The Impact of Workplace Wellness on Health Care Spending, Health, and Employment Outcomes: A Randomized Controlled Trial[†]

Zirui Song[‡], Katherine Baicker[§]

Analysis Plan—Phase 2
January 7, 2020

ClinicalTrials.gov Identifier: NCT03167658

American Economic Association RCT Registry ID: AEARCTR-0000586

[†] We thank Ozlem Barin and Erica Paulos for excellent research assistance in the preparation of this analysis plan, and Jose Zubizarreta and Sherri Rose for their guidance on statistical methodology. We thank Anna Sinaiko for helpful comments on a prior version of the analysis plan. We are also grateful to Kathryn Clark, Yuanxiaoyue Yang, Jack Huang, Harlan Pittell, and Josephine Fisher for excellent research assistance and to Bethany Maylone for expert project management in prior years. In addition, we thank our study partners, BJ's Wholesale Club and Wellness Workdays, for their collaboration and expertise in designing and fielding the intervention. We gratefully acknowledge funding from the National Institute on Aging, the Abdul Latif Jameel Poverty Action Lab North America, and the Robert Wood Johnson Foundation.
[‡] Harvard Medical School, Massachusetts General Hospital
[§] University of Chicago Harris School of Public Policy, Harvard T.H. Chan School of Public Health, NBER

Table of Contents

This document describes the second phase of our planned analysis of a randomized controlled trial of a workplace wellness program, which pertains to outcomes over three years. We have previously planned, archived, implemented, and published analysis covering the first 18 months of this intervention.[1,2]

I. Introduction

Workplace wellness programs have become increasingly popular across the U.S. Centered on awareness, education, and the promotion of healthy behaviors for disease prevention, workplace wellness programs comprised a $7.8 billion industry in 2016. In the face of rising health care costs for their employees, about 50 percent of small firms and 84 percent of large firms in the U.S. now offer a wellness program, often focused on smoking cessation, weight management, or behavioral or lifestyle coaching; firms also frequently offer employees a health risk assessment and biometric screening.[3] In addition to this substantial private sector investment, the growth of workplace wellness has been aided by public investments such as the Affordable Care Act, which provided incentives for the establishment of such programs. Despite the importance of workplace wellness programs for U.S. workers, employees, and the government, little rigorous evidence exists on the effect of such programs on health and economic outcomes.

Prior studies of workplace wellness programs, largely observational in nature, have been plagued by selection bias, lack of control groups, and small samples. Participants in wellness programs and firms offering them are likely different from their non-participants in important observed and unobserved ways that affect health outcomes. Thus, it has been difficult to identify the effect of such programs using observational studies comparing participants to non-participants. Moreover, meta-analyses have produced widely varying estimates of program benefits relative to costs.

Through a partnership with a large multi-state U.S. employer (BJ's Wholesale Club) and an experienced and award-winning wellness vendor (Wellness Workdays), we implemented a randomized controlled trial of a workplace wellness program beginning in 2015. We have already analyzed results from the first 18 months of the intervention (phase 1), in which we found no statistically significant effects on health care spending or objective measures of health, but did find significant effects on self-reported health behaviors, including regular exercise and active weight management.[2] Our study, which randomized worksites into treatment and control arms, complemented another randomized controlled evaluation of a workplace wellness program at the University of Illinois, which randomized employees at the individual level, and also found no effects on spending or health outcomes after the first year.[4]

---

[1] The Impact of Employee Wellness Programs. NIH ClinicalTrials.gov. U.S. National Library of Medicine. 2017 May 30. (https://clinicaltrials.gov/ct2/show/NCT03167658); The Impact of Employee Wellness Programs: A Randomized Controlled Trial. American Economic Association Randomized Controlled Trial Registry. 2015 Feb 3. (https://www.socialscienceregistry.org/trials/586/history/26720)

[2] Song Z, Baicker K. Effect of a Workplace Wellness Program on Employee Health and Economic Outcomes: A Randomized Clinical Trial. JAMA. 2019 Apr 16;321(15):1491-1501.

[3] The Kaiser Family Foundation. Employer Health Benefits: 2019 Annual Survey. 2019. (https://www.kff.org/health-costs/report/2019-employer-health-benefits-survey/)

[4] Jones D, Molitor D, Reif J. What do Workplace Wellness Programs do? Evidence from the Illinois Workplace Wellness Study. Q J Econ. 2019 Nov;134(4):1747-1791.

This second analysis plan details the evaluation of this intervention and specific methodology through the three-year period spanning January 2015 through December 2017. This analysis evaluates the impact of the workplace wellness program on employee health care spending and utilization, health outcomes, employment, and productivity.

This analysis plan seeks to pre-specify the analysis before comparing outcomes for treatment and control groups, in order to minimize issues of data mining and specification searching. To create this document, we examined data on outcomes for the control group and performed limited comparisons of non-outcome variables between the treatment and control groups (such as pre-randomization demographics). However, we have not conducted any analysis of differences in outcomes between the treatment and control groups post-treatment over this three-year window. Institutional review board approval was granted and maintained through Harvard University.

II. Treatment

The treatment was a longitudinal multi-component workplace wellness program designed to improve the health and wellbeing of workers. It took place at BJ's Wholesale Club, the largest warehouse retail corporation in the Eastern U.S. and third largest warehouse retail company in the country, with approximately 25,000 employees serving 9 million members. BJ's operates about 200 worksites or "clubs" from Maine to Florida and has a demographically and socioeconomically diverse workforce across a variety of work settings.

The wellness program took place in 2 phases. Phase 1 of the treatment period spanned 18 months, from January 2015 through June 2016. In phase 2 of the study, the treatment period was extended for another 18 months, from January 2015 through December 2017 and is the focus of this analysis plan (**Table 1**). In phase 2, 5 additional treatment worksites and 5 additional primary control worksites were added to the experiment. This wellness program was designed and implemented by a third-party vendor, Wellness Workdays. Wellness Workdays is a wellness vendor that delivers and manages wellness programs across many industries, including finance, manufacturing, banking, higher education, and legal across a number of states.

The wellness program consisted of a personal health assessment (PHA), in-person screenings, and multiple program modules. Each module took place over 4-10 consecutive weeks. The modules centered on themes such as team-based and individual wellness challenges, nutrition, stress reduction, and physical activity, as well as workplace culture. For each module, participants could earn modest financial rewards ($25-50) for completion.

The 3-year intervention period comprised 12 modules. Eight modules were offered during the first 18 months (phase 1): "Take Charge of Your Health," which taught proactive strategies for participating in health and health care; "Nutrition for a Lifetime," which aimed to help employees achieve and maintain a healthy weight through nutrition; "Club Cardio Challenge" (consisting of 2 modules), which focused on cardiovascular activity; "Maintain Don't Gain," which combined principles of healthy nutrition with physical activity; "Power Down the Pressure," which taught methods for stress management; "Weight Loss Boot Camp," which

focused on nutrition and exercise methods for weight loss; and "Movin' in May," which once again focused on physical activity with active tracking of progress.

Four modules were offered during the next 18 months (phase 2): "Healthy and Fit," which combined healthy eating patterns and regular exercise to avoid weight gain during the holiday season; "Step Challenge," which was a team-based module, focusing on increasing physical activity by using a fitness device to track steps; and two consecutive "Nutrition for a Lifetime" modules, which were designed to help employees achieve and maintain a healthy lifestyle by combining healthy nutrition, exercise, stress management and sleep patterns. In addition, participants had the opportunity to log their step counts using a wearable device and this information was gathered between January 2017 and July 2017. Please refer to Appendix 1 for detailed information on the components of the wellness program by module, including requirements and incentives. Employees had opportunities to receive incentive payments through completion of the PHA and the biometric screenings, in addition to participation in the individual modules of the program.

In each treatment worksite, a Registered Dietitian employed by Wellness Workdays coordinated and led the wellness programming. The Registered Dietitians engaged employees in the wellness program modules, educated them about the content of the program, and led employees in various creative activities such as group fitness and cooking demonstrations. Each Registered Dietitian had the flexibility to tailor the day-to-day programming around the themes of the modules. The Registered Dietitians spent approximately 8 hours per week at each worksite.

III. Randomization

The wellness program was implemented in a randomly selected subset of BJ's Wholesale Clubs. Each club is a standalone worksite, with an average of 118 employees at any given time. At the beginning of the study, there were 201 BJ's worksites in the U.S. along the East coast, extending from Maine to Florida. We eliminated 41 worksites because they were geographically remote or had employee pools with substantially different insurance coverage from the others, leaving 160 worksites in our sample.

Among these worksites, we randomly selected 25 "treatment" worksites that would receive the wellness program and 25 "primary control" worksites for phase 2.[5] These primary control worksites participated in data collection through PHAs and in-person biometric screenings, but did not receive the wellness program treatment. **Figure 1** shows the locations of the 25 treatment and 25 control worksites across the Eastern U.S. The remaining 110 worksites served as "secondary controls." which were included in analyses of administrative data. Thus, the primary data described below was collected from the 50 treatment and primary control worksites; administrative data were compiled for all 160 worksites. **Figure 2** shows a Consolidated Standards of Reporting Trials (CONSORT) flow diagram for our phase 2 evaluation, including the sample sizes of treatment and control groups.

[5] For the first 18 months of the study (phase 1), there were 20 treatment and 20 primary control sites.

Randomization at the worksite level allows us to assess the effects of interventions that aim to affect workplace culture or operate on a group level, such as using common break rooms to stage nutrition demonstrations, team-based activities, or worksite level prizes.

IV. Data

Our analysis focuses on outcomes from five data categories. The Table below displays the source of each data set and the study population from which it is derived.

Summary of the Components of Data Collected				
Data	**Source**	**Treatment Worksites (25)**	**Primary Control Worksites (25)**	**Secondary Control Worksites (110)**
Administrative Data				
Employment records	BJ's	All employees	All employees	All employees
Claims data (medical and pharmaceutical)	Cigna (via BJ's)	Employees insured by Cigna	Employees insured by Cigna	Employees insured by Cigna
Primary Data				
Biometric screening data	Wellness Workdays	Employees completing screening	Employees completing screening	None (by design)
Personal Health Assessment	Wellness Workdays	Employees completing survey	Employees completing survey	None (by design)
Participation in the treatment	Wellness Workdays	All employees	None (by design)	None (by design)

A. Administrative Data

Administrative data consist of employment records and medical and pharmaceutical claims data. Employment records are available for all employees across treatment and control worksites (both primary and secondary controls). Medical and pharmaceutical claims data are available through Cigna for BJ's employees who are insured through a Cigna plan. In cross-section, approximately 37 percent of all BJ's employees were insured through Cigna during June 2016 (the midpoint of phase 2). Of note, BJ's as an organization is self-insured (i.e. it bears risk of health care spending of its enrolled members), with Cigna as the administrator of its health plans. We used all administrative data available for the entire study period.

A1. Employment records

BJ's provided us with data from their earnings database, which includes data elements that we use both to define our sample (based on hire and termination dates) and to measure key outcomes, such as absenteeism and performance reviews. Information included:

- **Dates and location of employment actions**: Dates and sites of hire, rehire, termination, transfer, and performance review. These data capture employment history across clubs. A small proportion of employees (4.5%) appeared in more than 1 worksite during the study period. We defined these workers' treatment or control assignment using the randomized assignment of their initial worksite, as the decision to leave (or stay) at a worksite could be endogenous. Most employees have one performance review per calendar year.

- **Demographic characteristics**: Characteristics of the working population include age, sex, and race. These are discussed in greater in the section on Analysis below. Worksite level analyses, as described below, were augmented using 2015 estimates of county-level demographic characteristics from the U.S. Census Bureau matched to the locations of each worksite.

- **Employment characteristics**: Characteristics of an individual's job include type of hired arrangement (full time salaried, full time hourly, and part time hourly) and type of job. The types of jobs were grouped into three categories: sales, non-sales (including laborers and helpers, operatives, and service workers), and other (including administrative support, craft workers, and first/mid-level office professionals).

- **Hours and earnings**: Each employee's hours (including working hours, vacation, sick time, personal time) and earnings in each pay period. Each observation is at the employee, pay-period, earnings-type level. These data include a handful of negative payments/hours that indicate corrections.

We used these data to study several outcomes:

- **Absenteeism:** We calculated absenteeism as an employee's number of sick plus personal hours, divided by the sum of an employee's number of sick, personal, and worked hours. This gives the ratio of absence relative to how much an employee usually works. Vacation and holiday time are excluded from both the numerator and denominator.

