (treatment group; mean [SD] cluster size, 202 [59.4])
4037 Individuals[b]
2266 Employed at beginning of the study period[c]
2303 Employed at the end of the study period[c]
10.4 Mean No. of months employed during the study period[c]
20 Worksites randomized as primary control sites (mean [SD] cluster size, 205 [45.2])
4106 Individuals[b]
2336 Employed at beginning of the study period[c]
2416 Employed at the end of the study period[c]
10.4 Mean No. of months employed during the study period[c]
120 Worksites randomized as secondary control sites (mean [SD] cluster size, 207 [60.9])
24831 Individuals[b]
13766 Employed at beginning of the study period[c]
14168 Employed at the end of the study period[c]
10.2 Mean No. of months employed during the study period[c]
1080 Responded to the Personal Health Assessment (survey)[d]
1066 Participated in the clinical biometrics[d]
1005 Insured by Cigna
1020 Responded to the Personal Health Assessment (survey)[d]
1006 Participated in the clinical biometrics[d]
986 Insured by Cigna
No Personal Health Assessment (survey) or clinical biometrics offered
5640 Insured by Cigna

[a] Randomization took place at the level of the worksite.

[b] All individuals were assigned to treatment or control status based on the first worksite in which they appeared during the treatment period. The number of employees in each arm of the trial represents the number of unique individuals employed in the company's workforce in one of the 160 worksites during the study period.

[c] There was natural employment turnover at the company during the study period, so the number of employees at the beginning of the study period, the number at the end of the study period, and the mean number of months each worker was employed during the study period are also shown.

[d] Only workers employed at the end of the study period were eligible to complete the survey and clinical biometrics (they could choose to participate in either, both, or neither).

## Outcomes

Prespecified outcomes were collected across 4 domains, of which 2 were gathered in person and 2 derived from administrative data (eTables 2-3 in Supplement 2). Two in-person primary data domains were collected in the 20 treatment and 20 primary control sites at the end of the study period. Self-reported health and behaviors data were collected via personal health assessment surveys and included measures such as exercise, diet, smoking, and alcohol use.[17] Clinical measures of health data were obtained from clinical biometric screenings by registered nurses and included blood pressure, body mass index, blood glucose levels, and cholesterol levels. No imputation was done for any unanswered survey items or unmeasured biometrics. We assessed potential selection into in-person data collection by comparing baseline characteristics of employees who participated in surveys and biometrics to those who did not.

Administrative data, gathered for all 160 treatment and control worksites, comprised employment records and health insurance claims collected continuously over the study period. Information on health care spending and utilization was gathered for the subset of workers enrolled in employer-sponsored insurance plans through Cigna, the third-party administrator for this self-insured firm. About half of stably employed workers (defined below) and a third of workers at any time were enrolled in Cigna; there were no missing data for these workers. Employment outcomes were gathered from employment records and included absenteeism and tenure (data available for all employees), as well as available work performance evaluations for the 73% of employees during the study period who had an evaluation.

## Statistical Analyses

After randomization of the worksites (with the number of treatment sites limited by the study budget), we conducted initial power calculations before implementing this randomized clinical trial or collecting outcome data. Power calculations were made using data from secondary data sources, including the National Health and Nutrition Examination Survey, the Behavioral Risk Factor Surveillance System, the Medical Expenditure Panel Survey, and commercial insurance claims, to generate benchmark means and standard deviations, using standard assumptions about intracluster correlation and power. Details on these power calculations and estimated detectable differences are provided in eMethods 2 in Supplement 2.

In our primary analyses, we estimated the effect of working at a treatment worksite on outcomes, regardless of participation in the wellness program.[18] For administrative outcomes, we compared all employees at treatment sites with all employees at control sites (an intention-to-treat design); for in-person primary data outcomes, we analyzed employees who were available at the 18-month mark (analogous to intention-to-treat, assessed in the available population).

We used an individual-level linear model with an indicator for employment in a treatment site as the key independent variable. This captured the effect of offering the opportunity to participate in the wellness program. As with most retailers, there was natural employee turnover during the study period. We included all individuals employed at any point during the study period. While administrative outcomes were available for all individuals, in-person primary data were gathered from participants in screenings and biometric collection at the 18-month mark. The study sample also included those who worked full time and those who worked only part time. The model weighted each individual by exposure to the program, as measured by the work schedule and share of the treatment period the individual was employed, described in eMethods 2 in Supplement 2.

To improve the precision of estimates and balance between treatment and control groups, we controlled for age, sex, age-sex interactions, race, and initial employment characteristics (not plausibly affected by the program) and also included minimum variance weights constructed to make the distribution of age, sex, and race representative of the entire study population—a method that performs better than a model-based approach that fits a propensity score.[19-21] Data on race/ethnicity, used to describe the population and compare demographics between study groups, were gathered from voluntary survey responses of study participants to multiple-choice options presented by the investigators.

Because multiple measures within an outcome domain may reflect the same fundamental outcome, we prespecified standardized treatment effects across categories of clinical measures of health, self-reported health behaviors, and mental health and well-being. The standardized treatment effect is a summary measure of closely related outcomes and denotes the mean change across all of the components in the domain, measured in units of standard deviations (that is, the size of the estimated effect for an outcome relative to standard deviation of that outcome, averaged across all of the outcomes in the domain).

Because of the potential for type I error due to multiple comparisons, as prespecified, we also adjusted for multiple inference within outcome domains and reported both standard, per-comparison *P* values and adjusted, "family-wise" *P* values that accounted for multiple inference, using a conservative approach of grouping together a wide range of outcomes following the Westfall and Young method with 1000 bootstrap replications.[22] Standard errors were clustered by worksite.

In addition to the effect of working in a treatment site, the effect of actually participating in the program may be of interest. Since participation was voluntary (and thus potentially related to health or health behaviors), a simple comparison of participants to non-participants risks producing biased estimates. We used a standard 2-stage least squares instrumental variables approach to estimate the local average treatment effect of program participation, with randomization into treatment as the instrument for participation. Our primary definition of participation was the completion of at least 1 program module, but we also tested robustness to other definitions (completion of at least 3 modules and number of modules completed).

We assessed the heterogeneity of program effects by age and sex among prespecified and key outcomes by testing for differences in the coefficient of interest using an interaction term between treatment status and the demographic characteristic of interest. We also conducted a number of secondary analyses. First, we estimated program effect on a prespecified cohort of stably employed workers who were employed for at least 13 consecutive weeks prior to the intervention. Second, we evaluated aggregate employment and claims outcomes at the worksite level. Third, we analyzed key outcomes using only the exposure weights. Fourth, we estimated logistic models for binary outcomes.

To assess endogenous selection into program participation, we compared the baseline characteristics of program participants to those of non-participants in treatment sites. To assess endogenous selection into participation in primary data collection, we compared baseline characteristics of workers who elected to provide clinical data or complete the health risk assessment to those of workers who did not, separately within the treatment group and the control group. This enabled us to assess any potential differential selection into primary data collection. Additionally, to examine differences in findings between our randomized trial approach and a standard observational design (and thereby any bias that confounding factors would have introduced into naive observational estimates), we generated estimates of program effects using ordinary least squares to compare program participants with nonparticipants (rather than using the variation generated by randomization).

