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

Zirui Song[‡], Katherine Baicker[§]

Analysis Plan—Phase 1
March 15, 2018

[†] We are grateful to Jose Zubizarreta and Anna Sinaiko for helpful comments and to Kathryn Clark and Yuanxiaoyue Yang for excellent research assistance in the preparation of this analysis plan. We are also grateful to Jack Huang and Harlan Pittell for excellent research assistance in prior years and to Bethany Maylone for expert project management. 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 and Massachusetts General Hospital
[§] University of Chicago Harris School of Public Policy, Harvard School of Public Health, and NBER

Table of Contents

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, over 80 percent of large firms in the U.S. now offer a wellness program, frequently comprising a health risk assessment, biometric screenings, and a focus on topics such as weight loss, physical activity, and smoking cessation.[1] In addition to this massive private sector investment, the growth of workplace wellness has also been aided by public investments such as funds included in the Affordable Care Act. Despite the attention and investment in 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. Evidence from the few experimental or quasi-experimental studies is mixed.[2] Participants in wellness programs and firms offering them are likely different from 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.[3]

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. The analysis plan below details the implementation of the first phase of this intervention (January 2015 through June 2016) and the evaluation methodology. 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 post-treatment outcomes between the treatment and control groups. Institutional review board approval was granted and maintained through Harvard University.

[1] Kaiser/HRET Survey of Employer-Sponsored Health Benefits, 2016.

[2] See, for example, Jones D, Molitor D, Reif J. What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study. NBER Working Paper No. 24229. 2018; Fries et al. Randomized controlled trial of cost reductions from a health education program: the California Public Employees' Retirement System (PERS) study. Am J Health Promot. 1994;8 (3):216-23; and Leigh et al. Randomized controlled study of a retiree health promotion program. The Bank of America Study. Arch Intern Med. 1992;152 (6):1201-6.

[3] For reviews of prior experimental, quasi-experimental, and observational studies, see, for example, Baicker K, David C, Song Z. Workplace wellness programs can generate savings. Health Affairs, 2010;29(2): 304–311; the RAND Corporation, Mattke S, Schnyer C, Van Busum KR. A Review of the U.S. Workplace Wellness Market. 2012. (https://www.rand.org/pubs/occasional_papers/OP373.html).

II. Treatment: Workplace Wellness Program

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

The treatment took place in 2 phases. Phase 1 of the treatment period spanned 18 months, from January 2015 through June 2016 – and is the subject of this analysis plan (**Table 1**). In an ongoing Phase 2 of the study, the treatment was extended for another year and to additional clubs. Data for this extended treatment will be collected and analyzed separately. This treatment 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 treatment consisted of the opportunity to participate in a personal health assessment, in-person screenings, and multiple program modules. Each module took place over 4-7 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. Phase 1 comprised the following 8 modules: "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" (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. Across the program, employees had opportunities to receive incentives through completion of the personal health assessment, the biometric screenings, and participation in the individual modules of the program. Employees earned a $50 BJ's gift card for completing both the biometric screening and personal health assessment in each round of screenings and typically received a $25 BJ's gift card for completion of a module, with employees who had Cigna insurance coverage able to earn an additional incentive for some of the modules. Please refer to Appendix 1 for detailed information on the components of the wellness program by module, including requirements and incentives. **Table 2** shows average participation by module across the treatment clubs for Phase 1 of the wellness program.

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

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 108 employees per club. Many aspects of typical workplace wellness programs, including those studied in this context, focus on changing the workplace environment (such as changing the snacks in the breakroom or providing informational posters or seminars) and on team-based interventions (such as team step challenges) that would not be possible to evaluate with individual-level randomization within the worksite. We thus randomized our wellness intervention at the club level.

At the beginning of the study, there were 201 BJ's clubs in the U.S. along the East coast, extending from Maine to Florida. We eliminated 41 clubs from our sample because they were geographically remote or had employee pools with substantially different insurance coverage from the others, leaving 160 clubs in our sample.

Among these clubs, we randomly selected 20 "treatment" clubs that would receive the wellness program and 20 "primary control" clubs for Phase 1. Data from personal health assessments and in-person biometric screenings were collected in all of the treatment and primary control clubs. **Figure 1** shows the locations of these 40 clubs. The remaining 120 clubs served as "secondary controls" in Phase 1, and were included in analyses of the administrative data that were available for all clubs.

In Phase 2 of the treatment, we expanded fielding of the wellness program to 5 additional randomly selected treatment clubs and 5 additional randomly selected control clubs. During Phase 2 of the treatment, the 25 treatment clubs received an additional 4 modules of wellness programs.

IV. Data

This analysis draws on data from 5 categories. The table below displays the source of each data set and the study population for which it is available.

Summary of the Components of Data Collected				
Data	**Source**	**Data Availability**		
		Treatment Clubs (20 clubs)	**Primary Control Clubs (20 clubs)**	**Secondary Control Clubs (120 clubs)**
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	3rd party vendor (via Wellness Workdays)	Employees completing screening	Employees completing screening	None (by design)
Personal Health Assessment	3rd party vendor (via 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 include data on employment history and earnings and are available for all employees across treatment and control clubs (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, the large majority of whom are full-time employees. BJ's is a self-insured company (i.e. it bears risk for the health care spending of its employed population), with Cigna as the administrator of its health plans. In cross-section, approximately 35 percent of all BJ's employees are insured through Cigna during the study period.