- **Performance review:** Employee performance reviews were coded on a 5-point scale, with 1 representing the best performance rating and 5 the worst. We averaged performance review scores (weighted by the duration of time over which a score held) and created a binary indicator where a score of less than 3 was coded as good performance and 3 and greater as poor performance.

- **Employment tenure:** We calculated each worker's length of tenure at BJ's and how many hours he or she worked (including the nature of those hours, such as regular hours and overtime). We defined tenure as the difference between the hire date and the latest

termination date. If there was no hire date recorded in the data, we used the first appearance in our database. If there was no termination or retirement date in the data, we used the end of the study period.

Table 2 provides a summary of demographic and employment characteristics of the population. The data are presented at the employee level (overall and for analytical subsets) and at the worksite level. **Table 3** provides summary statistics for tenure, performance review, and absenteeism gathered among the control worksites during the treatment period.

A2. Medical and pharmaceutical claims data

Health care claims data were provided at the individual employee level by Cigna and were used to calculate spending and utilization variables. These administrative data are available for all worksites (treatment, primary control, and secondary controls), but only for the roughly 37% of employees who were enrolled in a BJ's employer-sponsored Cigna health insurance plan at any given time. Full time workers were more likely to have Cigna coverage than part-time workers. We analyzed medical claims for BJ's employees (but not their dependents, who were not directly exposed to the treatment).

We aggregated medical spending and utilization at the employee level across our 36-month treatment period. To standardize utilization outcome variables across employees who may have been employed and in our sample for different lengths of time, we normalized outcomes for partial-year enrollees to annual values. For each employee, we exclude any claims with service dates prior to either the employee's hire date or the start of Cigna eligibility status as well as any claims with service dates more than 30 days after the employee's termination or end of Cigna enrollment.

We examined the following outcomes at the individual level as well as the worksite level. In worksite-level analyses, results are shown per 2000 work hours for all the continuous outcomes to normalize across sites of different sizes. An exception was the number of distinct medications, which we constructed at the worksite level using a weighted average of individual values based on hours worked at the worksite.

- **Total medical spending**: We defined total medical spending per year as the sum of all payments made for health care – deductibles, copayments, coinsurance, insurance payments, and "amount paid by other carrier" – that appear on an employee's claims.

- **Medical utilization**: We defined three medical utilization variables (i.e. counts of services provided):

    - *Office Visits:* We defined an office visit as a claim line with site of care as "office" or "on-campus outpatient hospitals" and service type as "physician visit." We counted as one office visit such an occurrence at the level of a unique combination of patient, service date, and provider specialty. In other words, if an office code appears on multiple claim lines on the same date and billed by the same provider specialty, we counted these as a single visit. We do not include office visits that occurred with the site of care as

"inpatient hospital." Similarly, we defined an urgent care visit as a claim line with a site of care as "urgent care facility" and identified preventive care visits by using preventive visit CPT codes in claims.

- *Hospitalizations*: We defined a unique hospitalization as a set of contiguous days on which a patient has a claim line with site of care "inpatient hospital."

- *Emergency Room visits*: ER visits are identified from claim lines with site of care "Emergency Room – Hospital" and service type "emergency facility" or "emergency medical care." We include the service type restriction so that we avoid double counting an ER visit when a claim for imaging, labs, or prescription drug is received a day or two later than the actual ER visit. Similar to hospitalizations above, we treat claims for the same or continuous days as a single emergency room visit.

- **Total prescription drug spending**: We defined total prescription drug spending as the sum of all payments, inclusive of cost-sharing, that appear on an employee's claims at the annual level, scaled using the method described above.

- **Prescription drug utilization**: We defined two pharmaceutical utilization variables:

  - *Number of Distinct Medications:* We defined number of distinct medications as the total count of each unique prescription drug molecule. Unlike the other medical and pharmaceutical outcomes, this outcome is not annualized.

  - *Number of Medication Months:* We used the total quantity and duration of medication supplied to obtain the number of medication months. Because a person can be taking multiple medications at the same time, the number of medication months can be greater than the number of months a person is enrolled.

Table 4 provides summary statistics for medical and pharmaceutical spending and utilization among the control worksites during the treatment period. While substantial additional granularity is available in the claims data, our sample sizes do not in general support condition-specific analyses.

B. Primary Data

Primary data consist of biometric data collected during in-person screenings conducted by registered nurses and self-reported data gathered from PHA surveys. The biometric data includes blood pressure, height and weight (inputs into the calculation of BMI), and blood measurements of cholesterol and blood sugar. PHAs contain self-reported information on health behaviors. Unlike the administrative data, these primary data are available for individuals in the 25 treatment worksites and the 25 primary control worksites who completed the PHA and biometric data collection between 08/01/2017 and 10/31/2017.

B1. Biometric screening data

In the treatment and primary control worksites, we conducted biometric screenings after the conclusion of the last wellness module. The screenings were conducted by registered nurses in the worksites, who met with employees individually. We defined all of the following biometric outcomes as continuous variables.

- Total cholesterol
- High-density lipoprotein (HDL) cholesterol
- Blood glucose (mg/dl)
- Systolic blood pressure
- Body mass index (BMI): weight in kilograms divided by the square of height in meters

Table 5 provides summary statistics for biometric screening data from employees who were screened in the control worksites.

B2. Personal Health Assessment Data

At the conclusion of the wellness modules, we also fielded PHA surveys in each of the treatment and primary control worksites. Employees were asked to complete a paper survey with questions relating to their health behaviors, mental health, and general wellbeing. We use this dataset to assess the impact of the wellness program on employees' self-reported health behaviors and perceived health status.
Based on an examination of the distribution of PHA responses collected from the control group, we examine the following outcome variables.

- **Health behaviors:**

  Screenings and Exams

  - *Percent of recommended tests received*: We pooled commonly recommended tests (based on age and sex) covered in the PHA and determined the share of those tests that respondents reported obtaining. These tests are cholesterol level, fasting blood glucose level, blood pressure, dental exam, colon cancer screening, mammogram, and pap smear.

  Physical activity

  - *Regular exercise:* We defined this as responding yes to the question "Do you engage in regular exercise according to any of the definitions listed?" The provided definitions of regular exercise read "Regular exercise means doing: moderate physical activity that increases your breathing rate and causes you to break a light sweat (such as brisk walking, golf, or raking leaves) for at least 150 minutes (2 hours and 30 minutes) each week OR vigorous physical activity that causes big increases in your breathing and heart rate and makes conversation difficult (such as jogging or running) for at least 75 minutes (1 hour and 15 minutes) each week OR a mix of moderate and

vigorous physical activity that is equal to at least 150 minutes of moderate activity, such as 90 minutes of moderate activity and 30 minutes of vigorous activity each week."[6]

- *Number of hours sitting per day:* This continuous variable was defined as an individual's answer to the categorical question "How many hours per day do you sit? Please consider time at work and at home and include activities such as sitting in front of a computer or television." We use the midpoint value of the chosen category.

Nutrition

- *Number of non-zero calorie drinks per day:* This continuous variable was defined as the difference between an individual's answers to two questions: "How many naturally or artificially sweetened beverages do you consume per day?" and "How many of these beverages are diet or zero calories?"

- *Read the Nutrition Facts panel*: We defined this as responding yes to the question "Do you read the Nutrition Facts panel on food labels?"

- *Consume at least 2 cups of fruit and 2.5 cups of vegetables per day*: We defined this as responding yes to the question "Do you eat at least 2 cups of fruit and $2^1/_2$ cups of vegetables per day?"

Weight Management

- *Actively managing weight:* We defined this as responding yes to either the question "In the past month, have you been actively trying to lose weight?" or "In the past month, have you been actively trying to keep from gaining weight?" or both.

Tobacco use

- *Smoking:* We defined this as responding yes to the question "Do you currently smoke?" or if the answer to "How many cigarettes do you smoke during a typical day?" was greater than zero. The indicator is coded as 0 if the answer to the first question was "non-smoker."

[6] Interpretation of these self-reports can be informed by supplemental data collected from a subset of the treatment group who used an ActivBand wearable step-recording device. This provides an opportunity for us to explore the correlation between actual physical activity and self-reported levels of physical activity on the PHA, although such correlations are merely suggestive, given that the subset of employees with an ActivBand is endogenous based on completion of certain program modules and that the measurement was not comprehensive. Overall, 985 employees in the treatment sites used the ActivBand between January and July of 2017; the average number of steps per person per month was 96,121. On average, individuals who reported engaging in regular exercise logged 32,144 steps more per month than those who reported not engaging in regular exercise (S.E. 8,869; $p<0.001$). Individuals who reported actively managing weight logged 5,021 steps more per month than those who reported otherwise (S.E. 10,063; $p=0.62$). Individuals who reported 3.5 hours or less of sitting per day logged 6,730 steps more per month than those who reported more than 3.5 hours of sitting per day (S.E. 9,732; $p=0.49$).

- **Health and Well-Being:**

    - *SF-8 scores*: We used the 8-question Short Form (SF-8) health survey to measure self-reported functional health and well-being. Each question of the SF-8 uses a 5- or 6-point Likert scale. Its standardized scoring system combines responses into a score that can be interpreted as a continuous variable, with higher scores denoting better self-reported health. Using these scoring methods, we create the two standard scores from this survey: the physical summary score and the mental summary score.[7]

    - *Unmanaged stress:* We defined this as responding no to the question "Do you effectively practice stress management in your daily life?" The question defines stress management to include regular relaxation, physical activity, talking with others, or making time for social activities.

    - *Unmanaged depression:* We defined this as responding "no" or "I have never been depressed" to the question "Depression prevention means using effective methods to keep depression from occurring, or if it does occur, to keep it as mild and brief as possible. Effective methods include controlling negative thinking every day, engaging in healthy, pleasant activities on most days, exercising for 30 minutes or more on most days, practicing stress management on most days, and getting professional help when needed. Do you effectively practice depression prevention in your daily life?"

    - *Stress at work*: We defined this as answering the question "During the last 30 days, how often have you found yourself stressed or worried about problems as work?" with "sometimes," "fairly often," or "very often." (rather than "almost never" or "never").

Table 6 provides summary statistics for employees who completed the PHA in the control worksites.

B3. Program participation data

At the individual employee level, we examine three different definitions of participation based on the number of modules completed. Each module had its own set of requirements that defined completion and incentives attached to participation or completion. Details of the modules and their requirements are provided in Appendix 1.

- **Modules completed**: We defined a continuous variable denoting the number of modules completed, which ranges from 0 to 12 over the course of the study period. This serves as the primary definition of participation for this analysis.

[7] 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. Lincoln RI: QualityMetric Incorporated, 2001.

- **Participation indicator**: We defined a binary indicator of participation based on completing at least one module. This is a secondary definition for this analysis.[8]

- **High participation**: We defined another secondary binary indicator based on completing 3 or more modules. We chose 3 modules as the threshold for high participation based on the distribution of the number of modules completed.

Table 7 shows summary statistics on participation overall and by module across the treatment worksites for phase 1 and for phase 2 of the wellness program.

V. Study Sample

A. Main sample and subsamples

We designed our analyses based on two study samples. Our main study sample comprised all employees who worked for any duration of time at BJ's Wholesale Club during the study period. In other words, we included any worker who worked at BJ's for any length of time during the study period. This is the most inclusive sample.

A potential drawback of these sample inclusion criteria is the possibility of endogenous entry or exit based on the treatment itself. For example, a worker's decision to join or leave employment at BJ's Wholesale Club could be a function of the availability of, or exposure to, the wellness program. Therefore, we also defined a relatively stable subsample of employees who were continuously employed at BJ's in the months before the treatment. This "stably employed" subsample comprised all employees working at BJ's during, at minimum, the 13 consecutive weeks immediately prior to the onset of the treatment—and thus cannot have entered because of the treatment. Exit from this sample might still be endogenous in theory, however. We will therefore gauge the potential bias introduced by any observed differences in exit between treatment and control sites using a bounding exercise. Additional details on the construction of this subsample are provided below and in the appendix.

B. Balance between treatment and control

We tested balance between treatment and control groups on observable baseline characteristics. We examined balance on demographic variables (age, sex, and race) for all of the analytical samples. We also examined balance on job and employment characteristics in the pre-intervention period for the stably employed subsample, as they by construction were in jobs with characteristics defined prior to the intervention and thus exogenous to the treatment status.