Two-tailed tests were used, with a significance level of *P* = .05. Detailed methods are available in eMethods 3 in Supplement 2.

## Results

### Population and Participation

The study population included 4037 individuals at the 20 treatment worksites, 4106 at the 20 primary control worksites, and 24 831 at the 120 secondary control worksites. Their demographic and employment characteristics, with balance weights, are shown in **Table 1**.

About 20% of the population was black and 18% Hispanic. Full-time workers comprised about 60% of the study population. Mean earnings for full-time salaried workers was slightly under $50 000 per year, and full-time hourly workers earned about half that amount. Population characteristics without balance weights are shown in eTable 4 in Supplement 2.

Program participation increased from 12.2% in the first module to, on average, 30.6% in the subsequent modules

Table 1. Characteristics of the Study Population[a]

| | All Employees, No. (%) | | |
|---|---|---|---|
| | Treatment Worksites (n = 4037) | Primary Control Worksites (n = 4106) | Primary and Secondary Control Worksites (n = 28 937) |
| **Demographic Characteristics** | | | |
| Age, mean (SD), y | 38.8 (0.7) | 38.3 (0.5) | 38.7 (0.2) |
| Sex, % | | | |
| Male | 2104 (53.7) | 2151 (54.5) | 15 597 (54.2) |
| Female | 1933 (46.3) | 1955 (45.5) | 13 339 (45.8) |
| Race/ethnicity, % | | | |
| Black | 797 (19.8) | 1004 (20.1) | 7218 (20.7) |
| White | 2601 (56.3) | 2203 (57.9) | 14 754 (55.3) |
| Hispanic | 402 (17.9) | 720 (17.1) | 5161 (17.8) |
| Other | 237 (6.0) | 179 (5.0) | 1803 (6.2) |
| | Stably Employed Subsample (n = 1892) | Stably Employed Subsample (n = 1930) | Stably Employed Subsample (n = 13 452) |
| **Employment Characteristics[b]** | | | |
| Worker type, % | | | |
| Full-time salaried | 232 (15.5) | 222 (15.2) | 1605 (16.4) |
| Full-time hourly | 700 (44.9) | 743 (47.0) | 5113 (46.2) |
| Part-time hourly | 960 (39.6) | 965 (37.8) | 6734 (37.4) |
| Annual earnings, mean (SD), $ | | | |
| Full-time salaried | 49 340 (1116.8) | 47 669 (698.4) | 48 467 (298.7) |
| Full-time hourly | 25 727 (682.6) | 24 528 (436.1) | 25 296 (173.8) |
| Part-time hourly | 10 301 (180.5) | 9981 (100.7) | 10 034 (48.2) |
| Job category, % | | | |
| Sales workers | 720 (34.3) | 741 (32.5) | 5085 (31.9) |
| Laborers/helpers | 345 (20.1) | 351 (20.6) | 2495 (20.6) |
| Operative workers[c] | 309 (16.1) | 291 (15.4) | 2063 (15.9) |
| Service workers | 225 (11.6) | 254 (13.1) | 1688 (12.1) |
| Mid-level officials | 184 (11.5) | 172 (11.4) | 1262 (12.4) |
| Administrative support | 70 (4.3) | 71 (4.4) | 570 (5.1) |
| Other | 39 (2.0) | 50 (2.6) | 289 (2.0) |
| Employer-sponsored insurance[d] | | | |
| Ever enrolled in 2014, % | 762 (50.0) | 748 (50.1) | 5052 (48.3) |
| Months enrolled in 2014, mean | 11.5 | 11.5 | 11.5 |

[a] All individuals were assigned to treatment or control status based on the first worksite in which they appeared during the treatment period. Characteristics were weighted by exposure to the wellness program based on duration of employment and hours worked and a weight that balances treatment and control on demographics. Age was defined as of December 2014, the year before the intervention.

[b] Employment characteristics were derived from the stably employed subsample and measured at the time of an individual's first appearance in the data.

[c] The operative workers category included cake decorator, stock/cart retriever, gas station team member, bakery clerk, bakery supervisor, meat clerk or cutter, deli supervisor or clerk, produce clerk, merchandise specialist, order picker packer, order forklift driver, order fulfillment specialist, produce inspector, pallet forklift operator, and lead storage forklift operator.

[d] BJ's Wholesale Club, a self-insured employer, offered employer-sponsored health insurance through a third-party administrator, Cigna. Enrollment data were taken from 2014, the year before the intervention. About half of stably employed employees and a third of all employees were enrolled in Cigna.

(eTable 5 in Supplement 2). Overall, 35.2% of individuals ever employed in treatment sites completed at least 1 module and 21.4% completed at least 3 modules (mean of 1.3 modules). Among those who completed at least 1 module, 60.9% completed at least 3 modules, with a mean of 3.7 modules completed. Participation in the personal health assessment survey and biometric screening at the 18-month mark (June 2016) was 25.8% and 25.5%, respectively, among individuals ever employed in the 20 treatment or 20 primary control worksites during the study period. Among individuals employed in June 2016, mean participation in surveys and screenings was 42.4% and 42.8%, respectively, across these 40 worksites.

Tables show effects of working at a treatment worksite and of participating in the wellness program on main outcomes (**Table 2**, **Table 3**, **Table 4**, **Table 5**). Full results across domains and for alternative populations are shown in eTables 6-10 in Supplement 2.

## Self-reported Health and Behaviors

Effects on self-reported health and behaviors are shown in Table 2. The number of individuals providing these outcomes ranged between 1722 and 2022 (35.3% to 41.4% of individuals employed in June 2016). Randomization into a treatment worksite led to a higher proportion who reported engaging in regular exercise by 8.3 percentage points (95% CI, 3.9-12.8; unadjusted $P < .001$, adjusted $P = .03$) (treatment group, 69.8% vs control group, 61.9%), and who reported actively managing their weight by 13.6 percentage points (95% CI, 7.0 to 20.2; unadjusted $P < .001$, adjusted $P = .02$) (treatment group, 69.2% vs control group, 54.7%).