A1. Employment records

Data on employment and earnings provided by BJ's enabled us to define our sample (based on hire and termination dates) and measure key employment-related outcomes, such as absenteeism and performance reviews. Employment history data capture all employment-related events associated with an employee, such as a hire or termination. Earnings data capture the number of hours worked and dollars earned by an employee in a given pay-period for a specific type of earnings. Variables in the employment records fall into the 4 general categories below.

- **Actions:** In the employment history data, employment-related actions include hire, rehire, termination, transfer, and performance review.

- **Locations and dates:** These data describe the worksites where employees worked and the start and end dates of their employment.

- **Demographic variables:** These include date of birth, gender, and race/ethnicity. These are discussed in greater detail in the Analysis section below.

- **Earnings:** These data include number of hours worked and dollars earned by an employee in a given pay-period for a specific type of earnings (regular time, overtime, etc.).

We used these data to study several categories of outcomes:

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

- **Performance review:** BJ's rates employee performance on a 5-point scale (1 through 5), with 1 representing the best performance rating and 5 the worst. Most employees have one performance review per calendar year (although not always in the year in which they are hired or terminated). 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 greater than or equal to 3 as poor performance.

- **Employment tenure:** Using data on action dates and earnings, we defined when an individual was employed by 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, with a maximum tenure of the entire length of the study.

Table 3 provides summary statistics for tenure, performance review, and absenteeism gathered among the control clubs during the treatment period.

A2. Medical and pharmaceutical claims data

Medical claims data were provided at the individual employee level by Cigna and were used to calculate spending and utilization variables. To standardize these outcome variables to a defined period of time at the individual level, we used the enrollment file detailing the length of Cigna coverage (measured in months) for each BJ's employee covered by Cigna. These administrative data are available for all clubs (treatment, primary control, and secondary controls), but only for employees who were enrolled in a BJ's employer-sponsored Cigna health insurance plan. Full-time workers were more likely to have Cigna coverage than part-time workers.

We analyzed medical claims for BJ's employees only (excluding their dependents, who were not directly exposed to the treatment). For each employee, we included claims with service dates during the intervention period, including an additional 30 days to capture potential billing delays.

We considered the entire treatment period (January 2015 through June 2016 for Phase 1 of the wellness program) as a whole. We aggregated medical spending and utilization at the employee level across this 18-month treatment period, normalized to daily rates based on the number of days employees were insured during the treatment period. We rescaled these outcomes to annual or monthly averages for ease of interpretation. We examined the following outcomes at the individual level as well as the club level.

- **Total medical spending**: We defined total medical spending per year as the sum of all payments, including deductibles, copayments, coinsurance, insurance payments, and any amount paid by another carrier, that appear on an employee's claims.

- **Medical spending by site of care**: We used the site of care variable in the claims data to categorize medical spending by different types of sites. These sites of care are mutually exclusive and exhaustive. The sites of care include: office, inpatient hospital, emergency department, outpatient hospital, urgent care, and other (home + SNF + missing site of care).

- **Out-of-pocket medical spending**: We also examine out-of-pocket spending, defined as the sum of the deductibles, copayments, and coinsurance. Out-of-pocket spending is a subset of total medical spending.

- **Utilization**: We defined a number of utilization variables:

    - *Physician office visits*: We defined an office visit as a claim line with site of care as "office" and service type as "physician visit." We consider multiple claims with the same patient, service date, and provider specialty as a single office visit. We did not include office visits that occurred with the site of care as "outpatient hospital." We examined a binary indicator for whether a subject had any office visits and the total number of office visits.

    - *Hospitalizations*: We defined hospitalizations based on days on which a patient has a claim line with site of care "inpatient hospital" and service type "hospital visit," treating claims from the same or continuous days as a single hospitalization (being careful that a

missing day of hospital-specific claims did not break what otherwise appeared to be one continuous stay into two). We examined both a binary indicator for having any hospitalizations and the total number of hospitalizations.

- *Emergency Room visits*: ER visits were identified from claim lines with site of care "Emergency Room – Hospital" and service type "emergency facility" or "emergency medical care," again ensuring that we do not double count ER visits when a claim for imaging, labs, or prescription drugs is received a day or two later than the actual ER visit. Similar to hospitalizations above, we treated claims for the same or continuous days as a single ER visit. We examined both a binary indicator for having any ER visits and the total number of ER visits.

- *Urgent care visits:* Urgent care visits were identified from claim lines with site of care "Urgent Care Facility." We treated all claims from a particular day for a particular patient as one visit. We examined both a binary indicator for having any urgent care visits and the number of urgent care visits.

- *Preventive care visits*: We identified preventive care visits using CPT codes corresponding to "preventive medicine services." These included 99384-99387 and 99394-99397. We considered multiple claim lines on same day with same provider specialty to be one visit. We examined both a binary indicator for having any preventive care visits and the number of preventive care visits.

While substantial additional granularity is available in the claims data, our sample sizes do not in general support condition-specific analyses. **Table 4** provides summary statistics for medical spending and **Table 5** provides summary statistics for utilization among employees insured at the control clubs during the treatment period. At the club level, control means were calculated as a weighted average across individuals, weighted by hours worked in the club.