As **Table 8** demonstrates at the employee level, our randomly assigned worksites were balanced on many employee characteristics, but not all. Notably, while treatment and control site workers

[8] This indicator was our primary definition of participation in our prior phase 1 analysis. Given that we expanded the sample size for phase 2 (with a different total number of modules available), we believe the continuous measure is preferable here.

were similar on age and gender, a greater proportion of treatment site workers were white (and a lower proportion black and Hispanic) than individuals in control worksites. To address imbalance on some observable characteristics, we followed two common strategies. First, we weighted the treatment and control groups to be more balanced on observed characteristics (such that each group mirrors the overall employee population). Second, we include observable characteristics as covariates in the regression analyses below. Neither of these strategies, however, guarantees that there are no idiosyncratic differences in unobserved confounders.

The sample weights that we generate to address the imbalance on some demographics were constructed to balance treatment and control samples on age, sex, and race while minimizing variance of the weights for a given level of balance. A given level of balance is defined as the tolerance of a certain magnitude of difference in the standardized means between treatment and control. The tradeoff for more precise balance is a larger variance in the weights. In our base analysis, we will use weights that tolerate a difference in the standardized means in age, sex, and race up to 0.001 standard deviations between treatment and control. In sensitivity analyses, we will loosen the tolerance in constructing these weights to 0.01 standard deviations between treatment and control. The balance weights are calibrated to match the characteristics of the overall study population. This method has been shown to perform better than a model-based approach that fits a propensity score.[9] Once we deploy the sample balance weights, all demographic characteristics are balanced. Job characteristics, including prevalence of Cigna insurance, are balanced in all analytical subsamples.

Table 9 examines individual-level characteristics collapsed to the worksite level, as well as county-level characteristics, using 2015 data from the U.S. Census Bureau. This suggests that the employee pools at the worksites roughly reflect the characteristics of the local population. The fact that treatment and control groups were located in idiosyncratically different parts of the country can be appreciated from **Figure 1**, where, for example, worksites in Ohio and Virginia were disproportionately randomly assigned to treatment and worksites in Florida were more likely to be randomly assigned to control.

VI. Statistical Analyses

Analyses are conducted at the individual employee level and at the worksite level. The treatment period in our phase 2 analysis is defined as the 36-month period spanning January 2015 through December 2017. (In comparison, the phase 1 analysis spanned January 2015 through mid-2016).

A. Intention-to-treat analysis

In the individual-level intention-to-treat (ITT) analysis, our goal is to estimate the average effect of a worker being randomized into a treatment worksite on outcomes of interest. Given that the

[9] Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association. 2015 Sep;110(511):910-922; Wang X, Zubizarreta JR. Minimal approximately balancing weights: Asymptotic properties and practical considerations. Biometrika. 2017;103(1):1-22; Hirshberg DA, Zubizarreta JR. On two approaches to weighting in causal inference. Epidemiology. 2017;28(6):812-816.

treatment is implemented at the worksite level and that there are likely fixed costs of the program, the ITT estimates (and the worksite level analysis) are informative.

We use a base specification that includes a dose-response measure of the effect of exposure to the intervention (as measured by time worked). This specification captures the overall effect of randomization into treatment as well as how the effect of randomization to a treatment site varies with additional exposure to the treatment.

$$Y_{ijt} = \beta_0 + \beta_1 TREATMENT_{jt} + \beta_2 TIME\_WORKED_{ijt} + \beta_3 TREATMENT_{jt} * TIME\_WORKED_{ijt} + \beta_4 \boldsymbol{X}_{ijt} + \varepsilon_{ijt} \quad (1)$$

In this representative estimating equation, Y_{ijt} denotes an outcome of interest, such as medical spending, for individual *i* who is employed in worksite *j* at time *t*. $TREATMENT_{jt}$ is a binary indicator of whether the individual's worksite was randomized into treatment or control. We assigned each individual to the worksite at which he or she was employed when the treatment began in January 2015. $TIME\_WORKED_{ijt}$ is the share of total available work hours in the study period that the individual actually worked; it is expressed in percentage terms. The effect of randomization to a treatment worksite is allowed to vary based on exposure to the treatment (measured via time worked during the treatment period). The treatment effect is captured by $\beta_1 + \beta_3$ * Exposure. The average effect of randomization into the treatment group can be summarized by assessing this at the average level of exposure. While we will produce estimates of the average effect of randomization into treatment, as captured in the layout of Tables 3-6, we will additionally show results for both β_1 and β_3 for each of our results in a set of parallel tables.

$\boldsymbol{X}_{ijt}$ represents a vector of covariates that may help improve precision as well as account for chance differences in characteristics between treatment and control groups, including:

- Age indicators: <20 years (omitted), 20-29, 30-39, 40-49, 50-59, and 60 and greater.
- Sex indicator: male (omitted), female
- Age-sex interactions
- Race: white (omitted), black, Hispanic, and other
- Employment characteristics (measured at baseline for the stably employed subsample):
  - Part-time (omitted), full-time
  - Salaried (omitted), hourly
  - Sales worker (omitted), non-sales worker, and other worker.

ε_{ijt} is the idiosyncratic error term. The models are weighted using the sample weights described above. Standard errors were clustered at the worksite level.

B. Local average treatment effect (LATE)

While our ITT analysis explores the effect of being randomized into a treatment worksite, a related but distinct question is: what is the effect of participating in the wellness program on the outcomes of interest? The answer to this second question will be different if not all employees in treatment worksites elect to participate. Because the choice to participate in the program is endogenous, simply including participation as a right-hand side variable may produce biased

estimates of the effect of the treatment. Therefore, we model the impact of participation on outcomes using a two-stage least squares (2SLS) specification:

$$Y_{ijt} = \gamma_0 + \gamma_1 PARTICIPATION_{ijt} + \gamma_2 \boldsymbol{X}_{ijt} + \mu_{ijt} \quad (3)$$

Where the endogenous *PARTICIPATION* variable is estimated via the first stage regression:

$$PARTICIPATION_{ijt} = \pi_0 + \pi_1 TREATMENT_{jt} + \pi_2 TIME\_WORKED_{ijt} + \pi_3 TREATMENT_{jt} * TIME\_WORKED_{ijt} + \pi_4 \boldsymbol{X}_{ijt} + \nu_{ijt} \quad (4)$$

We include the exposure interaction in the first stage, parallel to our ITT estimates. The coefficient γ_1 in equation (3) identifies the effect of participation on outcomes.

On the left-hand side of the first stage, this base specification uses the continuous definition of participation, with values from 0-12, which indicates the number of modules an individual participated in during the 36-month study period. **Table 10** shows the results of first stage estimates for this and for our alternative definitions of participation (including indicators for completion of 3 or more modules of the program or completion of at least 1 module of the program). In robustness specifications, we use these alternative definitions of participation.

If no one in the control group receives the treatment, we can interpret the 2SLS estimates of the LATE as a treatment on the treated (TOT). Empirically, we expected this to be essentially true by construction because the control worksites did not receive the wellness modules. However, because we assigned employees to worksites based on their initial locations of employment at the beginning of the study period, a small number of employees who moved from control worksites to treatment worksites during the intervention period did receive an opportunity to participate in the program. This explains the fact that the control group means in **Table 10** are nearly, but not exactly, zero.

C. Addressing endogeneity concerns

The key identifying assumption is that the only mechanism through which being employed in a treatment worksite affects outcomes is through the availability of the wellness program. There are two threats to this assumption. First, to the extent that covariates were not balanced at baseline and those covariates affect the outcomes of interest, that imbalance could bias our estimates. We control for observed covariates and apply weights to balance covariates between treatment and control. Despite these adjustments, unobserved covariates might still be imbalanced.

Second, to the extent that the wellness program affected hiring or tenure of employees, their presence in the data set may be endogenously determined. We address the issue of endogenous entry by pre-specifying a subsample of those who were already employed at the time the intervention began. We defined this more restrictive "stable employment" subsample as comprising employees who had worked continuously for 13 weeks prior to the wellness program. This comes at a cost of reducing the overall sample size. Our goal in choosing the definition for this subsample was to balance the costs and benefits, as detailed in Appendix 2.

In this stably employed subsample, however, exit from the firm's workforce is still potentially endogenous. We deal with this concern in two ways. First, we will test balance on exit between treatment worksites vs. from control worksites overall and for subgroups. Indeed, tenure and time worked are among our primary outcomes. To the extent that we observe any unbalanced exit, we will perform a bounding exercise (assuming that those who exited have outcomes working against any observed effects of the intervention).

D. Subgroup analyses

We pre-specified three subgroup analyses. We selected these subgroups based on differences in key outcomes within the control group for those with different baseline characteristics, such as differences in medical spending by age or in reported exercise by sex (**Table 11**).

We will analyze how the effect of the wellness program varies along 3 pre-specified dimensions: age (less than 40 years or 40 and above), sex (male or female), and employment status (full-time vs. part-time). In our model, these subgroup analyses are implemented using a set of two-way interaction terms and a triple interaction term. Equation 5 below shows these interaction terms in our ITT framework, using the subgroup analysis on age as the representative equation, where $age40_{it}$ is defined as 1 if the individual is age 40 years or older, and 0 otherwise.

$$\begin{aligned} Y_{ijt} = {} & \beta_0 + \beta_1 TREATMENT_{jt} + \beta_2 TIME\_WORKED_{ijt} + \beta_3 Age40_{it} \\ & + \beta_4 TREATMENT_{jt} * TIME\_WORKED_{ijt} + \beta_5 TREATMENT_{jt} * Age40_{it} \\ & + \beta_6 TIME\_WORKED_{ijt} * Age40_{it} + \beta_7 TREATMENT_{jt} * TIME\_WORKED_{ijt} * Age40_{it} \\ & + \rho \boldsymbol{X}_{ijt} + \varepsilon_{ijt} \end{aligned} \quad (5)$$

In this model, the coefficient on the triple interaction term, β_7, represents the marginal effect of an additional unit of exposure (as measured by time worked) among employees 40 years or older who were randomized into treatment. For those under age 40, the effect of the wellness program is represented by $\beta_1 + \beta_4$*Exposure. For those over age 40, the effect of the wellness program is represented by $\beta_1 + \beta_4$*Exposure + $\beta_5 + \beta_7$*Exposure, or $(\beta_1 + \beta_5) + (\beta_4 + \beta_7)$*Exposure. The average effects for each group can be derived by evaluating this at the average level of exposure. We will evaluate whether any differences are statistically significant.

E. Worksite-level analyses

We complement our analyses at the individual level with analyses at the worksite level. Worksite-level data were generated by aggregating values for individuals up to their initial worksite. This aggregation was weighted by the number of hours each employee worked during the treatment period. The estimating equation below captures the framework of the worksite-level analyses.

$$Y_{kt} = \beta_0 + \beta_1 TREATMENT_{kt} + \beta_2 \boldsymbol{X}_{kt} + \varepsilon_{kt} \quad (6)$$

In equation (6), the subscript *k* denotes a worksite. Analogous to the individual-level analyses, Y_{kt} represents an outcome for worksite *k* at time *t*. $TREATMENT_{kt}$ is a binary indicator of

randomization into treatment or control, with β_1 indicating the average worksite-level effect of being randomized into treatment. The vector of covariates $\boldsymbol{X}_{kt}$ comprises the same set of observable characteristics for individuals, aggregated to the worksite level using weighted averages (generating percentages for binary variables). The worksite-level regressions are weighted by worksite size, as measured by the total hours worked by all employees at a worksite, with standard errors adjusted for heteroscedasticity. For estimates of program effect that are significant in the individual-level analysis, we will produce parallel LATE estimates at the workplace level, analogously using treatment status as an instrument for participation.

Although the employees at each worksite could change over time, the worksites themselves were all stably represented in our data. However, at the 18-month mark, 5 worksites previously in the control group entered the treatment group. In our base specification, we will assess effects for the 20 worksites that were continuously treated throughout the 3-year period (omitting the additional 5 from the analysis altogether). As an alternative specification, we will also evaluate effects at the worksite level when considering the 5 additional worksites in the treatment group.

F. Sensitivity analyses

We conduct several types of sensitivity analyses. First, in the statistical analyses above, our base regression models use linear regression even for binary outcomes. The OLS estimator has been shown to perform better than other non-linear approaches in estimating the average population effect, particularly with large samples.[10] However, OLS also has limitations, such as its sensitivity to outliers. To test the robustness of our main results to changes in functional form, we will fit logit models for binary outcome variables.

Second, we will present LATE results with alternative definitions of participation. As defined above, these include a binary indicator of participation based on completing at least one module, and a high participation indicator defined as completing at least 3 modules.