For some outcomes, such as smoking and alcohol use, randomization into treatment had a statistically significant effect by traditional *P* values, but statistical significance was not robust to multiple inference adjustment. For rates of smoking, the unadjusted treatment group mean was

Table 2. Mean Values and Effect of Program on Self-reported Health and Behaviors[a]

| | Group Mean (SD) | | Effect of Availability of Wellness Program (Assessed in the Population Available) | | | Effect of Participation in Wellness Program (Local Mean Treatment Effect) | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Treatment[b] | Control[c] | Effect (95% CI)[d] | *P* Value | Adjusted *P* Value | Effect (95% CI)[d] | *P* Value | Adjusted *P* Value |
| Screening and examinations | | | | | | | | |
| Annual examination, % | 65.6 (47.5) | 65.5 (47.6) | −1.3 (−7.0 to 4.5) | .66 | >.99 | −1.6 (−8.7 to 5.5) | .66 | >.99 |
| Flu shot, % | 33.5 (47.2) | 35.2 (47.8) | −2.4 (−8.3 to 3.5) | .42 | >.99 | −3.1 (−10.4 to 4.2) | .41 | >.99 |
| % of other recommended tests received | 59.9 (31.4) | 55.9 (31.0) | 3.2 (0.0 to 6.4) | .05 | .69 | 4.1 (0.1 to 8.1) | .05 | .71 |
| Mental health and well-being | | | | | | | | |
| PHQ-2 score of ≥3, %[e] | 7.6 (26.6) | 8.5 (28.0) | −0.1 (−3.5 to 1.5) | .44 | >.99 | −1.2 (−4.3 to 1.8) | .43 | >.99 |
| SF-8 score[f] | | | | | | | | |
| Physical summary score | 50.5 (8.0) | 50.8 (7.7) | −0.2 (−0.8 to 0.5) | .66 | >.99 | −0.2 (−1.0 to 0.7) | .66 | >.99 |
| Mental summary score | 50.9 (9.1) | 51.2 (9.1) | −0.4 (−1.2 to 0.5) | .44 | >.99 | −0.4 (−1.6 to 0.8) | .43 | >.99 |
| Unmanaged stress, % | 39.1 (48.8) | 41.8 (49.3) | −2.7 (−7.7 to 2.3) | .28 | .99 | −3.5 (−9.7 to 2.7) | .27 | .99 |
| Stress at work, % | 56.2 (49.6) | 55.7 (49.7) | 2.0 (−2.6 to 6.6) | .40 | >.99 | 2.5 (−3.2 to 8.3) | .39 | >.99 |
| Good-quality, adequate amount of sleep, % | 52.2 (50.0) | 54.1 (49.9) | −2.1 (−6.0 to 1.8) | .29 | .99 | −2.7 (−7.6 to 2.2) | .29 | .99 |
| Regular exercise, % | 69.8 (46.0) | 61.9 (48.6) | 8.3 (3.9 to 12.8) | <.001 | .03 | 10.6 (5.3 to 16.0) | <.001 | .03 |
| ≥3 d/wk of moderate exercise, % | 68.0 (46.7) | 64.0 (48.0) | 4.1 (−0.6 to 8.8) | .09 | .85 | 5.3 (−0.6 to 11.) | .08 | .84 |
| No. of d/wk intentionally increase activity | 3.2 (2.3) | 3.0 (2.4) | 0.1 (−0.1 to 0.3) | .44 | >.99 | 0.1 (−0.2 to 0.4) | .44 | >.99 |
| No. of hours sitting per day | 3.5 (1.9) | 3.5 (1.7) | 0.0 (−0.2 to 0.2) | .83 | >.99 | 0.0 (−0.2 to 0.3) | .83 | >.99 |
| Nutrition | | | | | | | | |
| No. of meals eaten out | 1.8 (1.4) | 1.8 (1.6) | −0.1 (−0.2 to 0.1) | .48 | >.99 | −0.1 (−0.3 to 0.1) | .47 | >.99 |
| No. of naturally or artificially sweetened drinks per day | 1.9 (1.9) | 1.8 (1.9) | 0.1 (−0.1 to 0.3) | .34 | >.99 | 0.1 (−0.1 to 0.4) | .33 | >.99 |
| Read the Nutrition Facts panel, % | 63.3 (48.2) | 58.7 (49.3) | 4.4 (−1.0 to 9.8) | .11 | .91 | 5.6 (−1.1 to 12.3) | .10 | .91 |
| Consume at least 2 cups of fruit and 2.5 cups of vegetables per day, % | 62.5 (48.4) | 57.5 (49.5) | 3.3 (−1.1 to 7.7) | .14 | .93 | 4.2 (−1.2 to 9.6) | .13 | .92 |
| Choose whole grain and reduced-fat foods more often than the regular variety, % | 35.7 (47.9) | 33.2 (47.1) | 1.2 (−3.2 to 5.6) | .58 | >.99 | 1.6 (−3.9 to 7.0) | .58 | >.99 |
| Weight management | | | | | | | | |
| Considering losing weight in the next 6 mo, % | 66.8 (47.1) | 56.3 (49.6) | 9.5 (3.7 to 15.4) | .002 | .09 | 12.1 (4.8 to 19.4) | .001 | .11 |
| Actively managing weight, % | 69.2 (46.2) | 54.7 (49.8) | 13.6 (7.0 to 20.2) | <.001 | .02 | 17.2 (9.1 to 25.4) | <.001 | .01 |
| Smoker, % | 17.3 (37.9) | 24.6 (43.1) | −6.9 (−12.9 to −0.9) | .03 | .52 | −8.8 (−16.3 to −1.3) | .02 | .53 |
| No. of alcoholic drinks per week | 4.0 (6.3) | 4.6 (7.4) | −0.6 (−1.1 to 0.0) | .04 | .69 | −0.7 (−1.4 to −0.0) | .04 | .68 |
| Medical utilization | | | | | | | | |
| No. of physician visits in last 12 mo | 1.6 (1.1) | 1.5 (1.1) | 0.0 (−0.1 to 0.1) | .98 | >.99 | 0.0 (−0.1 to 0.2) | .98 | >.99 |
| Any physician visit in last 12 mo, % | 75.8 (42.9) | 75.5 (43.0) | −0.6 (−5.3 to 4.1) | .80 | >.99 | −0.7 (−6.6 to 5.1) | .80 | >.99 |
| Any emergency visit in last 12 mo, % | 22.6 (41.9) | 25.8 (43.8) | −3.5 (−8.0 to 1.0) | .13 | .92 | −4.5 (−10.1 to 1.2) | .12 | .92 |
| Ever hospital patient in the last 12 mo, % | 15.0 (35.7) | 17.5 (38.0) | −2.9 (−7.0 to 1.1) | .15 | .93 | −3.7 (−8.6 to 1.3) | .14 | .93 |
| Days spent in hospital | 0.4 (1.3) | 0.4 (1.4) | −0.1 (−0.2 to 0.1) | .28 | .99 | −0.1 (−0.3 to 0.1) | .27 | .99 |
| No. of different prescriptions in last 12 mo | 1.3 (1.6) | 1.3 (1.6) | −0.1 (−0.2 to 0.1) | .43 | >.99 | −0.1 (−0.3 to 0.1) | .43 | >.99 |
| Any prescriptions in last 12 mo, % | 52.6 (50.0) | 52.8 (49.9) | −1.8 (−6.0 to 2.5) | .41 | >.99 | −2.2 (−7.5 to 3.0) | .40 | >.99 |
| Standardized treatment effect[g] | | | | | | | | |
| Mental health and well-being | | | 0.0 (0.0 to 0.0) | .97 | | 0.0 (−0.1 to 0.1) | .97 | |
| Health behaviors | | | 0.1 (0.0 to 0.1) | .001 | | 0.1 (0.0 to 0.1) | .001 | |

[a] This table reports intent-to-treat and local average treatment effect estimates of the wellness program at the employee level. All regressions included demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, Cigna coverage status, and job characteristics at the time they initially appeared in the data including full-time status, paid hourly status, and job category) and clustered SEs at the worksite level. Regressions and means were weighted by the combination of a weight for exposure to the program and a weight that balances treatment and control samples on demographics (age, sex, and race). Sample includes data from the 20 treatment worksites and 20 primary control worksites at the end of study. Because the No. of respondents varied, sample sizes of regressions ranged from 1722 to 2022 (35.3%-41.4% of those employed at the 40 worksites in June 2016).