Similar to medical claims, we used prescription drug claims to examine drug spending and utilization. We used the same method to scale drug utilization and spending as that described for medical spending and utilization above. We examined the following outcomes at the individual level as well as the club level.

- **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.

- **Out-of-pocket prescription drug spending**: Analogous to medical claims, we defined out-of-pocket spending as the sum of the deductibles, copayments, and coinsurance. Out-of-pocket spending is a subset of total prescription drug spending.

- **Utilization**: We defined a number of utilization variables:

- *Number of distinct drugs*: We defined number of distinct drugs as the count of different drug types (drug types are identified by generic names of drugs) a patient ever had during the study period.

- *Total quantity of prescriptions*: We defined total quantity of prescriptions as the sum of all prescription-months (e.g. one drug with three monthly fills was counted as three prescription-months). We also examined a binary indicator for having any prescription drugs, and a measure of the number of distinct drugs.

- *Quantity of medications by health condition*: We analyzed categories of common conditions and grouped medications by drug class into the following 8 health conditions: asthma, cardiovascular, diabetes, hyperlipidemia, mental health, pain, antibiotics, and other. These conditions were selected because they were more likely to be affected by the wellness treatment. We also examined binary indicators for having any prescription drugs for each of the 8 conditions.

Table 6 provides summary statistics for pharmaceutical spending and utilization (both total and by category of medication) among the control clubs during the treatment period. Again, at the club level, control means were weighted averages across individuals, weighted by hours worked.

B. Primary Data

Primary data consist of biometric data collected during in-person screenings conducted by registered nurses (employed by a third-party vendor) and self-reported data gathered from concurrently administered personal health assessment surveys. For completing this primary data collection, an employee received a $50 gift card. The participation rate was 52% in treatment clubs and 49% in control clubs. The biometric data included blood pressure, height and weight (enabling calculation of BMI), and blood measurements of cholesterol and blood sugar. Personal health assessments contain self-reported information on health behaviors, health, and wellbeing. These primary data are available for the individuals in the 20 treatment clubs and the 20 primary control clubs who completed the screenings during the summer of 2016.

B1. Biometric screening data

In the treatment and primary control clubs, we conducted biometric screenings at the conclusion of the wellness modules. The screenings were conducted by registered nurses and took place in the clubs. Unlike the administrative data above, biometric data are only available for employees in the treatment and primary control clubs who opted to complete the screening.

- **Total cholesterol**: We examined both total cholesterol as a continuous variable and a binary indicator of high cholesterol, defined as total cholesterol ≥200 mg/dl.

- **High-density lipoprotein (HDL) cholesterol**: We examined both HDL ("good cholesterol") as a continuous variable and a binary indicator for low HDL defined as HDL <40 mg/dl.

- **Blood glucose**: We examined blood glucose as a continuous variable in units of mg/dl.

- **Systolic and diastolic blood pressure**: We examined both blood pressure as a continuous variable and a binary indicator for high blood pressure or hypertension, with hypertension defined as systolic blood pressure ≥140 mmHg or diastolic blood pressure ≥90 mmHg.

- **Body mass index (BMI)**: We calculated BMI as weight in kilograms divided by the square of height in meters, and examined both a continuous measure and a binary indicator for obesity, with obesity defined as BMI ≥30.

Table 7 provides summary statistics for biometric screening data among employees who were screened in the control clubs.

B2. Personal Health Assessment Data

At the time of the biometric screenings, we also administered personal health assessment surveys in each of the treatment and primary control clubs. Employees were asked to fill out a paper survey, in which they were asked a variety of questions relating to their medical history, screenings and exams, emotional health, sleep, physical activity, nutrition, weight management, tobacco use, and alcohol use. We use this dataset to assess the impact of the wellness program on employees' health behaviors and self-reported health status. Again, these data are only available for employees in the treatment and primary control clubs who completed the personal health assessment.

Based on an examination of the distribution of PHA responses collected from the control group, we will examine the following outcomes.

- **Screenings and exams**:

  - *Annual exam*: We defined having an annual exam as a binary indicator with 1 equal to answering yes to the question "Have you had a physical exam or check-up by your healthcare provider (physician or nurse practitioner) in the last 12 months?"

  - *Flu shot*: We defined flu shot as a binary indicator with 1 equal to answering yes to the question "Do you receive the influenza vaccine (flu shot) annually?"

  - *Percent of other recommended tests received*: We considered other commonly-recommended tests (based on respondents' age and gender) discussed in the PHA as a group and determined the share of those tests that were obtained by the respondent, based on self-reports. These other recommended tests are cholesterol level, fasting blood glucose level, blood pressure, dental exam, colon cancer screening (for individuals aged 50-85), mammogram (for women aged 50-75), and pap smear (for women aged 21-65).

- **Mental health and well-being**:

  - *PHQ-2 score*: We used the PHQ-2, a pair of rapid depression screening questions that is commonly used in the primary care setting, to calculate a score for everyone who was screened. Individuals with a score of 3 or higher (the recommended cut-off) were flagged as possibly having depression.