Additionally, our phase 1 analysis of the first 18 months used a different baseline specification that did not allow for treatment effects to vary with individuals' exposure (rather, weighting individuals by exposure). It also included a different sample, as 5 additional worksites from the previous secondary control group were added to the treatment group and another 5 to the primary control group at the 18-month mark. Therefore, to facilitate comparisons of results between this phase 2 analysis to those from phase 1, we will present results using the phase 1 statistical model and set of weights on the phase 2 sample, as well as results using the phase 2 model and weights on the phase 1 sample.

Specifically, we conduct the ITT and LATE analyses using the phase 1 specifications that include just a treatment dummy. The ITT specification is:

$$Y_{ijt} = \beta_0 + \beta_1 TREATMENT_{jt} + \beta_2 \boldsymbol{X}_{ijt} + \varepsilon_{ijt} \quad (7)$$

[10] See, for example, Buntin and Zaslavsky, 2004; Manning, Basu, and Mullahy, 2005; Ellis and McGuire 2006; Jiang, Ellis, and Kuo, 2008.

The LATE specification is:

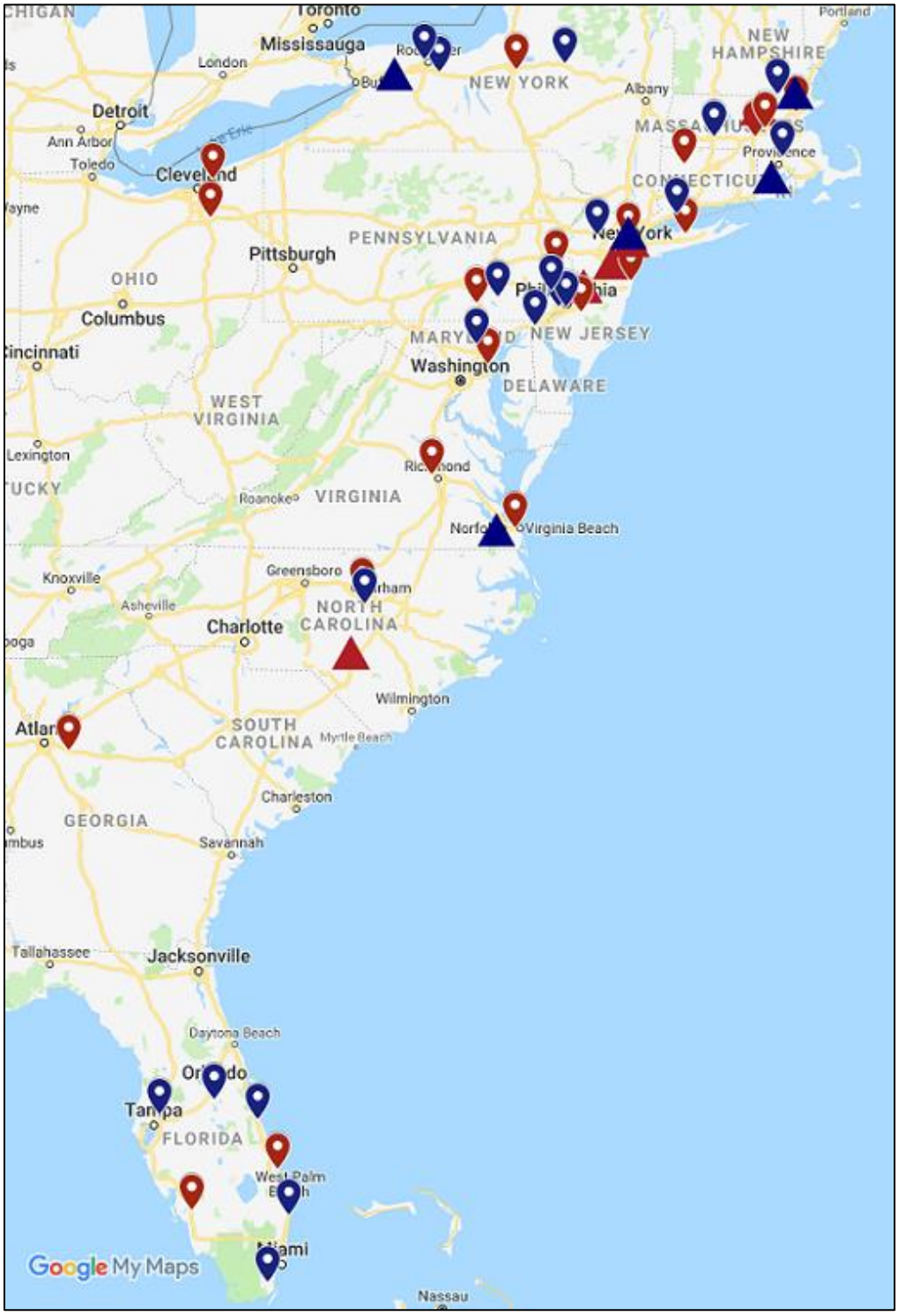
$$Y_{ijt} = \gamma_0 + \gamma_1 PARTICIPATION_{ijt} + \gamma_2 \boldsymbol{X}_{ijt} + \mu_{ijt} \quad (8)$$

With first stage regression:

$$PARTICIPATION_{ijt} = \pi_0 + \pi_1 TREATMENT_{jt} + \pi_2 \boldsymbol{X}_{ijt} + v_{ijt} \quad (9)$$

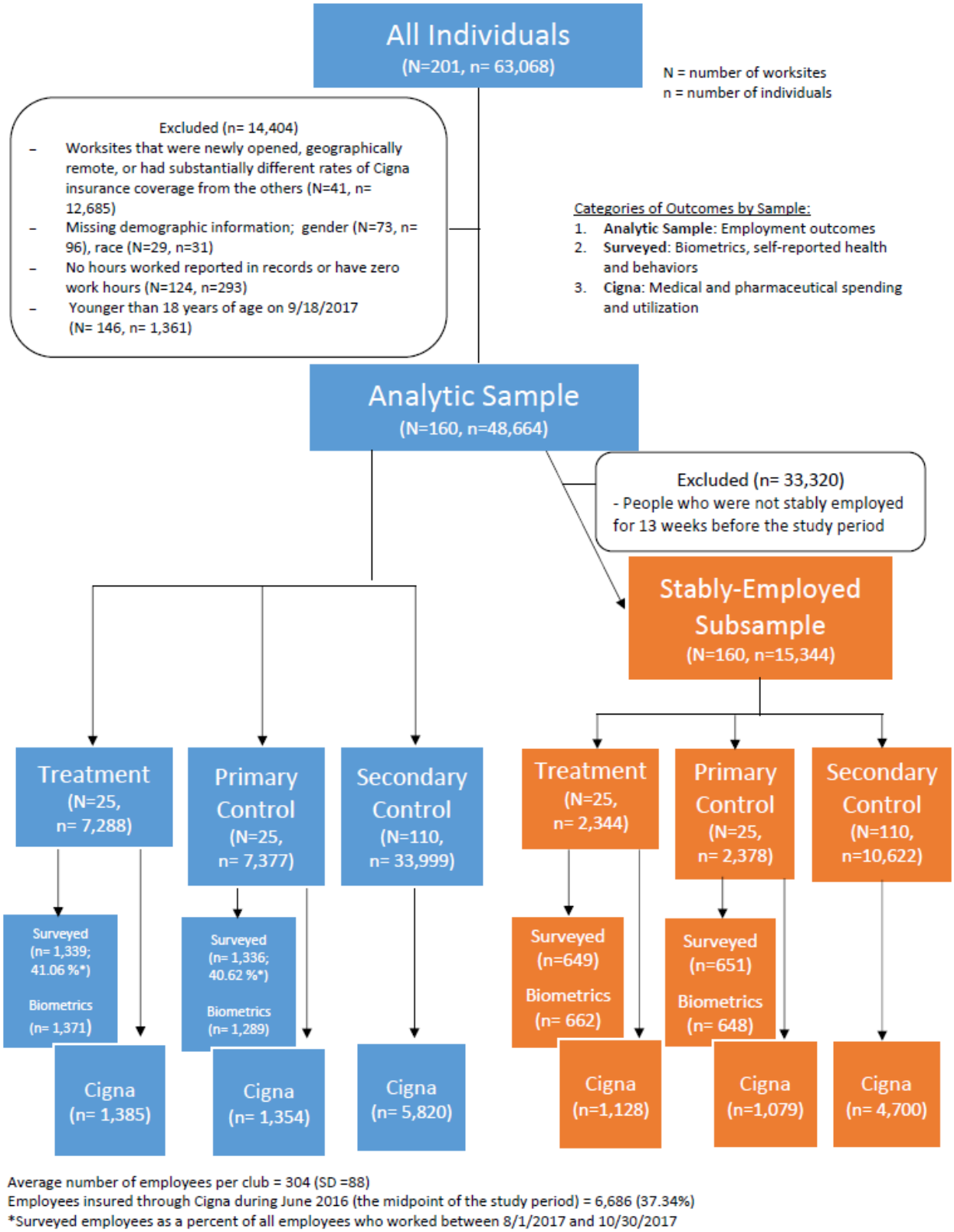
We conduct this analysis first for our full sample, and then for the subsample of treatment and control sites that match the phase 1 sample.

Figure 1: Locations of Treatment and Primary Control Worksites, Phase 2



Notes: This map shows the 25 treatment and 25 control worksites in Phase 2 of the treatment. Red markers designate treatment worksites; red triangles are the 5 additional treatment worksites that were added in Phase 2. Blue markers designate primary control worksites; blue triangles are the 5 additional primary control worksites that were added in Phase 2.

Figure 2. CONSORT Flow Diagram



Notes: This CONSORT diagram shows the flow of the trial and sample sizes through phase 2.

Table 1: Timeline of the Workplace Wellness Program

	Phase 1																																			
	Program announced ↓					Registered Dietitians begin working in the treatment worksites ↓																		Phase 2													
Year	2015												2016												2017											
Month	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Program Announcement																																				
Module 1. Take Charge of Your Health Round 1 & 2																																				
Module 2. Nutrition for a Lifetime																																				
Module 3. Club Cardio Challenge Round 1																																				
Module 4. Club Cardio Challenge Round 2																																				
Module 5. Maintain Don't Gain																																				
Module 6. Power Down the Pressure																																				
Module 7. Weight Loss Boot Camp																																				
Module 8. Movin' in May																																				
Surveys and Biometrics																																				
Module 9. Healthy and Fit																																				
Module 10. Nutrition for a Lifetime Round 1																																				
Module 11. Nutrition for a Lifetime Round 2																																				
Module 12. Step Challenge																																				
Surveys and Biometrics																																				

Notes: This table presents a graphical illustration of Phase 1 and Phase 2 of the wellness programs. The treatment began in 2015 with announcements of the wellness program worksite assignments (treatment worksites) in January followed by administration of the personal health assessments (PHAs) and in-person screenings in February. Phase 1 comprised 8 modules and concluded at the end of June 2016. After Phase 1, PHAs and in-person screenings were conducted during the summer of 2016. Afterwards, Phase 2 of the wellness program began in the fall of 2016, with PHAs and in-person screenings in fall of 2017.

Table 2. Summary of Demographic Characteristics for Employees in Control Worksites

	Employee-level			Stably Employed Subsample			Worksite-level		
	All	**PHA**	**Cigna**	**All**	**PHA**	**Cigna**	**All**	**PHA**	**Cigna**
Age (yrs)	34.5	39.8	45.3	39.6	44.3	45.3	39.4	42.4	45.2
Female (%)	47.2	56.9	45.6	47.0	62.5	45.9	46.2	57.6	45.8
Race (%)									
White	52.0	57.4	69.1	60.2	60.7	68.0	58.7	59.0	68.5
Black	25.9	19.1	14.4	19.0	14.6	15.0	20.7	18.2	14.5
Hispanic	16.7	17.8	13.4	17.3	19.6	14.0	16.4	17.7	14.1
Other race	5.4	5.6	3.1	3.5	5.1	3.1	4.1	5.1	2.9
Employment (%)									
Full-time salary	4.7	7.0	21.3	11.6	12.1	21.7	11.5	11.4	21.8
Full-time hourly	37.4	44.3	66.6	38.6	52.4	66.8	49.0	53.6	66.4
Part-time hourly	57.9	48.7	12.1	49.8	35.5	11.4	39.5	35.0	11.8
Worker Type (%)									
Sales worker	43.2	45.3	23.2	38.4	39.0	22.3	36.2	39.3	23.3
Non-sales worker	45.9	41.8	52.4	46.4	42.8	52.8	47.4	43.7	51.6
Other worker	10.9	13.0	24.4	15.3	18.2	25.0	16.5	16.9	25.0

Notes: Table lists demographic characteristics for the sample covered by Cigna weighted by months of Cigna coverage. About a third of the total sample has Cigna coverage. Age is defined as age at the mid-point of the treatment period (June 2016). This is different from the balance table where age is defined as of December, 2014 (pre-treatment). Thus the means of age in this table are larger than those in the balance table across all samples.