[b] Number of observations for each outcome (range, 864-1013 [36.2%-42.4% of those employed at worksites in June 2016]).

[c] Primary control worksite number of observations for each outcome (range, 858-1009 [34.4%-40.4% of those employed at worksites in June 2016]).

[d] For variables measured as percentages in group means, the difference attributable to the wellness program is expressed in percentage points.

[e] The Patient Health Questionnaire 2 (PHQ-2) asks about frequency of depressed mood and anhedonia over the past 2 weeks. Score range, 0-6 (≥3 suggests major depressive disorder).

[f] Medical Outcomes Study 8-Item Short-Form Health Survey (SF-8) scores (range, 0-100) were normalized (mean [SD], 50 [10]) in the general US population. Higher scores indicate better self-reported health-related quality of life.

[g] SE is shown in parentheses. Mental health and well-being was calculated using table outcomes under the mental health and well-being section; health behaviors was calculated using outcomes under screenings and examinations, sleep, exercise, nutrition, weight management, smoker, and alcohol use.

Table 3. Mean Values and Effect of Program on Clinical Measures of Health[a]

| | Group Mean (SD) | | Effect of Availability of Wellness Program (Assessed in the Population Available) | | | Effect of Participation in Wellness Program (Local Mean Treatment Effect) | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Treatment[b] | Control[c] | Effect (95% CI)[d] | *P* Value | Adjusted *P* Value | Effect (95% CI)[d] | *P* Value | Adjusted *P* Value |
| Continuous measures | | | | | | | | |
| Total cholesterol, mg/dL | 180.9 (44.4) | 177.6 (41.5) | 2.6 (−5.8 to 11.0) | .54 | .99 | 3.3 (−7.1 to 13.1) | .53 | .99 |
| HDL cholesterol, mg/dL | 52.7 (15.9) | 53.0 (16.4) | −0.3 (−2.4 to 1.7) | .75 | >.99 | −0.4 (−3.0 to 2.2) | .75 | >.99 |
| Glucose, mg/dL | 104.6 (39.8) | 101.9 (33.5) | 1.4 (−4.0 to 6.8) | .61 | >.99 | 1.8 (−5.0 to 8.6) | .61 | >.99 |
| Blood pressure, mm Hg | | | | | | | | |
| Systolic | 124.9 (17.0) | 124.3 (16.9) | 0.2 (−1.7 to 2.2) | .81 | >.99 | 0.3 (−2.1 to 2.7) | .80 | >.99 |
| Diastolic | 80.3 (11.0) | 79.7 (10.6) | 0.5 (−0.8 to 1.8) | .46 | .98 | 0.6 (−0.1 to 2.2) | .45 | .98 |
| BMI | 29.9 (7.1) | 29.7 (7.1) | 0.1 (−0.6 to 0.8) | .79 | >.99 | 0.1 (−0.7 to 0.1) | .78 | >.99 |
| Binary measures, % | | | | | | | | |
| High total cholesterol (≥200 mg/dL) | 30.3 (46.0) | 29.3 (45.6) | 0.1 (−8.0 to 8.1) | .99 | >.99 | 0.1 (−10.0 to 10.1) | .99 | >.99 |
| Low HDL cholesterol (<40 mg/dL) | 20.3 (40.2) | 22.3 (41.7) | −1.1 (−5.8 to 3.6) | .65 | >.99 | −1.4 (−7.4 to 4.5) | .64 | >.99 |
| Hypertension (systolic BP ≥140 or diastolic BP ≥90 mm Hg) | 26.5 (44.2) | 23.1 (42.2) | 2.7 (−2.4 to 7.8) | .30 | .93 | 3.4 (−2.9 to 9.8) | .29 | .92 |
| Obesity (BMI ≥30) | 43.5 (49.6) | 43.0 (49.5) | 0.6 (−3.7 to 4.8) | .80 | >.99 | 0.7 (−4.6 to 6.0) | .79 | >.99 |
| Standardized treatment effect for clinical outcomes[e] | | | 0.0 (−0.1 to 0.0) | .37 | | 0.0 (0.1 to 0.0) | .36 | |

Abbreviations: BMI, body mass index; HDL, high-density lipoprotein.

SI conversion factors: To convert cholesterol values to mmol/L, multiply by 0.0259; to convert glucose to mmol/L, multiply by 0.0555.

[a] This table reports intent-to-treat and local average treatment effect estimates of the wellness program at the employee level. All regressions included demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, Cigna coverage status, and job characteristics at the time they initially appeared in the data including full-time status, paid hourly status, and job category) and clustered SEs at the worksite level. Regressions and means were weighted by the combination of a weight for exposure to the program and a weight that balances treatment and control samples on demographics (age, sex, and race). This sample includes biometric data collected at the 20 treatment worksites and 20 primary control worksites at the end of the study. Due to clinical biometrics varying in number of participants, sample sizes of the regressions ranged from 2082 to 2139 (42.6%-43.8% of individuals employed at these 40 primary worksites in June 2016). These numbers exceeded the participants in clinical biometrics in the Figure because some individuals from secondary control worksites unexpectedly took part in the measurement of biometrics.

[b] Number of observations for each outcome (range, 1036-1065 [43.4%-44.6% of those employed at worksites in June 2016]).

[c] Primary control worksite number of observations for each outcome (range, 1046-1074 [41.9%-43.0% of those employed at worksites in June 2016]).

[d] For variables measured as percentages in group means, the difference attributable to the wellness program is expressed in percentage points.

[e] Calculated using only the continuous outcome variables.

17.3% and control group was 24.6% (adjusted difference, −6.9 percentage points [95% CI, −12.9 to −0.9 percentage points; unadjusted *P* = .03, adjusted *P* = .52]). For number of alcoholic drinks per week, the unadjusted treatment group mean was 4.0 and control group was 4.6 drinks (adjusted difference, −0.6 drinks [95% CI, −1.1 to 0.0 drinks; unadjusted *P* = .04, adjusted *P* = .69]). Other outcomes in this domain were not significantly affected by randomization into treatment (all *P* values >.05) (Table 2).