  - *SF-8 score*: We used the eight question SF-8 health survey to measure self-reported functional health and well-being. We examined two validated scales from this survey as continuous variables: the physical summary score and the mental summary score.[4]

  - *Stress at work*: We defined stress at work using answers to the question "How often have you found yourself stressed or worried about problems as work?" Answer choices that we defined as 1 indicating stress included "sometimes," "fairly often," or "very often." Answer choices "almost never" or "never" were coded as 0.

  - *Unmanaged stress*: We examined a binary indicator for the presence of unmanaged stress. Individuals were asked whether they had stress in their life, and: "Stress management includes regular relaxation, physical activity, talking with others or making time for social activities. Do you effectively practice stress management in your daily life?" Those who declared that they had no stress or answered "yes" to this question were coded as 0, while those who had stress and answered "no" were coded as 1.

- **Sleep**: We created a binary indicator for getting an adequate amount of good quality sleep per night. This variable is based on responses to two questions: "Do you consider [the amount of sleep you reported getting] adequate for you?" and "Do you consider the quality of your sleep to be good?" Individuals who responded yes to both questions were coded as 1, while individuals who responded no to one or both questions were coded as 0.

- **Physical activity**:

  - *Regular exercise*: We defined regular exercise as answering yes to the question "Do you engage in regular exercise according to any of the definitions listed?" The provided definition 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."

[4] 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.

- *3+ days moderate exercise*: We defined this binary indicator as answering the question "During a typical week, on how many days do you do moderate physical activity or exercise that causes light sweating or slight to moderate increases in your breathing or heart rate (pulse)?" with a number greater than or equal to 3 days.

- *Number of days per week intentionally increase activity*: This continuous variable was defined as an individual's answer to the question "During a typical week, on how many days do you intentionally increase your activity level by going for walks, parking farther away, or taking the stairs rather than the elevator?"

- *Number of hours sitting per day*: This continuous variable was defined as an individual's answer to the 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."

- **Nutrition**:

  - *Number of meals eaten out*: This continuous variable was defined as an individual's answer to the question "During a typical week, how many meals do you eat at a fast-food, casual dining, or sit down restaurant?"

  - *Number of sweetened drinks per days*: This continuous variable was defined as an individual's answer to the question "How many naturally or artificially sweetened beverages do you consume per day? Please include regular and diet soft drinks, energy, and sports drinks."

  - *Read the Nutrition Facts panel*: We defined this binary indicator as 1 if individuals responded 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 binary indicator as 1 if individuals responded yes to the question "Do you eat at least 2 cups of fruit and $2^1/_2$ cups of vegetables per day?"

  - *Choose whole grain foods and reduced fat foods more often than the regular variety*: We defined this binary indicator as 1 if individuals responded yes to both questions "Do you choose 100% whole grain bread, pasta, rice, cereal and crackers more often than the regular (white) variety?" and "Do you choose low fat or reduced fat items more often than regular or full-fat products?" Individuals who responded no to one or both questions were coded as a 0.

- **Weight Management**:

  - *Considering losing weight*: We defined this as responding yes to the question "Are you seriously considering trying to lose weight to reach your goal in the next 6 months?"

- *Actively managing weight*: We defined this as responding yes to either or both of the questions "In the past month, have you been actively trying to lose weight?" and "In the past month, have you been actively trying to keep from gaining weight?" A response of no to both questions was coded as a 0.

- **Smoking**: This binary indicator was coded as 0 for people who responded that they had never smoked, were not regular smokers, or had quit smoking, and was defined as 1 for those who reported that they had smoked in the past and had not quit smoking.

- **Alcohol use**: This continuous variable was defined using two questions: "How many drinks do you have on a typical weekend day?" and "How many drinks do you have on a typical weekday?" The response to the first question was multiplied by two and the response to the second multiplied by five, and these two were added together to get a number of drinks per week.

- **Medical utilization**:

  - *Doctor visits in the last 12 months*: We defined the number of doctor visits as an individual's response to the question "In the last 12 months, how many times did you go to a doctor's office, clinic, or other health care provider to get care for yourself? Don't include emergency room or hospital visits. Your best estimate is fine." Responses are truncated at a number of 3 visits, with 3 representing 3 or more visits. We also examined a binary indicator for having any visit.

  - *Any ER visit in the last 12 months*: We defined this binary indicator as equal to 1 if the response to the question "In the last 12 months, how many times did you go to an emergency room to get care for yourself? Your best estimate is fine." was greater than or equal to 1 visit.

  - *Days spent in a hospital*: This continuous variable was defined as the response to the question "In the last 12 months, how many total days did you spend in a hospital? Your best estimate is fine."

  - *Ever a hospital patient in the last 12 months*: We defined this binary indicator as equal to 1 if days spent in a hospital was greater than or equal to 1 day.

  - *Number of different prescriptions in the last 12 months*: This continuous variable was defined as the response to the question "In the last 12 months, how many different prescription medications did you regularly take every day? Your best estimate is fine." Responses are capped at 6 prescriptions, with 6 representing 6 or more prescriptions.

  - *Any prescriptions in the last 12 months*: We defined this binary indicator as equal to 1 if number of different prescriptions in the last 12 months was greater than or equal to 1 prescription.

Table 8 provides summary statistics for the PHA survey data among employees who completed the PHA in the control clubs.