Table 3: Impact on Employment

	Employee-level			Stably Employed Subsample			Worksite-level		
	Mean Value in Control Group	Reduced Form (Linear)	2SLS (Linear)	Mean Value in Control Group	Reduced Form (Linear)	2SLS (Linear)	Mean Value in Control Group	Reduced Form (Linear)	2SLS (Linear)
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Absenteeism (%)	2.23			3.39			2.90		
	(2.30)			(1.97)			(0.35)		
	41,376			13,000			135		
Performance Review (% ≤ 3)	46.59			63.64			65.94		
	(49.88)			(48.11)			(12.41)		
	27,150			12,916			135		
Tenure§	416.69			786.57			76.08		
	(399.87)			(376.96)			(4.35)		
	41,376			13,000			135		

Notes: Table reports the coefficient on TREATMENT from estimating equation (1) by OLS (column 2), and the coefficient on PARTICIPATION from estimating equation (2) by IV (column 3). Standard errors are listed in parentheses. Column 1 reports the mean of each employment outcome in the control group for each sample (with standard deviation in parentheses, followed by the number of observations for each outcome). All regressions will include demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, Cigna coverage status, full-time status, paid hourly status, and job category) and cluster standard errors at the worksite (for employee-level regressions). Employee-level and stably employed subsample control means are weighted by a weight that balances treatment and control on demographic characteristics. Worksite-level control means are not traditionally weighted: individual-level records are collapsed at the worksite-level and weighted by the total number of hours worked at each worksite. For worksite level results on tenure, the outcome was defined as the percent of total days at the worksite during the study period that was worked by employees at the worksite.

§ Tenure was defined as the number of days worked during the treatment period for the employee-level mean and stably employed subsample mean; it was defined as the percent of the entire treatment period worked by employees in the worksite-level mean.

Table 4: Impact on Medical & Pharmaceutical Spending and Utilization

	Employee-level			Stably Employed Subsample			Worksite-level (per 2000 work hours)		
	Mean in Control Group	Reduced Form (Linear)	2SLS (Linear)	Mean in Control Group	Reduced Form (Linear)	2SLS (Linear)	Mean in Control Group	Reduced Form (Linear)	2SLS (Linear)
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Medical Spending									
Total Spending	4800.57			4721.85			5044.74		
	(18559.82)			(19056.92)			(2994.54)		
Medical Utilization									
Number of Office visits	3.90			3.95			4.21		
	(4.47)			(4.33)			(0.93)		
Number of Hospitalizations	0.09			0.08			0.08		
	(0.40)			(0.36)			(0.05)		
Number of ER Visits	0.28			0.26			0.27		
	(0.75)			(0.61)			(0.11)		
Pharmaceutical Spending									
Total Spending	1237.60			1227.32			1317.72		
	(6921.32)			(7115.67)			(1145.17)		
Pharmaceutical Utilization									
Number of Distinct Medications	5.49			5.91			6.00		
	(6.44)			(6.54)			(1.30)		
Total Medication Months	11.20			11.41			12.39		
	(19.31)			(19.37)			(4.38)		
N	7,174			5,779			135		

Notes: Table reports the coefficient on TREATMENT from estimating equation (1) by OLS (column 2), and the coefficient on PARTICIPATION from estimating equation (2) by IV (column 3). Standard errors are listed in parentheses. Column 1 reports the mean of each outcome in the control group for each sample (with standard deviation in parentheses). All regressions include demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, full-time status, paid hourly status, and job category) and cluster standard errors at the worksite (for employee-level regressions). Employee-level and stably employed subsample control means are weighted by a weight that balances treatment and control on demographic characteristics. Worksite-level control means are not traditionally weighted: individual-level records are collapsed at the worksite-level and weighted by the total number of hours worked at each worksite.

Table 5: Impact on Biometrics

	Employee-level			Stably Employed Subsample		
	Mean in Control Group (1)	Reduced Form (Linear) (2)	2SLS (Linear) (3)	Mean in Control Group (1)	Reduced Form (Linear) (2)	2SLS (Linear) (3)
Cholesterol (mg/dl)	179.87			184.88		
	(38.71)			(37.85)		
	1,392			739		
HDL (mg/dl)	49.47			50.51		
	(14.58)			(14.87)		
	1,366			720		
Glucose (mg/dl)	104.76			105.79		
	(35.71)			(38.17)		
	1390			736		
Systolic BP (mmHg)	124.84			127.22		
	(18.55)			(18.73)		
	1,400			738		
BMI	29.79			30.04		
	(7.27)			(6.71)		
	1,403			742		

Notes: Table reports the coefficient on TREATMENT from estimating equation (1) by OLS (column 2), and the coefficient on PARTICIPATION from estimating equation (2) by IV (column 3). Standard errors are listed in parentheses. Column 1 reports the mean of each biometric outcome in the control group for each sample (with standard deviation in parentheses, followed by the number of observations for each outcome). All regressions include demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, Cigna coverage status, full-time status, paid hourly status, and job category) and cluster standard errors at the worksite (for employee-level regressions). Control means are weighted by a weight that balances treatment and control on demographic characteristics.

Table 6: Impact on Self-Reported PHA Responses

	Employee-level			Stably Employed Subsample		
	Mean in Control Group	Reduced Form (Linear)	2SLS (Linear)	Mean in Control Group	Reduced Form (Linear)	2SLS (Linear)
	(1)	(2)	(3)	(1)	(2)	(3)
Domain 1: Health Behaviors						
Screenings and Exams						
Percent of recommended tests received	58.00			64.03		
	(32.51)			(30.80)		
	1,454			747		
Physical Activity						
Regular exercise (%)	53.79			54.56		
	(49.87)			(49.83)		
	1,440			737		
Number of hours sitting per day	3.73			3.71		
	(1.78)			(1.77)		
	1,444			741		
Nutrition						
Number of non-zero calorie drinks per day	1.36			1.21		
	(1.71)			(1.60)		
	1,443			742		
Read the Nutrition Facts panel (%)	58.91			61.66		
	(49.22)			(48.65)		
	1,443			739		
Consume at least 2 cups of fruit and 2.5 cups of vegetables per day (%)	54.13			57.60		
	(49.85)			(49.45)		
	1,441			738		
Weight Management						
Actively managing weight (%)	61.52			61.40		
	(48.67)			(48.72)		
	1,417			726		

Tobacco Use		
Smoker (%)	14.67	13.00
	(35.40)	(33.65)
	1,444	741
Domain 2: Health and Well-Being		
SF-8 score – physical summary score	51.56	51.65
	(6.78)	(6.76)
	1,427	735
SF-8 score – mental summary score	51.87	52.26
	(8.47)	(8.07)
	1,427	735
Unmanaged stress (%)	30.68	31.77
	(46.13)	(46.59)
	1,441	738
Unmanaged depression (%)	30.58	29.42
	(46.09)	(45.60)
	1,444	740
Stress at work (%)	39.69	44.35
	(48.94)	(49.71)
	1,447	743

Notes: Table reports the coefficient on TREATMENT from estimating equation (1) by OLS (column 2), and the coefficient on PARTICIPATION from estimating equation (2) by IV (column 3). Standard errors are listed in parentheses. Column 1 reports the mean of each self-reported health outcome in the control group for each sample (with standard deviation in parentheses, followed by the number of observations for each outcome). All regressions include demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, Cigna coverage status, full-time status, paid hourly status, and job category) and cluster standard errors at the worksite (for employee-level regressions). Control means are weighted by a weight that balances treatment and control on demographic characteristics.

Table 7: Participation Rates

Panel A. Average Participation Rates by Module

	Phase 1								Phase 2			
Module	Take Charge of Your Health	Nutrition for a Lifetime	Club Cardio Challenge Round 1	Club Cardio Challenge Round 2	Maintain Don't Gain	Power Down the Pressure	Weight Loss Boot Camp	Movin' in May	Healthy and Fit	Nutrition for a Lifetime Round 1	Nutrition for a Lifetime Round 2	Step Challenge
Participation rate (%)	12.4	25.7	37.8	28.6	31.6	33.4	28.7	28.5	26.3	32.6	27.2	29.7

Notes: The average participation rate for each module was calculated as the percentage of individuals who completed a module out of all who were eligible to complete a module during the time that the module was running. Participation is equivalent to completion of a module, with an incentive of a gift card for completion of the module. Descriptions of each module and their criteria for completion are provided in Appendix 1. Employees could only participate in the Take Charge of Your Health module once, though it was run twice. Club Cardio Challenge had two rounds and completion of either round 1 or round 2 earned a gift card; completion of both rounds did not earn an additional gift card, but rather an entry into a raffle for a Fitbit, unless the employee had Cigna health insurance, in which case they could complete both rounds of Club Cardio Challenge for an additional fitness reimbursement. Nutrition for a Lifetime occurred in two rounds; individuals who completed round 1 earned an ActivBAND or gift card. Completion of round 1 and round 2 earned an additional gift card or weight management reimbursement for Cigna members. Values were weighted by the number of days an individual was working during a given module's timeframe.

Panel B. Intensity of Participation by Definition

Definition of Participation	Employees at Treatment Worksites (N = 7288)	Participants at Treatment Worksites (N = 2071)
Completed any module (%)	28.4	100.0
Modules completed (#)	1.17	4.11
3 or more modules (%)	16.7	58.8

Notes: These summary statistics show the intensity of participation in the treatment group according to the 3 definitions of participation. In contrast to Panel A above, the sample (denominator) for these calculations was all employees at treatment worksites (first column) and all employees at treatment worksites who completed any module (second column) throughout the study period. Because of this difference in the denominators, participation rates for a given module in Panel A could exceed the average participation study-wide in Panel B.

Table 8: Balance Between Treatment and Control—Employee Level

Panel A.1: All Employees – Unweighted

	(1) Treatment (n=7288)	(2) Primary Control (n=7377)	(3) Primary + Secondary Control (n=41376)	(4) (1) vs (2) P value	(5) (1) vs (3) P value
Demographics					
Age (yrs)	33.5	32.8	33.0	0.333	0.358
Female (%)	46.9	46.8	45.6	0.942	0.444
Race (%)				<0.001	<0.001
Black	21.1	25.2	26.7		
White	57.7	51.6	48.8		
Hispanic	14.2	17.4	17.7		
Other	7.0	5.8	6.9		

Panel A.2: All Employees – Weighted

	(1) Treatment (n=7288)	(2) Primary Control (n=7377)	(3) Primary + Secondary Control (n=41376)	(4) (1) vs (2) P value	(5) (1) vs (3) P value
Demographics					
Age (yrs)	33.0	32.9	33.1	0.867	0.934
Female (%)	45.9	46.9	45.8	0.614	0.964
Race (%)				0.999	>0.999
Black	25.8	24.4	25.9		
White	50.0	52.9	50.1		
Hispanic	17.3	16.9	17.2		
Other	7.0	5.8	6.8		

Notes: Demographic characteristics are plausibly unaffected by the treatment. Data are from the Team Member database supplied by BJ's and based on the first entry for an individual during the treatment period. Age is defined as of December, 2014 (pre-treatment). Column 1 reports the means for employees in the treatment group while columns 2 and 3 report the means for the primary control employees and all control employees (primary and secondary), respectively. Treatment status is defined by the first worksite an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between the employees at treatment worksites and the employees at primary control worksites and column 5 reports the p-value for the comparison between the employees at treatment worksites and all employees at control worksites. In Table A.1, regressions are unweighted and standard errors are clustered by worksite. In Table A.2, all regressions are weighted by a weight that balances treatment and control on demographic characteristics and standard errors are clustered by worksite.

Panel B: Personal Health Assessment (PHA) Subsample

	(1) Treatment (n=1339)	(2) Primary Control (n=1336)	(3) (1) vs (2) P value
Demographics			
Age (yrs)	36.8	37.9	0.251
Female (%)	56.0	57.3	0.621
Race (%)			0.999
Black	21.4	18.6	
White	55.4	57.5	
Hispanic	16.7	18.2	
Other	6.5	5.7	

Notes: Employees are included if they answered at least 1 question on the PHA. Demographic characteristics are plausibly unaffected by the treatment. Demographics are taken from the Team Member database supplied by BJ's and based on the first entry for an individual during the treatment period. Age is defined as of December, 2014 (pre-treatment). Column 1 reports the means for employees in the treatment group while column 2 reports the means for the primary control employees. Treatment status is defined by the first worksite an employee appears in during the treatment period. Column 3 reports the p-value for the comparison between the employees at treatment worksites and the employees at primary control worksites. All regressions are weighted by a weight that balances treatment and control on demographic characteristics and cluster standard errors by worksite.