In the standardized treatment effect, health behaviors were 0.07 SD better (95% CI, 0.02 to 0.10; *P* = .001) in the treatment sites. There was no detectable effect on the standardized treatment effect for mental health and well-being (0.001 SD [95% CI, −0.05 to 0.05; unadjusted *P* = .97]). (As a single index, standardized treatment effects do not have adjusted *P* values.)

## Clinical Measures of Health

Results for clinical measures of health are shown in Table 3. The number of individuals providing these outcomes ranged between 2082 and 2139 (42.6% to 43.8% of individuals employed in June 2016). High cholesterol levels (30.3% in the treatment group vs 29.3% in the control group), hypertension (26.5% in the treatment group vs 23.1% in the control group), and obesity 43.5% in the treatment group vs 43.0% in the control group) did not differ between groups. Randomization into a treatment worksite did not have a detectable effect on any clinical measures of health (all *P* values >.05) or their standardized treatment effect (−0.03 SD [95% CI, −0.09 to 0.03; unadjusted *P* = .37]).

## Health Care Spending and Utilization

Results for health care spending and utilization are shown in Table 4. The sample size was 7631 or 23.2% of all employees during the study period, with no missing data among these individuals with employer-sponsored insurance. Medical spending averaged $3583 per employee per year in the treatment group vs $3953 in the control group. Pharmaceutical spending was a mean of $1412 per employee per year in the treatment group vs $1215 in the control group. Medical cost-sharing averaged $780 per year in the treatment group and $778 in the control group. Pharmaceutical cost-sharing was

Table 4. Mean Values and Effect of Program on Health Care Spending and Utilization[a]

| | Group Mean (SD) | | Effect of Availability of Wellness Program (Intention to Treat) | | | Effect of Participation in Wellness Program (Local Mean Treatment Effect) | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Treatment[b] | Control[c] | Effect (95% CI)[d] | *P* Value | Adjusted *P* Value | Effect (95% CI)[d] | *P* Value | Adjusted *P* Value |
| Medical spending, $[e] | | | | | | | | |
| Total | 3583 (11 318) | 3953 (14 697) | −425.57 (−1266 to 415) | .32 | .95 | −670.13 (−1954 to 614) | .31 | .95 |
| Spending by site of care, $ | | | | | | | | |
| Office | 1934 (6079) | 2133 (7362) | −222.01 (−723 to 279) | .38 | .97 | −349.59 (−1119 to 419) | .37 | .97 |
| Inpatient hospital | 939 (6508) | 1151 (9228) | −234.10 (−706 to 238) | .33 | .96 | −368.63 (−1092 to 355) | .32 | .95 |
| Emergency department | 615 (2289) | 527 (1750) | 78.49 (−103 to 260) | .39 | .97 | 123.60 (−159 to 407) | .39 | .97 |
| Urgent care | 20 (71) | 26 (109) | −5.73 (−13 to 2) | .14 | .79 | −9.03 (−21 to 3) | .13 | .78 |
| Other | 75 (1092) | 117 (1336) | −42.22 (−105 to 20) | .18 | .88 | −66.48 (−162 to 29) | .17 | .86 |
| Out-of-pocket spending | 780 (1219) | 778 (1208) | −7.93 (−113 to 97) | .88 | >.99 | −12.49 (−175 to 151) | .88 | >.99 |
| Medical utilization | | | | | | | | |
| Physician visits | 3.4 (4.1) | 3.2 (4.1) | 0.11 (−0.2 to 0.4) | .44 | .97 | 0.17 (−0.3 to 0.6) | .44 | .97 |
| Hospitalizations | 0.1 (0.3) | 0.1 (0.3) | −0.02 (−0.03 to 0.0) | .08 | .67 | −0.02 (−0.1 to 0.0) | .07 | .64 |
| Emergency department visits | 0.3 (0.8) | 0.3 (0.7) | 0.02 (0.0 to 0.1) | .47 | .97 | 0.0 (−0.1 to 0.1) | .46 | .97 |
| Urgent care visits | 0.1 (0.4) | 0.1 (0.5) | −0.02 (−0.1 to 0.0) | .40 | .97 | −0.03 (−0.1 to 0.0) | .39 | .97 |
| Preventive care visits | 0.4 (0.6) | 0.4 (0.6) | 0.01 (−0.1 to 0.1) | .85 | >.99 | 0.01 (−0.1 to 0.1) | .85 | >.99 |
| Pharmaceutical spending, $[e] | | | | | | | | |
| Total spending | 1412 (5872) | 1215 (7424) | 179.40 (−245 to 603) | .40 | .99 | 282.50 (−378 to 943) | .40 | .99 |
| Out-of-pocket spending | 102 (162) | 94 (170) | 7.05 (−5 to 19) | .26 | .93 | 11.09 (−8 to 30) | .25 | .93 |
| Pharmaceutical utilization | | | | | | | | |
| Any medications, % | 60.9 (48.8) | 58.5 (49.3) | 2.09 (−1.3 to 5.5) | .23 | .93 | 3.29 (−2.0 to 8.6) | .22 | .93 |
| Distinct medications | 4.3 (4.8) | 4.0 (4.7) | 0.25 (−0.1 to 0.6) | .12 | .80 | 0.40 (−0.1 to 0.9) | .12 | .80 |
| Medication months (≤18) | 11.8 (19.9) | 11.0 (19.7) | 0.60 (−0.9 to 2.1) | .41 | .99 | 0.95 (−1.3 to 3.2) | .42 | .99 |
| By clinical category | | | | | | | | |
| Any asthma medications, % | 13.8 (34.5) | 11.8 (32.2) | 2.05 (−0.7 to 4.8) | .15 | .85 | 3.22 (−1.1 to 7.6) | .15 | .85 |
| Asthma medication, mo | 0.5 (1.8) | 0.5 (2.5) | 0.01 (−0.2 to 0.1) | .87 | >.99 | −0.02 (−0.2 to 0.2) | .86 | >.99 |
| Any cardiovascular medications, % | 23.0 (42.1) | 22.3 (41.6) | 0.40 (−2.4 to 3.2) | .78 | >.99 | 0.63 (−3.7 to 5.0) | .78 | >.99 |
| Cardiovascular medication, mo | 2.6 (6.7) | 2.6 (6.5) | −0.01 (−0.5 to 0.5) | .98 | >.99 | −0.01 (−0.7 to 0.7) | .98 | >.99 |
| Any diabetes medications, % | 7.9 (26.9) | 7.1 (25.6) | 0.56 (−1.2 to 2.3) | .53 | >.99 | 0.89 (−1.9 to 3.6) | .53 | >.99 |
| Diabetes medication, mo | 1.0 (4.4) | 1.0 (4.5) | 0.06 (−0.2 to 0.4) | .71 | >.99 | 0.09 (−0.4 to 0.6) | .71 | >.99 |
| Any hyperlipidemia medications, % | 14.1 (34.8) | 14.0 (34.7) | −0.27 (−2.4 to 1.9) | .80 | >.99 | −0.43 (−3.8 to 2.9) | .80 | >.99 |
| Hyperlipidemia medication, mo | 1.1 (3.4) | 1.1 (3.5) | 0.0 (−0.2 to 0.2) | .70 | >.99 | −0.06 (−0.4 to 0.2) | .69 | >.99 |
| Any mental health medications, % | 18.8 (39.1) | 17.4 (37.9) | 1.20 (−2.2 to 4.6) | .49 | >.99 | 1.89 (−3.4 to 7.2) | .48 | >.99 |
| Mental health medication, mo | 1.8 (5.3) | 1.6 (5.3) | 0.13 (−0.2 to 0.5) | .47 | >.99 | 0.20 (−0.4 to 0.8) | .47 | >.99 |
| Any pain medications, % | 20.1 (40.1) | 17.6 (38.1) | 2.43 (−0.4 to 5.2) | .09 | .71 | 3.82 (−0.4 to 8.1) | .08 | .68 |
| Pain medication, mo | 0.8 (2.7) | 0.8 (2.7) | 0.02 (−0.1 to 0.2) | .76 | >.99 | 0.0 (−0.2 to 0.3) | .75 | >.99 |
| Any antibiotics, % | 12.9 (33.5) | 12.8 (33.5) | −0.18 (−2.8 to 2.4) | .89 | >.99 | −0.28 (−4.4 to 3.8) | .89 | >.99 |
| Antibiotics medication, mo | 0.4 (1.3) | 0.4 (1.6) | 0.03 (−0.1 to 0.1) | .57 | >.99 | 0.05 (−0.1 to 0.2) | .57 | >.99 |
| Any other medications, % | 38.0 (48.6) | 34.3 (47.5) | 3.42 (0.3 to 6.5) | .03 | .43 | 5.39 (0.6 to 10.2) | .03 | .45 |
| Other medication, mo | 3.6 (7.5) | 3.1 (7.1) | 0.42 (−0.3 to 1.1) | .22 | .93 | 0.66 (−0.4 to 1.7) | .22 | .93 |