B3. Program participation data

One of our key independent (right-hand-side) variables is participation in the wellness program. Each module had its own set of requirements that defined completion of the module, along with financial incentives attached to participation or completion, as described in Appendix 1. We examined 3 different participation metrics based on the number of modules of the wellness program completed.

- **Participation indicator**: We defined a binary indicator of participation based on completing at least one module (any module) in Phase 1 of the wellness program. This is our primary definition of participation.

- **High participation**: We defined a second binary indicator based on completing 3 or more modules in Phase 1. We selected 3 modules as the cutoff given that, conditional on completing at least one module, 3 was the median of the distribution of modules completed in the study population.

- **Modules completed**: We also examined a continuous measure of the number of modules completed in Phase 1, which ranges from 0 to 8.

Table 13 provides control means of the above three definitions of Participation and first stage estimates of the impact of Treatment on these alternate definitions of Participation.

V. Study Sample

A. Main sample and subsamples

Our analysis draws on two study samples. Our main study sample comprised all employees who worked at BJ's Wholesale Clubs during any part of the study period. Each worker was weighted by the share of the study period s/he was employed at BJ's (described below).

One potential drawback of this broad sample inclusion criterion is the potential for endogenous entry or exit based on the treatment itself. For example, a worker's decision to join or exit BJ's could be a function of the availability of the wellness program. We can assess endogenous entry and exit directly in the data by testing whether the treatment affects job tenure, as in Table 3 (although this cannot address any endogenous change in the type of workers attracted to BJ's).

To (partially) address the issue of endogenous tenure, we also defined an alternative sample based on a reasonably stable subsample of employees who were continuously employed at BJ's in the 13 weeks immediately preceding the treatment. This "stably employed" subsample is thus immune from endogenous entry or composition, as it is defined based on presence in the sample in advance of the intervention. Members of this subsample were also much more likely to be employed at the end of the study period, as shown below. The choice of inclusion criteria (13 weeks of employment pre-randomization) was made to balance sample size and stability of employment going forward. Additional details on the construction of this subsample are provided in Appendix 2.

Exit from this sample may still potentially be endogenous, however, if the treatment affects whether workers remain employed at BJ's. There are two additional strategies for addressing this issue. First, we will assess the magnitude of any differential exit between treatment and control clubs empirically. We can then gauge the potential bias introduced from any observed differential exit using a bounding exercise. We similarly defined a subset of the stably employed subsample of employees who were continuously covered under Cigna insurance at BJ's in the 13 weeks immediately preceding the treatment at BJ's to address the issue of endogenous entry into Cigna insurance. This population is used to analyze claims-based outcomes that were only available for those with Cigna insurance.

Last, we also perform analysis at the club level. This level of analysis abstracts from individual-level employee turnover (under the assumption that the total size of the BJ's employee pool is exogenous), focusing on club-level employment and health care spending outcomes. To create aggregated data at the club level, we collapsed employee-level data to the club level as described in section VI below.

Table 9 provides a summary of demographic characteristics of the main sample and subsamples using data from control clubs.

B. Balance between treatment and control groups

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 key analytical samples. We also examined balance on baseline job and employment characteristics in the pre-intervention period for the stably employed subsample (as they by definition were in the data prior to the intervention). To augment club level analyses of balance, we added estimates of county-level characteristics from the 2015 U.S. Census Bureau matched to the locations of each club.

As **Table 10** demonstrates at the employee level, our randomly assigned clubs were balanced on some employee characteristics but not others. Notably, individuals in the treatment clubs were, on average, older (by 1-2 years, depending on the sample) and more likely to be white (by about 6-9 percentage points) than individuals in control clubs. This was consistent across the overall sample and subsamples of employees who completed the PHA and were enrolled in Cigna. On the other hand, observable employment characteristics (worker type, annual compensation rate, hours worked per week, and job category) were largely balanced between treatment and control for the "stably employed" subsample for whom they were observed at baseline.

Balance at the club level was analyzed using club-level measures derived from a weighted average of individual-level measures, weighted by individuals' hours worked. This imbalance at the individual level was also present after collapsing employees to the club level (**Table 11**).

To assess whether the imbalance was related to the overall demographics of where treatment clubs and control clubs were located in the U.S., we examined similar population characteristics of the counties in which the clubs were located using 2015 data from the U.S. Census Bureau. This showed a similar imbalance relative to that seen at the individual level, suggesting that the imbalance was at least in part driven by underlying differences in population characteristics of the areas where the clubs were located, from which the workforce was presumably drawn. This is also consistent with qualitative assessment of Figure 1, which shows, for example, treatment clubs were disproportionately represented in Ohio and Virginia, while control clubs were more likely to be located in Florida.

To deal with the imbalance on observable characteristics, we followed two strategies. First, in all specifications we controlled for the baseline demographic characteristics. Second, in our primary analyses we weighted the treatment and control groups on observed age, sex, and race so that both samples resembled the entire employee population. Neither of these strategies fully inoculates us against the possibility that, just as our random draw was by chance imbalanced on some demographic characteristics, our sample might be imbalanced on unobservable characteristics that are imperfectly correlated with observables for which we control and that also affect the outcomes of interest. While this is of course a possibility (as with many similar designs), we are less concerned given the balance on baseline employment characteristics.