Panel C: Cigna Insured Subsample

	(1) Treatment (n=1385)	(2) Primary Control (n=1354)	(3) Primary + Secondary Control (n=7174)	(4) (1) vs (2) P value	(5) (1) vs (3) P value
Demographics					
Age (yrs)	42.7	42.1	42.6	0.321	0.836
Female (%)	44.4	43.8	44.6	0.768	0.949
Race (%)				0.996	>0.999
Black	16.9	14.9	17.7		
White	62.7	66.6	61.2		
Hispanic	14.7	14.9	15.9		
Other	5.6	3.7	5.2		

Notes: Employees are included if they had at least 1 month of Cigna health insurance coverage. Demographic characteristics are plausibly unaffected by the treatment. Demographics are taken from the Team Member database supplied by BJ's and based on the first entry for an individual during the treatment period. Age is defined as of December, 2014 (pre-treatment). Column 1 reports the means for employees in the treatment group while columns 2 and 3 report the means for the primary control employees and all control employees (primary and secondary), respectively. Treatment status is defined by the first worksite an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between the employees at treatment worksites and the employees at primary control worksites and column 5 reports the p-value for the comparison between the employees at treatment worksites and all employees at control worksites. All regressions are weighted by a weight that balances treatment and control on demographic characteristics and cluster standard errors by worksite.

Panel D: Stably Employed Subsample

	(1)	(2)	(3)	(4)	(5)
	Treatment (n=2344)	Primary Control (n=2378)	Primary + Secondary Control (n=13000)	(1) vs (2) P value	(1) vs (3) P value
Demographics					
Age (yrs)	39.0	38.1	38.7	0.193	0.623
Female (%)	47.4	47.7	46.9	0.905	0.747
Race (%)				0.998	>0.999
Black	20.5	18.1	20.3		
White	57.9	60.5	56.2		
Hispanic	16.2	17.4	17.7		
Other	5.4	4.0	5.8		
Employment					
Worker type (%)				>0.999	>0.999
FT salary	11.7	11.8	12.0		
FT hourly	36.2	37.3	38.2		
PT hourly	52.1	50.9	49.9		
Annual rate ($)					
FT salary	49,304	47,910	48,301	0.178	0.218
FT hourly	24,933	24,237	24,773	0.285	0.764
PT hourly	10,142	9,944	10,027	0.220	0.430
Standard Hours Per Week				>0.999	>0.999
FT salary	40	40	40		
FT hourly	35.6	35.6	35.8		
PT hourly	20	20	20		
Job Category				>0.999	>0.999
Sales workers	38.1	38.3	37.8		
Laborers/Helpers	18.9	17.8	18.4		
Operatives	15.7	15.1	15.4		
Service workers	12.3	13.1	12.5		
First/Mid level officials	9.2	9.2	9.4		
Admin Support	3.8	3.9	4.2		
Other	2.1	2.6	2.2		
Health Insurance					
Ever Enrolled in Cigna (2014)	0.4	0.4	0.4	0.964	0.244
Months in Cigna	11.4	11.3	11.4	0.834	0.607
Total medical spending	5,476	3,634	4,659	0.135	0.493

Notes: Employees are included if they were part of the stably employed subsample. Health insurance variables are pre-randomization characteristics. Demographics and employment characteristics are taken from the Team Member database supplied by BJ's and based on the first entry for an individual during the treatment period. Age is defined as of December, 2014 (pre-treatment) and employees who were under 18 are excluded. Column 1 reports the means for employees in the treatment group while columns 2 and 3 report the means for the

primary control employees and all control employees (primary and secondary), respectively. Treatment status is defined by the first worksite an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between the employees at treatment worksites and the employees at primary control worksites and column 5 reports the p-value for the comparison between the employees at treatment worksites and all employees at control worksites. All regressions are weighted by a weight that balances treatment and control on demographic characteristics and cluster standard errors by worksite.

Table 9: Balance Between Treatment and Control—Worksite Level

	(1) Treatment (n=25)	(2) Primary Control (n=25)	(3) Primary + Secondary Control (n=135)	(1) vs (2) P value	(1) vs (3) P value
Employee Demographics					
Age (yrs)	38.5	37.6	38.1	0.165	0.458
Female (%)	48.3	46.6	45.7	0.348	0.065
Race (%)					
Black	17.6	19.4	20.4	0.723	0.463
White	65.6	60.0	57.2	0.427	0.098
Hispanic	11.4	16.1	16.6	0.321	0.150
Other	5.5	4.4	5.8	0.295	0.665
County-Level Demographics					
Age (yrs)	40.0	39.8	39.7	0.798	0.725
Female (%)	51.1	51.4	51.3	0.183	0.228
Race (%)					
Black	12.4	12.8	13.3	0.909	0.725
White	75.9	76.9	74.7	0.773	0.687
Hispanic	11.0	14.3	13.7	0.341	0.170
Other	11.7	10.2	12.0	0.396	0.833

Notes: Demographic characteristics are plausibly unaffected by the treatment. Employee demographics are taken from the Team Member database supplied by BJ's and based on the first entry for an individual during the treatment period. Age is defined as of December, 2014 (pre-treatment). ACS demographics are taken from the 2015 American Community Survey (ACS) Population Estimates for the county each worksite is located in. Worksite-level analyses are obtained by first calculating a weighted average for each worksite (weighted by an employee's hours worked during the treatment period). Column 1 reports the means for employees in the treatment group while columns 2 and 3 report the means for the primary control employees and all control employees (primary and secondary), respectively. Treatment status is defined by the first worksite an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between the employees at treatment worksites and the employees at primary control worksites and column 5 reports the p-value for the comparison between the employees at treatment worksites and all employees at control worksites.

Table 10: First Stage Estimates

Panel A: Employee Level—All

	All		Completed PHA		Cigna Enrolled	
	(1) Control Mean	(2) Estimated First Stage	(3) Control Mean	(4) Estimated First Stage	(5) Control Means	(6) Estimated First Stage
Completed any module (%)	0.21	27.53	2.29	63.06	0.91	57.31
		(1.25)		(1.98)		(2.81)
		[0.00]		[0.00]		[0.00]
Modules completed (#)	0.01	1.12	0.08	3.45	0.03	3.13
		(0.09)		(0.24)		(0.26)
		[0.00]		[0.00]		[0.00]
3 or more modules (%)	0.12	15.95	1.60	46.74	0.53	43.29
		(1.08)		(2.31)		(2.74)
		[0.00]		[0.00]		[0.00]
Average total incentive payment ($)	0.26	37.97	3.20	126.54	1.37	148.83
		(3.20)		(9.40)		(11.93)
		[0.00]		[0.00]		[0.00]
N	41,376	48,664	1,454	2,793	7,174	8,559

Note: Control means and first stage estimates of the impact of TREATMENT on alternate definitions of PARTICIPATION. All regressions include demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, full-time status, paid hourly status, and job category) and cluster standard errors at the worksite level. All samples other than the sample with Cigna coverage also include a control for whether or not the employee ever had Cigna coverage during the treatment period. Employee-level regressions are weighted by a weight that balances treatment and control on demographic characteristics. Standard errors shown in parenthesis and p-values in brackets.

Panel B: Employee-level—Stably Employed Sub-sample

	All		Surveyed		Cigna	
	(7) Control Mean	(8) Estimated First Stage	(9) Control Mean	(10) Estimated First Stage	(11) Control Means	(12) Estimated First Stage
Completed any module (%)	0.57	49.75	3.90	79.22	0.97	59.54
		(2.82)		(2.74)		(3.24)
		[0.00]		[0.00]		[0.00]
Modules completed (#)	0.02	2.56	0.15	5.41	0.04	3.41
		(0.21)		(0.39)		(0.29)
		[0.00]		[0.00]		[0.00]
3 or more modules (%)	0.35	35.04	2.96	67.78	0.61	45.79
		(2.47)		(3.39)		(3.15)
		[0.00]		[0.00]		[0.00]
Average total incentive payment ($)	0.75	97.62	5.92	217.91	1.54	163.36
		(8.45)		(15.74)		(13.80)
		[0.00]		[0.00]		[0.00]
N	13,000	15,344	747	1,396	5,779	6,907

Note: Control means and first stage estimates of the impact of TREATMENT on alternate definitions of PARTICIPATION. All regressions include demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, full-time status, paid hourly status, and job category) and cluster standard errors at the worksite level. All samples other than the sample with Cigna coverage also include a control for whether or not the employee ever had Cigna coverage during the treatment period. Employee-level regressions are weighted by a weight that balances treatment and control on demographic characteristics. Standard errors shown in parentheses and p-values in brackets.

Panel C: Worksite level

	All		Surveyed		Cigna	
	(1) Control Mean	(2) Estimated First Stage	(3) Control Mean	(4) Estimated First Stage	(5) Control Means	(6) Estimated First Stage
Completed any module (%)	0.22	28.19	3.07	62.25	0.91	57.29
		(1.27)		(1.89)		(2.83)
		[0.00]		[0.00]		[0.00]
Modules completed (#)	0.01	1.17	0.11	3.49	0.03	3.18
		(0.09)		(0.22)		(0.27)
		[0.00]		[0.00]		[0.00]
3 or more modules (%)	0.12	16.69	2.27	46.65	0.54	43.68
		(1.12)		(2.16)		(2.76)
		[0.00]		[0.00]		[0.00]
Average total incentive payment ($)	0.27	40.31	4.40	129.92	1.37	150.85
		(3.47)		(8.84)		(12.47)
		[0.00]		[0.00]		[0.00]
N	135	160	85	110	135	160

Note: Control means and first stage estimates of the impact of TREATMENT on alternate definitions of PARTICIPATION. All regressions include demographic and employment controls (age, sex, age-sex interactions, race/ethnicity, full-time status, paid hourly status, and job category) and adjust standard errors for heteroscedasticity. All samples other than the sample with Cigna coverage also include a control for the percent of workers per worksites every having Cigna coverage during the treatment period. Worksite level regressions weighted by worksite size (total hours worked in the worksite). Standard errors shown in parentheses and p-values in brackets.

Table 11: Heterogeneity

	Absenteeism (%)		Total medical spending ($)		Regular exercise (%)		Considering losing weight (%)	
	N	Control Mean	N	Control Mean	N	Control Mean	N	Control Mean
Gender								
Female	18886	2.40	3182	5772.21	813	49.75	753	71.21
Male	22490	2.08	3992	4019.28	627	59.04	583	49.30
Age								
Below 40	27495	2.07	2964	2627.73	744	54.38	685	60.48
40 and above	13881	2.54	4210	6323.57	696	53.16	651	62.91
Employment type								
Full-time	16901	2.47	5976	5162.78	765	53.59	712	62.41
Part-time	24475	2.06	1198	2993.43	675	54.01	624	60.82

	SF-8 physical summary score		SF-8 mental summary score		Non-zero calorie Drinks (No.)		BMI		Systolic BP (mmHg)	
	N	Control Mean	N	Control Mean	N	Control Mean	N	Control Mean	N	Control Mean
Gender										
Female	810	50.87	810	51.11	815	1.24	789	30.36	785	121.62
Male	617	52.47	617	52.87	628	1.53	614	29.04	615	128.97
Age										
Below 40	733	52.58	733	50.73	743	1.69	718	29.23	717	118.43
40 and above	694	50.49	694	53.06	700	1.02	685	30.37	683	131.52
Employment type										
Full-time	761	51.25	761	52.36	769	1.20	767	30.13	764	126.42
Part-time	666	51.91	666	51.31	674	1.55	636	29.37	636	122.93

Note: Table reports the control group sample size in column 1 and control mean in column 2 for each outcome. Control means are weighted by a weight that balances treatment and control on demographic characteristics.