[a] This table reports intent-to-treat and local average treatment effect estimates of the wellness program at the employee level. All regressions included demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, and job characteristics at the time they initially appeared in the data including full-time status, paid hourly status, and job category) and clustered SEs at the worksite level. Regressions in this table do not control for Cigna coverage status as the data come from medical and pharmaceutical claims of these employees; thus, all had Cigna coverage. Regressions and means were weighted by the combination of a weight for exposure to the program and a weight that balances treatment and control samples on demographics (age, sex, and race). This sample includes medical and pharmaceutical claims data collected from the 20 treatment and all 140 primary and secondary control worksites continuously across the study period. Because all individuals with Cigna coverage were included in these analyses, the sample size across all outcomes in this domain was 7631 (23.2% of individuals employed at any time in the 160 worksites in the study period).

[b] The number of observations for each outcome was 1005 (24.9% of those employed at worksites at any time in the study period).

[c] Primary and secondary control worksites number of observations for each outcome was 6626 (22.9% of those employed at worksites at any time in the study period).

[d] For variables measured as percentages in group means, the difference attributable to the wellness program is expressed in percentage points.

[e] Reported as dollars per individual per year, adjusted for inflation to 2016 dollars.

Table 5. Mean Values and Effect of Program on Employment Outcomes[a]

| | Group Mean (SD) | | Effect of Availability of Wellness Program (Intention to Treat) | | | Effect of Participation in Wellness Program (Local Mean Treatment Effect) | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Treatment[b] | Control[c] | Effect (95% CI)[d] | *P* Value | Adjusted *P* Value | Effect (95% CI)[d] | *P* Value | Adjusted *P* Value |
| Absenteeism, % of scheduled hours missed | 2.5 (1.6) | 2.6 (1.6) | −0.1 (−0.3 to 0.0) | .09 | .21 | −0.3 (−0.5 to 0.0) | .08 | .20 |
| Performance review, % with a score better than 3 out of 5[e] | 60.6 (48.9) | 60.5 (48.9) | −0.5 (−8.3 to 7.4) | .91 | .92 | −0.8 (−14.0 to 12.4) | .91 | .92 |
| Tenure, days employed during the treatment period[f] | 305.9 (213.1) | 308.8 (212.6) | −5.6 (−18.8 to 7.7) | .41 | .45 | −15.8 (−53.1 to 21.5) | .41 | .45 |

[a] This table reports intent-to-treat and local average treatment effect estimates of the wellness program at the employee level. All regressions included demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, Cigna coverage status, and job characteristics at the time they initially appeared in the data including full-time status, paid hourly status, and job category) and clustered SEs at the worksite level. Regressions and means for tenure were weighted to balance treatment and control groups on demographics. Regressions and means for absenteeism and performance review were weighted by the combination of this weight and a weight for exposure to the wellness program. Multiple inference adjustment was performed for absenteeism and performance review using the family-wise *P* values. Due to the difference in weights, tenure was excluded from multiple inference adjustment. These administrative data were collected from the 20 treatment and all 140 primary and secondary control worksites continuously across the study period. Due to variation in the number of performance reviews that employees received, including those who did not receive a performance review during the study period, the sample sizes were 32 974 (100% of all employees) for absenteeism and tenure, and 24 054 (73.0% of all employees) for performance reviews.

[b] The number of observations for each outcome ranged between 2975 and 4037 (73.7% to 100.0% of the employees who were employed at these worksites at any time during the intervention).

[c] Primary and secondary control worksites number of observations for each outcome (range, 21 079-28 937 [72.8%-100.0% of the employees who were employed at these control worksites at any time during the study].)

[d] For variables measured as percentages in group means, the difference attributable to the wellness program is expressed in percentage points.

[e] Score range, 1 to 5 (1 [best performance] to 5 [the poorest]). Given natural variation in the number of performance reviews received during the study period across individuals, this outcome averaged available performance review scores for each individual (weighted by the duration over which a score was held). Binary outcome measure score definitions (<3, good performance and ≥3, poor performance). About 40% of the employees scored an average performance (1-3).

[f] The study period was defined as January 4, 2015, through July 2, 2016. Thus, the maximum number of days employed (tenure) during the study period was 546.

a mean of $102 per year in the treatment group and $94 in the control group. Randomization into a treatment worksite did not have a detectable effect on health care spending or utilization (all *P* values >.05).

## Employment Outcomes

Table 5 shows results for absenteeism, work performance, and job tenure, derived from the full sample of 32 974 employees (for absenteeism and tenure, where there were no missing data) and 24 054 for work performance (73% of employees had a performance review). Workers were absent (sick or personal time) for a mean of 2.5% of scheduled hours in the treatment group vs 2.6% in the control group. Employees scored better than 3 out of 5 on their job performance review 60.6% of the time in the treatment group vs 60.5% in the control group. Workers were employed for a mean of 305.9 total days during the study period in the treatment group vs 308.8 in the control group. Randomization into treatment had no effect on absenteeism, work performance, or tenure (all *P* values >.05) (Table 5).