VI. Statistical Analyses

Analyses will be conducted at the individual employee level and at the club level.

A. Intent-to-treat analysis

In the intent-to-treat analysis at the individual level, our goal is to estimate the average effect of a worker being randomized into a treatment club vs. a control club on outcomes of interest. We use a model that includes a treatment indicator capturing whether an individual was employed at a treatment vs. a control club. Individual-level observations were weighted based on the share of the intervention period during which the individual was employed at BJ's – or exposure to the intervention (discussed below). The model aims to answer the question: what is the effect of offering an individual the opportunity to participate in a wellness program? It is worth noting that in our experimental setting, individuals who worked at a treatment club but did not elect to participate actively in any of the wellness programming may still be "exposed" to the intervention by, for example, seeing posters in the common areas, sampling the healthier food made available in break rooms, or hearing about activities from participants at the club.

$$Y_{ij} = \beta_0 + \beta_1 TREATMENT_j + \beta_2 \boldsymbol{X}_{ij} + \varepsilon_{ij} \quad (1)$$

In this representative estimating equation, Y_{ij} denotes an outcome of interest for individual *i* who is employed in club *j*, such as medical spending. $TREATMENT_j$ is a binary indicator of whether the individual's club was randomized into the treatment or control arm. A small share of employees (2.6%) appeared in more than one club during the study period. We defined each individual's treatment or control status using the status of the club where the individual was originally employed, given that subsequent movement between clubs could in theory be endogenous. Standard errors are clustered by club.

The coefficient on $TREATMENT_j$ (β_1) indicates the effect of being randomized into a treatment club, or the intent-to-treat (ITT). The ITT estimate is informative for employers considering implementing a wellness program. $\boldsymbol{X}_{ij}$ represents a vector of covariates that may help improve precision as well as account for chance differences in characteristics between treatment and control groups. These include:

- Age indicators: <20 years (omitted), 20-34, 35-49, 50-64, 65 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): full-time vs. part-time, employee type (salaried vs. hourly), job category (sales vs. non-sales vs. other)

We used two sets of weights in our primary analysis. First, each individual was assigned an "exposure" weight based on the extent of his or her exposure to the wellness program (i.e. the treatment) during the study period. Many BJ's employees joined or left BJ's employment during the course of this 18-month intervention. Moreover, many worked far less than full time.

Outcomes for individuals with minimal exposure to the intervention are unlikely to be responsive to their small amount of time spent in a treatment vs. control club. Exposure weights are one way to account for this; an alternative would be estimating a dose-response model. We calculated this exposure weight using data on duration of employment and hours worked provided by BJ's. We summed the number of hours actually worked during the treatment period and divided by the number of hours a full-time employee would have worked during the study period, with weights resulting between 0 and 1. For example, a half-time worker who was employed for half of the treatment period would be assigned a weight of 0.25. See **Table 12A** for summary statistics of these weights in the control group. Due to the potential endogeneity with the treatment, we did not examine the distribution across the treatment group prior to conducting the analysis. We will test the balance on exposure weights between control and treatment groups similarly to the balance tests in **Table 10**, though without the weight for exposure.

Second, given that the treatment and control groups were not perfectly balanced on the set of observable characteristics after randomization, we derived a second set of weights that achieve balance between treatment and control workers on age, sex, and race—attributes that are not plausibly affected by the intervention. These balance weights were constructed to balance the demographic characteristics between the treatment and control groups with minimum variance between the weights, and were calibrated to be representative of the demographic attributes of the entire study population. This method has been shown to perform better than a model-based approach that fits a propensity score.[5] See **Table 12B** for summary statistics of the balance weights between the treatment and control groups. In primary analyses, we use a composite weight constructed by multiplying the exposure weights and the balance weights together. In secondary analyses, we reassess a set of key outcomes using only the exposure weights.

B. Local average treatment effect

While our ITT analysis above explores the effect of being randomized into a treatment club, a related but distinct question is: what is the effect of participating in the wellness program on the outcomes of interest? This second question will produce a different estimate because not all employees in treatment clubs chose to participate. Some may not have found the wellness program appealing, for example. Because of this endogenous participation choice, comparing those who participate in treatment clubs to all employees in control group clubs may produce biased estimates of the effect of participation. We therefore model the impact of participation on outcomes using a two-stage least squares (2SLS) specification:

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

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

$$PARTICIPATION_{ij} = \pi_0 + \pi_1 TREATMENT_j + \pi_2 \boldsymbol{X}_{ij} + \nu_{ij} \quad (4)$$

[5] 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.

γ_1 is the local average treatment effect (LATE) of participating in the wellness program. **Table 13** shows the results of estimating equation (4) for alternative definitions of participation. Our preferred specification uses the binary indicator for whether a person ever participated in any module the wellness program during the study period. In alternative specifications, we apply alternative definitions of participation including and indicator for participating in 3 or more modules and a continuous measure of the number of modules completed.

If no one in the control group received the treatment, we might interpret the 2SLS LATE as a treatment on the treated (TOT). This is nearly true by definition because, by construction, the control clubs did not have access to the wellness program modules. However, because we assign employees to clubs based on their initial locations of employment at the beginning of the study period, a few individuals in the data who moved from control clubs to treatment clubs during the study period did receive an opportunity to participate in the program. This accounts for the fact that the control group means in **Table 13** are nearly, but not exactly, zero.