The Impact of Workplace Wellness on Health Care Spending, Health, and Employment Outcomes: A Randomized Controlled Trial

Appendices

Appendix 1: Description of Phase 2 Modules and Incentives

Appendix 2: Determination of the Stably Employed Subsample

Appendix 1: Description of Phase 1 Modules and Incentives

Module 1

Take Charge of Your Health Rounds 1 and 2 (2/23/2015-3/27/2015, 4/13/2015-5/15/2015)

- Summary: These two five-week programs were presented as a series of webinars with corresponding PowerPoints designed to help employees who participate take their health care into their own hands. Topics covered included:
    - how to choose a health plan and primary care physician,
    - what to expect from a routine visit,
    - routine tests and screenings and recommended frequencies,
    - how to get the most from a doctor's visit,
    - choosing generic medications over the corresponding brand name,
    - staying healthy by eating well, staying active, sleeping enough, and managing stress, and
    - primary care vs urgent care vs the emergency room and when to use each.
- Incentive: Employees who completed the webinars and returned the verification form received a $25 BJ's gift card.

Module 2

Nutrition for a Lifetime (6/1/2015-7/10/2015)

- Summary: This six-week program was presented as a series of webinars with corresponding PowerPoints designed to help employees who participate achieve and maintain a healthy weight for life through the four pillars of health: nutrition, exercise, stress management, and sleep. Topics covered included:
    - the negatives consequences of chronic stress and poor sleep habits and techniques to manage stress and improve sleep,
    - good nutrition, including an overview on the different food groups and the amounts of each recommended per day,
    - reasons for making exercise a priority and how to get the most out of a workout,
    - foods to limit and foods to increase in a diet,
    - appropriate portion sizing, especially for weight loss and weight maintenance, and
    - choosing the right fats and the importance of fiber.
- Incentive: Employees with Cigna coverage received a $150 Weight Management Reimbursement and all other employees received at $50 gift card if they completed 5 out of the 6 webinars and returned the verification form.

Modules 3 and 4

Club Cardio Challenge Rounds 1 and 2 (8/10/2015-9/25/2015, 9/26/2015-11/16/2015)

- Summary: These two seven-week programs were exercise-based. Employees were supposed to complete 20 minutes or more of cardiovascular exercise at least 3 days per week and track their activity in an exercise log.
- Incentive: Employees who completed 6 of the 7 weeks in either round 1 or round 2 earned a $25 BJs gift card. Employees who completed 12 out of 14 weeks over both rounds were eligible to enter a raffle at their worksite for a Fitbit. Employees with Cigna coverage who completed 12 out of 14 weeks received a $150 fitness reimbursement from Cigna on top of the raffle entry and gift card. Worksites were also in competition with the

top worksite based on % participation and the top worksite with the highest average weekly minutes of exercise reported each receiving a trophy to display in the worksite, winner buttons for employee lanyards, and bragging rights.

Module 5

Maintain Don't Gain (11/23/2015-12/20/2015)

- Summary: This four-week challenge helped employees track their weight each week and offered tips on how to add physical activity to a daily routine and substitutions for options with fewer calories when dining out.
- Incentive: Employees who completed at least 3 out of the 4 weeks of weight tracking and returned the verification form received a $25 BJ's gift card.

Module 6

Power Down the Pressure (1/18/2016-2/19/2016)

- Summary: This four-week program encouraged employees to learn effective methods for managing stress by asking them to complete at least one activity from a list of options for the week for at least 3 days of the week. Week 1 was called "Unplug" and included activities such as refraining from watching TV for a day or having an electronic-free meal with family or friends. Week 2 was titled "Boost Your Mood" and included activities like doing a random act of kindness, getting 8 hours of sleep, or spending time with a friend. Week 3 was "Exercise" and asked employees to take a new exercise class or do a 30-minute workout/activity outdoors. The final week was called "Relaxation and Meditation" and encouraged employees to keep a stress journal, color, and meditate.
- Incentive: Employees who completed all four weeks of the program by completing at least 3 days of stress management activities a week and returned the verification form received a $25 BJ's gift card.

Module 7

Weight Loss Boot Camp (3/14/2016-4/8/2016)

- Summary: This four-week program aimed to teach employees methods for losing weight. For each of the four weeks, employees had to complete four activities (eating five or more servings of fruits and vegetables, exercising for at least 30 minutes, avoiding sweetened beverages, and weighing themselves weekly) a minimum number of days each week, from two days the first week up to five days the final week.
- Incentive: Employees who completed all four weeks and return the verification form received a $25 BJ's gift card.

Module 8

Movin' in May (5/1/2016-5/31/2016)

- Summary: This four-week program encouraged employees to exercise for at least 30 minutes 3 days per week and track their exercise.
- Incentive: Employees who completed all four weeks of the challenge and returned the verification form were entered to win one of two $250 visa gift cards at their worksite.

Module 9

Healthy and Fit (11/14/2016-12/11/2016)

- Summary: This four-week program encouraged employees to stay healthy and fit through the holiday season. Weekly activities included; weight check-in, 10-minute wellness coaching sessions, de-stressing (yoga, meditation, planning for the week ahead…etc.), "me" time, and a soda free day. Activities that as practiced multiple times a week included; exercising for at least 30 minutes, 15 minutes walking breaks, packing a healthy lunch, and practicing gratitude and kindness.
- Incentive: Employees who earned at least 800 points during this program received a Fitrax ActivBand.

Modules 10 and 11

Nutrition for a Lifetime Rounds 1 and 2 (1/23/2017- 4/02/2017)

- Summary: NFL was a 10-week program which required completing at least 4 weeks of the program to receive a prize. Employees who completed 8 weeks of the program (4 + 4 weeks as two NFL rounds) received additional prizes. Each week a new module was available and employees were able to start the program until the 6th week. Weekly topics included:
    - Initial counseling session to set the goals
    - Healthy diet and nutrition webinars
    - Cardio Challenge (logging miles of exercise every day for at least week)
    - Fiber, fat facts, and good night's sleep webinars
    - Stress management
    - Healthy dining out
    - At home exercises
    - Final counseling session to continue the goals
- Incentive: Employees who completed 4 weeks out of 10 weeks earned a $25 BJs gift card or a Fitrax ActivBand. Employees who completed 8 weeks (conditional on having both counseling sessions) receive an additional $25 BJ's Gift Card. Employees with Cigna coverage who completed 8 weeks of the program qualified for the $150 weight management reimbursement benefit.

Module 12

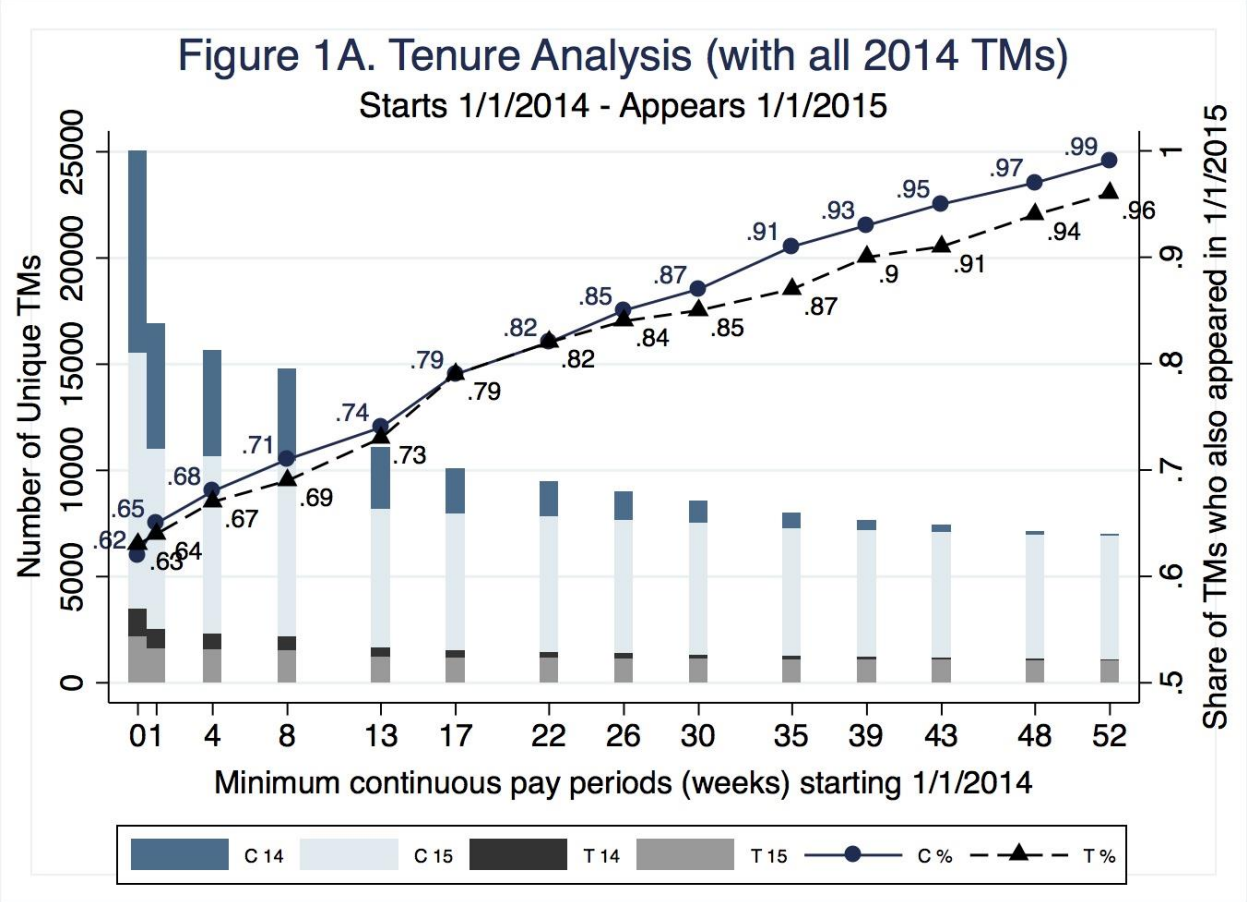
Step Challenge (5/8/2017-6/30/2017 or 7/30/2017 if Cigna eligible)

- Summary: The step challenge was an 8-week team-based program. Each team consisted of 4 team members and each member tracked their steps using an ActivBAND, Fitbit or a smart phone app. Each employee was required track at least 150,000 steps during the module to qualify as a team member.
- Incentive: Employees received both team-based and individual incentives based on their performance. At each worksite, 1st place team received a $100 gift card, 2nd place team received a $50 gift card, the 3rd place team received a $25 gift card and top 10 individuals received a $25 gift card which may be additional to the team based incentive. Cigna-Eligible employees needed to complete an additional 4 weeks (12 weeks in total) and at least 225,000 steps in total to qualify for a $150 fitness reimbursement from Cigna.

Appendix 2. Determination of the Stably Employed Subsample

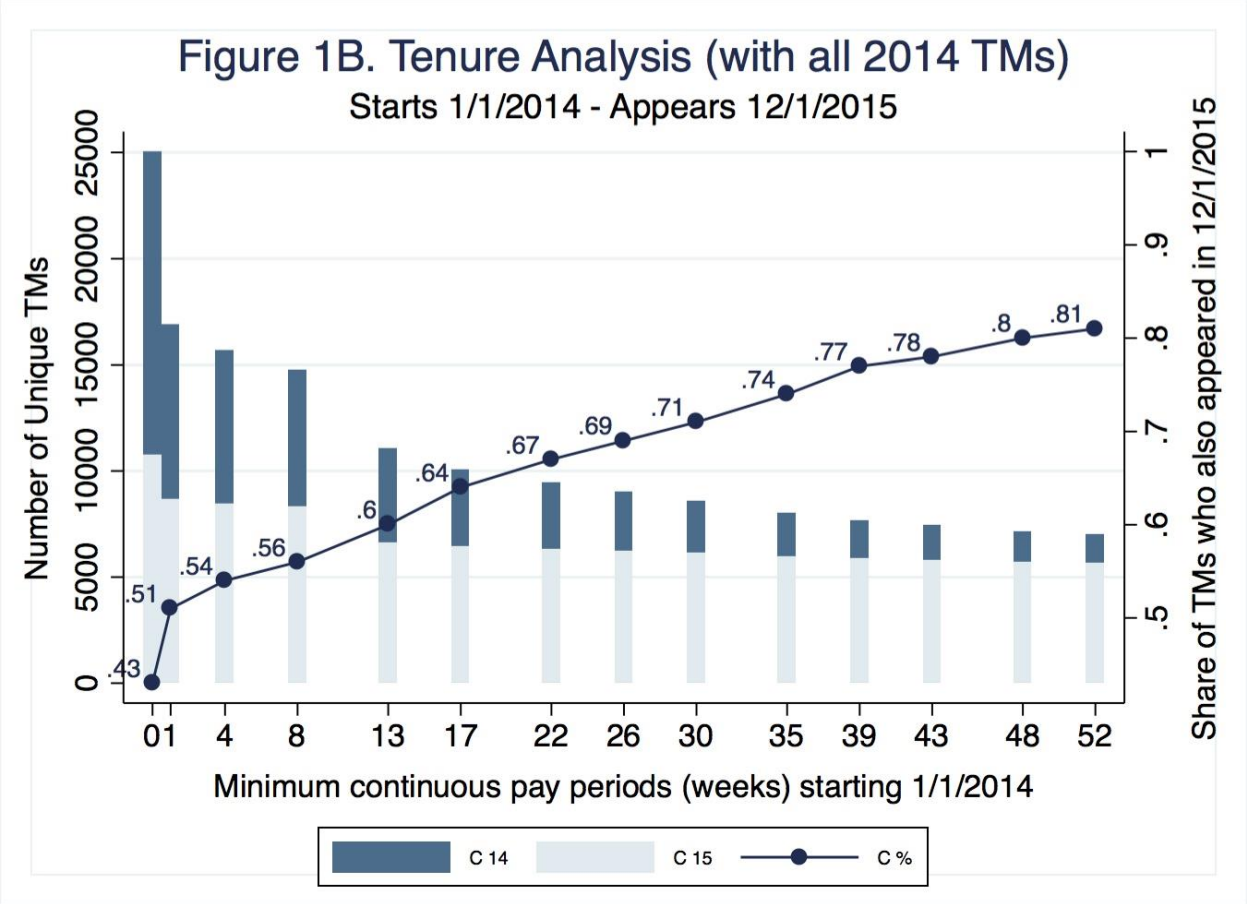
We conducted an analysis of the duration of employment (tenure), which informed our definition of the pre-specified stably employed subsample. This analysis of tenure is described below.