## Local Average Treatment Effects

For self-reported health and behaviors, participation in the wellness program (defined by participation in at least 1 module) led to a higher share who reported regular exercise (10.6-percentage point difference; 95% CI, 5.3-16.0; unadjusted *P* < .001, adjusted *P* = .03) and higher share actively managing weight (17.2-percentage-point difference; 95% CI, 9.1-25.4; unadjusted *P* < .001, adjusted *P* = .01) compared with the control group. No other outcome in this domain was significantly affected by program participation. The standardized treatment effect showed that health behaviors were 0.09 SD better (95% CI, 0.03-0.13; unadjusted *P* = .001) for wellness program participants (Table 2).

Participation in the program led to no detectable effects on clinical measures of health, with a standardized treatment effect of −0.04 SD (−0.12 to 0.04; unadjusted *P* = .36) (Table 3). There were also no detectable effects on health care spending or utilization or employment outcomes (all *P* values >.05) (Tables 4 and 5).

## Heterogeneity Analyses

Program effects among prespecified and key outcomes were not significantly different between men and women (*P* for interaction >.05; eTable 11A in Supplement 2). However, the increase in regular exercise was driven by workers aged 40 years or older (*P* for interaction = .01; eTable 11B in Supplement 2).

## Secondary and Sensitivity Analyses

When alternative definitions of participation were used, the effect of participation was numerically greater among participants who completed at least 3 modules than those who completed at least 1 module, although most estimates were not statistically significant (eTable 12 in Supplement 2).

Estimates using the stably employed subsample were similar to those from the full sample (eTables 5-9 in Supplement 2). Analyses of spending, utilizations, and employment outcomes at the worksite level yielded similar results to those obtained at the individual level (eTables 7-9 in Supplement 2). Estimates of program effect using only exposure weights produced similar estimates to the main findings (eTable 1 in Supplement 2). For binary outcomes, estimates using logistic regressions were similar to those using linear models (eTable 14 in Supplement 2).

### Selection Into Program Participation

Comparisons of preintervention characteristics between participants and nonparticipants in the treatment group provided evidence of potential selection effects. Participants were significantly more likely to be female, nonwhite, and full-time salaried workers in sales, although neither mean health care spending nor the probability of having any spending during the year before the program was significantly different between participants and nonparticipants (eTable 15 in Supplement 2). There was no evidence of differential selection into completion of surveys or biometrics between treatment and control groups on baseline covariates (eTable 16 in Supplement 2).

Moreover, an observational approach comparing workers who elected to participate with nonparticipants would have incorrectly suggested that the program had larger effects on some outcomes than the effects found using the controlled design, underscoring the importance of randomization to obtain unbiased estimates (eTable 17 in Supplement 2).

## Discussion

This randomized clinical trial of a multiyear, multicomponent workplace wellness program implemented in a middle- and lower-income population found that individuals in workplaces where the program was offered reported better health behaviors, including regular exercise and active weight management, but the program did not generate differences in clinical measures of health, health care spending or utilization, or employment outcomes after 18 months.

That the program affected self-reported health behaviors, but not health or economic outcomes, may be interpreted in several ways. Given that workplace wellness programs focus on changing behavior and that behavior change may precede improvements in other outcomes, these findings could be consistent with future improvements in health or reductions in spending. On the other hand, behavior change is likely easier to achieve than improvements in clinical or employment outcomes. Thus, there may remain no detectable effects on those outcomes, which would have implications for the return on investment in wellness programs.

The finding of no significant effects on clinical measures of health, health care spending, or employment outcomes is consistent with a recent trial of a wellness program implemented at the University of Illinois, which evaluated similar outcomes after 1 year.[15] However, our study found a sizeable and robust improvement in some self-reported health behaviors. Moreover, we found that participants did not have lower preintervention spending than nonparticipants, although there was selection on other dimensions. Unlike the Illinois study, this intervention was implemented at the worksite level (rather than varying across individuals within the same worksite), perhaps better facilitating changes in workplace culture and providing greater social supports for behavior change. This intervention was also fielded in a different population, set of geographies, and employment setting, making it difficult to isolate the causes of any differences in findings.

These findings stand in contrast with much of the prior literature on workplace wellness programs, which tended to find positive and often large returns on investment through, for example, reductions in absenteeism and health care spending.[3-9,23,24] Given that most prior studies were based on observational designs with methodological shortcomings such as potential selection bias, results based on random assignment of the intervention are likely more reliable.

### Limitations

This study has several limitations. First, although this population was diverse, results may not generalize to other workplace settings or populations. Second, the ability to detect treatment effects was limited by statistical power, despite prespecified strategies to maximize power. This challenge was augmented by our very conservative approach to multiple-inference adjustment, which grouped a wide array of outcomes (rather than narrowly construing related outcomes). It was further limited by employee turnover that restricted the workers present to participate in end-of-study primary data collection, although the mean duration of employment was similar among the 3 groups of the trial (Figure), suggesting that entry and exit from the sample was due to natural exogenous employment turnover, not the wellness program.

Third, not all employees contributed data for every outcome. Survey and biometric data were available only for individuals employed at the 18-month mark who chose to participate in primary data collection. However, there was no evidence of differential selection into completing the survey and screening. Claims data were available only for employees with Cigna coverage, although no data were missing in this sample. All individuals contributed employment outcomes, except performance reviews, which represented 74% and 73% of employees in the treatment and control groups, respectively. Overall, all available data on employees were analyzed; rates of missing data were similar between groups and may thus have affected the precision of estimates but do not seem to have adversely affected the validity of the findings.

Fourth, this study was unable to disentangle effects of particular elements of the wellness program, nor assess the effects of a differently configured wellness program. Rather, it evaluated the program as a package, with implementation that varied only idiosyncratically in small ways across worksites. Such design features are in fact common in most wellness programs.[3-6]

## Conclusions

Among employees of a large US warehouse retail company, a workplace wellness program resulted in significantly greater rates of some positive self-reported health behaviors among those exposed compared with employees who were not exposed, but there were no significant differences in clinical measures of health, health care spending and utilization, and employment outcomes after 18 months. Although limited by incomplete data on some outcomes, these findings may temper expectations about the financial return on investment that wellness programs can deliver in the short term.

ARTICLE INFORMATION

**Accepted for Publication:** March 6, 2019.

**Correction:** This article was corrected on April 16, 2019, for data errors in the Abstract and Figure and for omissions to the Additional Contributions section.

**Author Contributions:** Drs Song and Baicker had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
*Concept and design, acquisition, analysis, or interpretation of data, drafting of the manuscript, critical revision of the manuscript for important intellectual content, statistical analysis, obtained funding, administrative, technical, or material support, and supervision:* Both authors.

**Conflict of Interest Disclosures:** Dr Song reported no disclosures. Dr Baicker reported receiving personal fees from Eli Lilly outside the submitted work and reported serving on the board of directors of Eli Lilly.

**Funding/Support:** This work was supported by the National Institute on Aging (R01 AG050329; P30 AG012810 through the National Bureau of Economic Research), Robert Wood Johnson Foundation (grant 72611), and Abdul Latif Jameel Poverty Action Lab North America. BJ's Wholesale Club provided in-kind logistical and personnel support for the fielding of the wellness program.