C. Addressing the inclusion of multiple related outcomes

We have multiple measures that capture closely related outcomes. This introduces two issues: first, combining information from these metrics may increase power. Second, we need to account for the multiple estimates of closely related outcomes in our inferential statistics.

We assessed groups of related outcomes by pre-specifying three standardized treatment effects. Specifically, we generated standardized treatment effects for each of the following groups:

- Biometrics (systolic and diastolic BP, cholesterol, HDL, glucose, BMI)
- Health behaviors (all PHA outcomes except emotional health and medical utilization)
- Mental health and well-being (all of the mental health and well-being outcomes in the PHA)

We conduct multiple inference adjustment within categories of outcomes. We adjusted for the number of outcomes tested within domains – largely as defined by the outcomes grouped within a particular table.

For each outcome, we report standard, per-comparison p-values and adjusted "family-wise" p-values that take into account the multiple related outcomes we pre-specified within each outcome category. The adjusted p-value speaks to the probability of rejecting the null hypothesis (i.e. no effect of the intervention) on a given outcome under the null hypothesis that the intervention had no effect on any of the outcomes in that category. We used the Westfall and Young method for calculating these adjusted p-values (which, unlike the Bonferroni method, does not assume independence across the outcomes within a category).[6]

[6] See, for example, Westfall PH, Young SS. Resampling-based multiple testing: Examples and methods for p-value adjustment. Wiley & Sons, 1993, and Kling JR, Liebman JB, Katz LF. Experimental Analysis of Neighborhood Effects. Econometrica. 2007;75(1):83-119.

D. Pre-specified subgroup analyses

We will perform two subgroup analyses at the individual level. We will assess differences in the effect of the wellness program by age and sex—two dimensions along which we observed fairly substantial differences in means between the treatment and control groups (**Table 14**)—via interaction terms. Equation 5 shows this interaction in our base ITT framework for age (characterized by an indicator for being age 40 or over).

$$Y_{ij} = \beta_0 + \beta_1 TREATMENT_j + \beta_2 Age40_i * TREATMENT_j + \beta_3 \boldsymbol{X}_{ij} + \varepsilon_{ij} \quad (5)$$

The effect of the wellness program on those under 40 is estimated by β_1, while the effect for those 40 and older is estimated by the sum of the coefficients $\beta_1 + \beta_2$. Age categories continue to be included in covariates $\boldsymbol{X}$.

E. Club-level analyses

We complement our analyses at the individual level with analyses at the club level. Club-level data were generated by aggregating employees assigned to clubs based on their first appearance in the data. We regression-adjust for demographics at the individual level before aggregation, and weight individuals based on their hours worked to form club-level averages. The resulting club-level dataset comprised 160 data points, one for each club (20 intervention, 20 primary control, and 120 secondary control).

We focus on outcomes measured in administrative data for all employees, dictated by data availability but also representing the employer perspective on aggregate outcomes affected by the decision to have a wellness program. Our estimation equation is:

$$Y'_j = \beta_0 + \beta_1 TREATMENT_j + \varepsilon_j \quad (6)$$

In equation (6), the subscript j denotes a club. Y'_j represents a club-level average outcome. $TREATMENT_j$ is a binary indicator of randomization into treatment the treatment group, with β_1 indicating the average club-level effect of being randomized into treatment. Covariates were, as noted, incorporated at the individual level before aggregation. Standard errors were adjusted for heteroscedasticity.