Figure 1A looks at a number of scenarios where we take samples of treatment (T) and control (C) workers who were employed for varying numbers of consecutive weeks starting on 1/1/2014. In each scenario, we follow the samples of workers until they reach 1/1/2015 and look at how many of them are still employed. To be precise, for each restriction criterion of the number of consecutive weeks worked starting 1/1/2014 (X axis), the height of the dark blue bar (C 14) depicts the total number of control workers in the sample and the height of the light blue bar (C 15) depicts the total number of control workers who were still working on 1/1/2015. Analogously, the height of the black bar (T 14) represents treatment workers who started in the sample on 1/1/2014 and the height of the gray bar (T 15) represents treatment workers who were still working on 1/1/2015. Of note, the bars are overlapping for each X (i.e. they are not stacked; rather they all originate at 0). The solid and dotted lines merely reflect the percentages of C and T employees, respectively, who were still working at BJ's on 1/1/2015 (i.e. light blue bar divided by dark blue bar, gray bar divided by black bar).



Thus, for example, the interpretation of the bar when X=17 is as follows. There were about 10,000 employees in control worksites who were employed at BJ's on 1/1/2014 and who worked through the first 17 weeks of 2014 (dark blue bar). Among these employees, about 79% (or about 7,900) were still working on 1/1/2015 (light blue bar). The same retention of 79% was found among employees in treatment worksites—calculated using the gray (numerator) and black (denominator) bars. The bar originating at X=0 represents the case of no sample restrictions (i.e. all employees in the data).

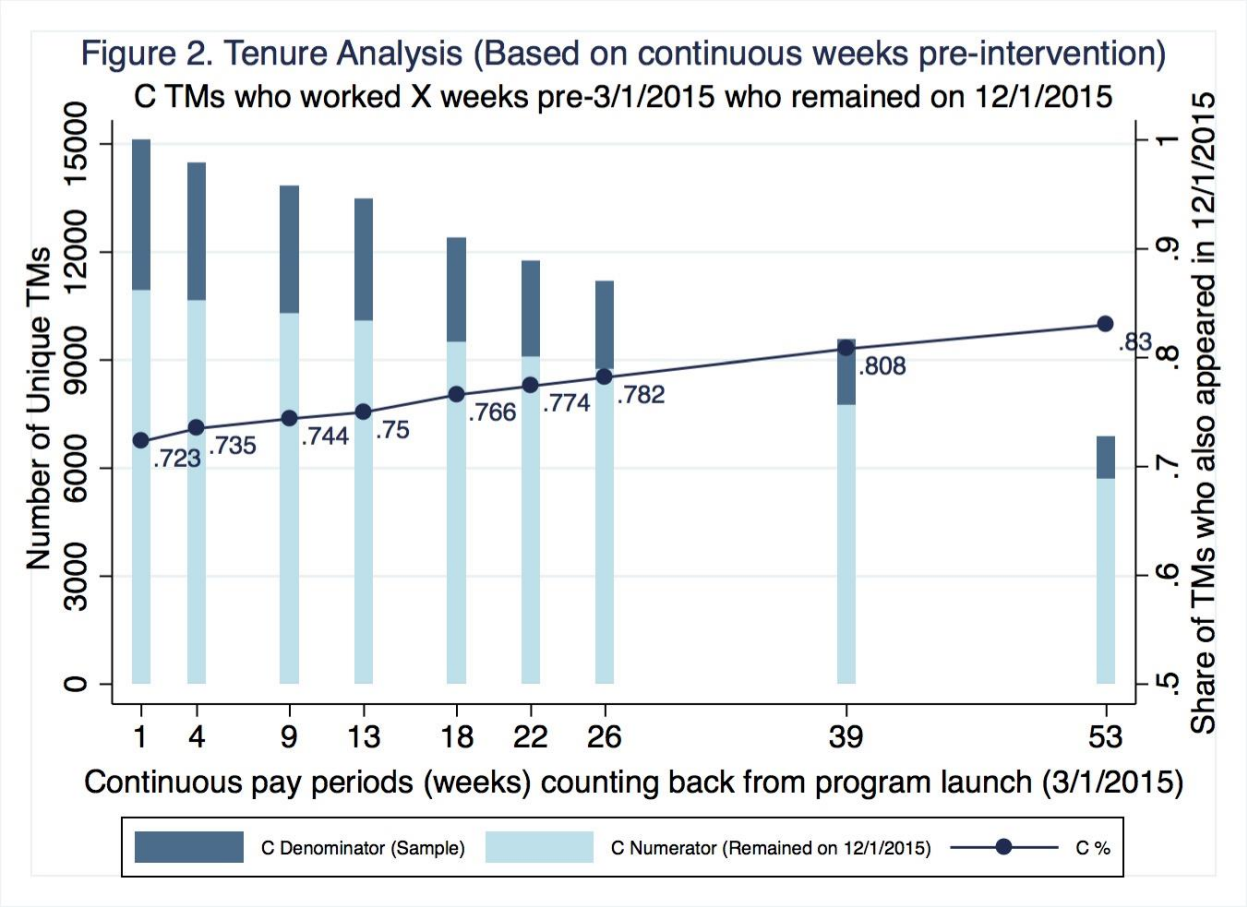
Similarly, Figure 1B shows a similar analysis when we extend the definition of retention to 12/1/2015. This graph contains only control worksite employees, because tenure itself may be affected by the wellness program and is an outcome we will examine formally in the analysis. To extend the above example of interpretation, of the 10,000 employees in control worksites who were employed through the first 17 weeks of 2014, about 64% were still employed on 12/1/2015. This decrease is from 79% at the beginning of 2015, implying that 15% of the sample (79% – 64% = 15%) were "lost" from the sample (e.g. terminated, left BJ's) during the first 11 months of 2015.



Analyzing samples defined with respect to 3/1/2015 (start of the first module): Figure 2 takes a different approach to looking at tenure. It looks at retention for samples of employees defined based on the number of continuous weeks worked immediately *before* the wellness treatment

launched (i.e. defined by counting backwards from 3/1/2015). Retention here is still defined as appearing in 12/1/2015. As above the figure contains only employees from control worksites.

As an example of interpretation, there were about 13,500 employees in control worksites who worked during the 13 continuous weeks (~3 months) before the start of the wellness program (counting back from 3/1/2015 -- i.e. Feb '15, Jan '15, and Dec '14). Among these 13,500 employees, about 10,000 remained actively working on 12/1/2015. This amounts to about a 75% retention rate.



We examined the rate of decline of this sample of employees among control worksites throughout the treatment. Figure 3 shows the rate of decline of the above control sample (those who worked for the 13 consecutive weeks leading up to the start of the treatment (3/1/2015), and illustrates the decline in the number and percent of this sample through 2017. The X axis shows the months elapsed since start of the treatment (0 is the end of February 2015, while 34 is the end of December 2017). This graph shows a smooth decline in the sample of employees to reach 51% by the end of December 2017

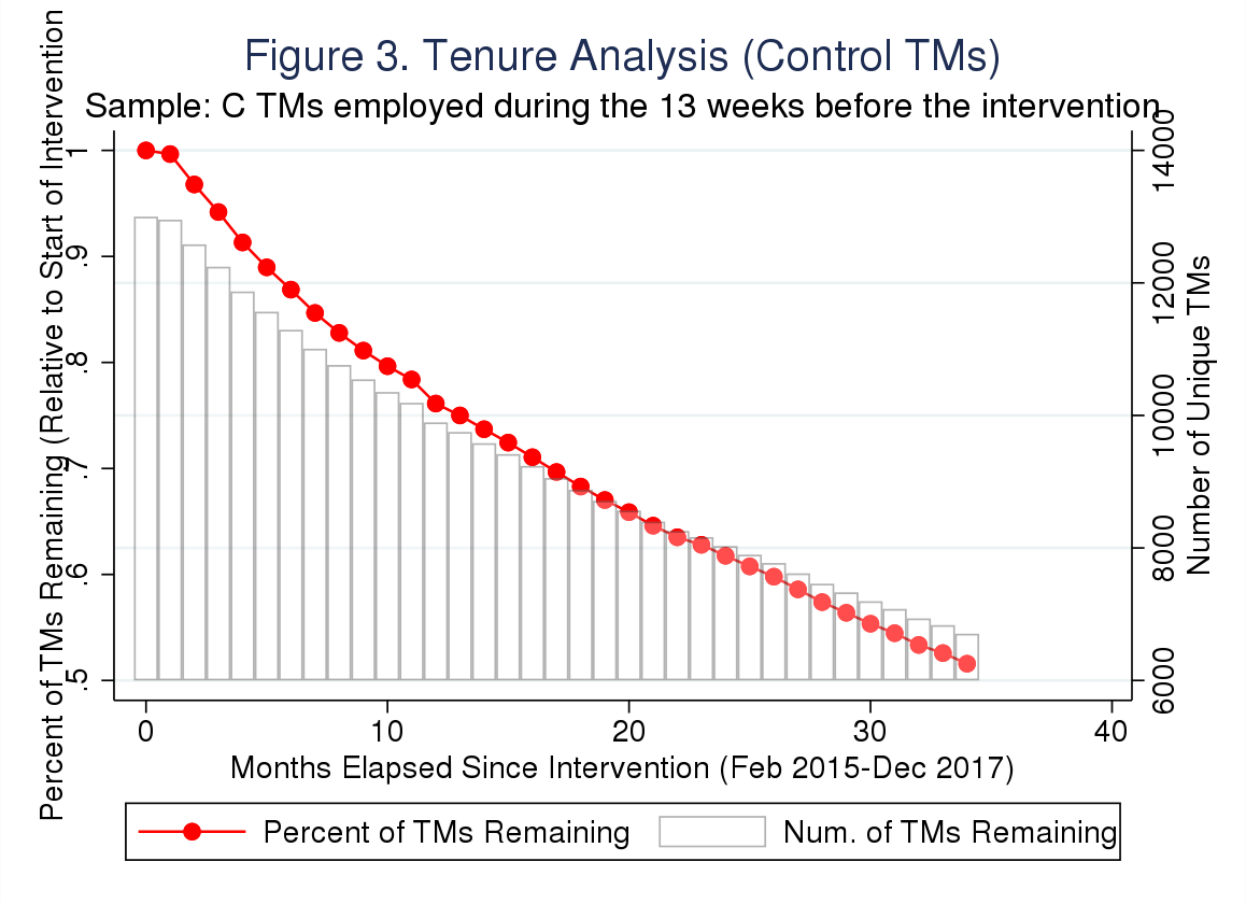


Figure 3. Tenure Analysis (Control TMs)
Sample: C TMs employed during the 13 weeks before the intervention
Percent of TMs Remaining (Relative to Start of Intervention
.5
.6
.7
.8
.9
1
Number of Unique TMs
6000
8000
10000
12000
14000
0
10
20
30
40
Months Elapsed Since Intervention (Feb 2015-Dec 2017)
Percent of TMs Remaining
Num. of TMs Remaining