**Role of the Funder/Sponsor:** The funders had no role in the design or conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

**Additional Contributions:** We thank José Zubizarreta, PhD, Harvard Medical School, for his statistical guidance and contributions to the sample weights, without financial compensation. We thank Sherri Rose, PhD, Harvard Medical School, for her statistical guidance on randomization, without financial compensation. We thank David Molitor, PhD, and Julian Reif, PhD, University of Illinois at Urbana-Champaign, for guidance on the statistical software for multiple inference adjustment they created in the University of Illinois wellness study, which was used in this study, without financial compensation. We thank Ozlem Blakeley, MA, an employee of Harvard Medical School, and Kathryn Clark, BA, BS, and Bethany Maylone, MEd, employees of the Harvard T.H. Chan School of Public Health, for research assistance and project management. We also thank Josephine Fisher, BA, Jack Huang, AB, Harlan Pittell, BS, and Artemis (Yuanxiaoyue) Yang, BA, employees of the Harvard T.H. Chan School of Public Health, for research assistance. We thank Luke Sonnet, BS, University of California, Los Angeles, for replicating the study results through the Abdul Latif Jameel Poverty Action Lab's Research Transparency and Reproducibility Initiative, without financial compensation. We thank the study partners, BJ's Wholesale Club, and Wellness Workdays for collaboration and assistance in the design and fielding of the workplace wellness program. We thank seminar participants at the 7th Conference of the American Society of Health Economists, the Department of Nutrition at the Harvard T.H. Chan School of Public Health, and the Department of Health Care Policy at Harvard Medical School for comments and suggestions, without financial compensation.

**Data Sharing Statement:** See Supplement 3.

## REFERENCES

**1**. Kaiser Family Foundation. 2018 Employer Health Benefits Survey. https://www.kff.org/health-costs/report/2018-employer-health-benefits-survey/. Published October 3, 2018. Accessed February 19, 2019.

**2**. Pollitz K, Rae M. Workplace wellness programs: characteristics and requirements. Kaiser Family Foundation. https://www.kff.org/private-insurance/issue-brief/workplace-wellness-programs-characteristics-and-requirements/. Published May 19, 2016. Accessed October 4, 2018.

**3**. Mattke S, Schnyer C, Van Busum KR. A review of the U.S. workplace wellness market. RAND Corporation. https://www.rand.org/pubs/occasional_papers/OP373.html. Published November 27, 2012. Accessed October 4, 2018.

**4**. Baicker K, Cutler D, Song Z. Workplace wellness programs can generate savings. *Health Aff (Millwood)*. 2010;29(2):304-311. doi:10.1377/hlthaff.2009.0626

**5**. Goetzel RZ, Henke RM, Tabrizi M, et al. Do workplace health promotion (wellness) programs work? *J Occup Environ Med*. 2014;56(9):927-934. doi:10.1097/JOM.0000000000000276

**6**. Goetzel RZ, Ozminkowski RJ. The health and cost benefits of work site health-promotion programs. *Annu Rev Public Health*. 2008;29:303-323. doi:10.1146/annurev.publhealth.29.020907.090930

**7**. Chapman LS; American Journal of Health Promotion Inc. Meta-evaluation of worksite health promotion economic return studies: 2005 update. *Am J Health Promot*. 2005;19(6):1-11. doi:10.4278/0890-1171-19.4.TAHP-1

**8**. Pelletier KR. A review and analysis of the clinical and cost-effectiveness studies of comprehensive health promotion and disease management programs at the worksite: update VIII 2008 to 2010. *J Occup Environ Med*. 2011;53(11):1310-1331. doi:10.1097/JOM.0b013e3182337748

**9**. Fries JF, Harrington H, Edwards R, Kent LA, Richardson N. Randomized controlled trial of cost reductions from a health education program: the California Public Employees' Retirement System (PERS) study. *Am J Health Promot*. 1994;8(3):216-223. doi:10.4278/0890-1171-8.3.216

**10**. Leigh JP, Richardson N, Beck R, et al; The Bank of American Study. Randomized controlled study of a retiree health promotion program. *Arch Intern Med*. 1992;152(6):1201-1206. doi:10.1001/archinte.1992.00400180067010

**11**. Volpp KG, John LK, Troxel AB, Norton L, Fassbender J, Loewenstein G. Financial incentive-based approaches for weight loss: a randomized trial. *JAMA*. 2008;300(22):2631-2637. doi:10.1001/jama.2008.804

**12**. Halpern SD, French B, Small DS, et al. Randomized trial of four financial-incentive programs for smoking cessation. *N Engl J Med*. 2015;372(22):2108-2117. doi:10.1056/NEJMoa1414293

**13**. Volpp KG, Troxel AB, Pauly MV, et al. A randomized, controlled trial of financial incentives for smoking cessation. *N Engl J Med*. 2009;360(7):699-709. doi:10.1056/NEJMsa0806819

**14**. Cahill K, Hartmann-Boyce J, Perera R. Incentives for smoking cessation. *Cochrane Database Syst Rev*. 2015;(5):CD004307.

**15**. Jones D, Molitor D, Reif J. What do workplace wellness programs do? evidence from the Illinois Workplace Wellness Study. NBER Working Paper Series. 2018; 24229.

**16**. Mello MM, Rosenthal MB. Wellness programs and lifestyle discrimination—the legal limits. *N Engl J Med*. 2008;359(2):192-199. doi:10.1056/NEJMhle0801929

**17**. Ware JE, Kosinski M, Dewey JE, Gandek B. How to score and interpret single-item health status measures: a manual for users of the SF-8 Health Survey. *QualityMetric Inc.* 2001;15(10):5.

**18**. DeMets DL, Cook T. Challenges of non-intention-to-treat analyses. *JAMA*. 2019;321(2):145-146. doi:10.1001/jama.2018.19192

**19**. Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. *J Am Stat Assoc*. 2015;110(511):910-922. doi:10.1080/01621459.2015.1023805

**20**. Wang X, Zubizarreta JR. Minimal approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*. 2017;103(1):1-22. doi:10.1093/biomet/asx011

**21**. Hirshberg DA, Zubizarreta JR. On two approaches to weighting in causal inference. *Epidemiology*. 2017;28(6):812-816. doi:10.1097/EDE.0000000000000735

**22**. Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for P Value Adjustment*. New York, NY: Wiley & Sons; 1993.

**23**. Ozminkowski RJ, Dunn RL, Goetzel RZ, Cantor RI, Murnane J, Harrison M. A return on investment evaluation of the Citibank, N.A., health management program. *Am J Health Promot*. 1999;14(1):31-43. doi:10.4278/0890-1171-14.1.31

**24**. Bly JL, Jones RC, Richardson JE. Impact of worksite health promotion on health care costs and utilization: evaluation of Johnson & Johnson's Live for Life program. *JAMA*. 1986;256(23):3235-3240. doi:10.1001/jama.1986.03380230059026