F. Sensitivity analyses and secondary analyses

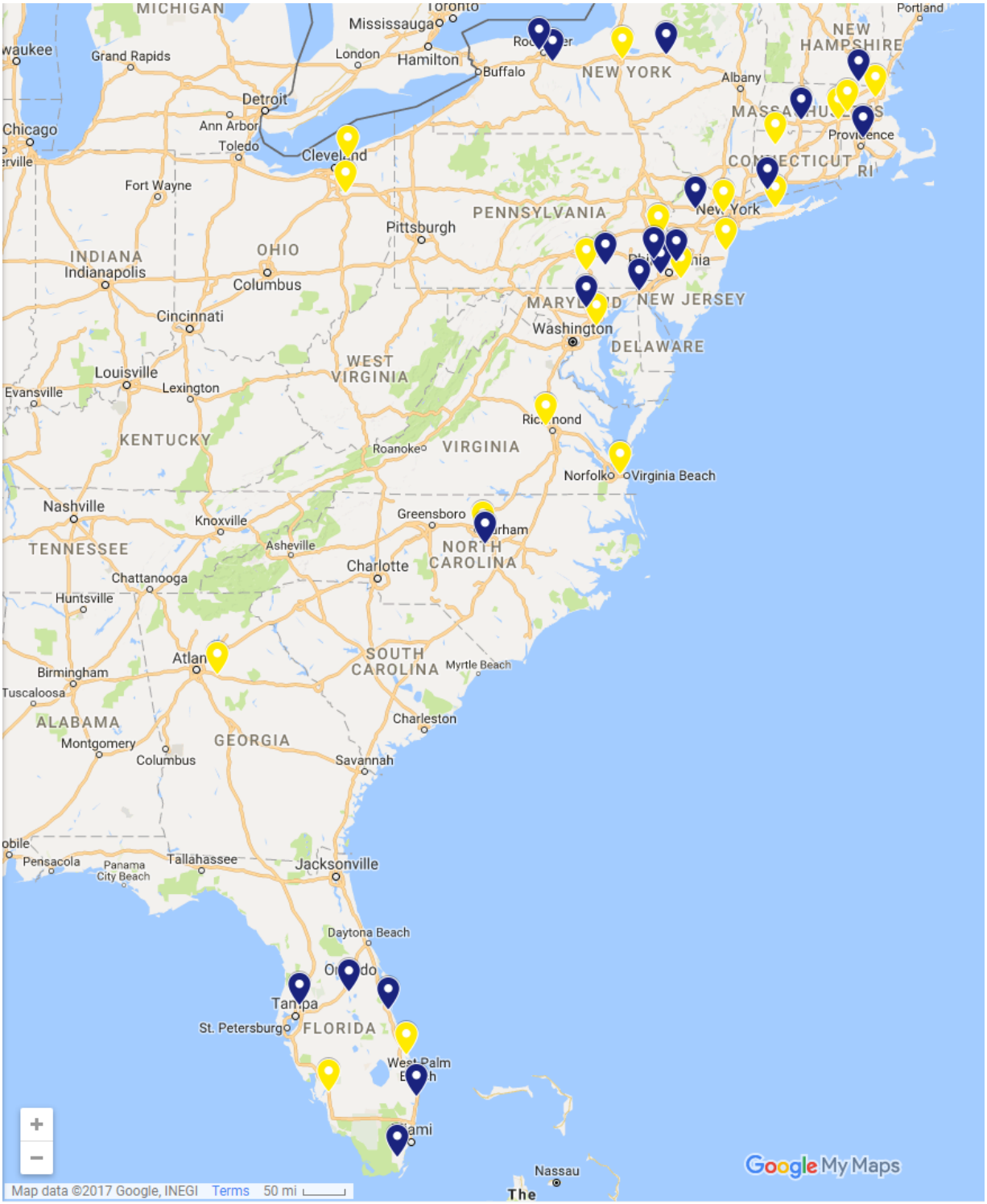
In the statistical analyses above, our base regression models use least squares specifications (OLS for ITT, 2SLS for LATE) for both continuous and binary outcomes. This approach has both strengths and weaknesses.[7] To test the robustness of our results, we estimate alternative functional forms, notably logit models for binary outcome variables.

[7] See, for example, Buntin MB, Zaslavsky AM. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. J Health Econ. 2004 May;23(3):525-42; Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. J Health Econ. 2005 May;24(3):465-88.

As noted above, in secondary analyses we reassessed a set of key outcomes using the exposure weights without the balance weights. The key outcomes were: total medical spending, total prescription drug spending, absenteeism, systolic blood pressure, BMI, annual exam (binary), SF-8 mental and physical health score, regular exercise, number of sweetened drinks per day, smoking (binary), and the number of alcoholic drinks consumed per week.

Given uncertainty about the functional form of the effect of participation in a multifaceted program on outcomes, we also test the sensitivity of our results using alternative definitions of participation. Specifically, we tested a definition of participation based on a threshold of completing at least 3 modules, as well as a continuous metric of participation as the number of modules completed. We present the results of these sensitivity and secondary analyses in additional tables that allow comparison to the main estimates.

Figure 1: Location of Treatment and Control Clubs



Notes: This map shows the 20 treatment and 20 control clubs in Phase 1 of the treatment. Yellow markers designate treatment clubs. Blue markers designate control clubs.

Table 1: Timeline of the wellness programs

Events	Phase 1																						Phase 2 begins ↓
	Program announced ↓					Registered Dietitians begin working in the treatment clubs ↓																	
Year	**2015**												**2016**										
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**
Screenings Round 1																								
Module 1. Take Charge of Your Health																								
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																								
Screenings Round 2																								

Notes: This table presents a graphical illustration of Phase 1 of the wellness program. The treatment began in 2015 with announcements of the wellness program club assignments (treatment clubs) in January followed by administration of the personal health assessments and in-person screenings in February. Phase 1 comprised 8 modules and concluded at the end of June 2016. After Phase 1, personal health assessments and in-person screenings were conducted during the summer of 2016. Afterwards, Phase 2 of the wellness program began in the fall of 2016. Due to an imbalance in the participation rates in the first screenings, data from Screenings Round 1 is excluded from our analysis.

Table 2: Average Participation Rates by Module, Phase 1

	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
Overall	12.2%	25.6%	37.7%	28.6%	31.6%	33.4%	28.7%	28.5%

Notes: Participation rate is calculated as the percentage of individuals who completed a module out of the number of employees eligible to complete a module during the time frame 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. 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. Numbers are weighted by the number of days an individual was working during a given module's timeframe.

Table 3: Impact on Employment

	Employee-level			Stably Employed Subsample			Club-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 (%)	2.63			2.86			2.57		
	(1.64)			(1.63)			(0.36)		
Performance Review (% ≥2)	60.48			66.49			59.04		
	(48.89)			(47.20)			(13.72)		
Tenure (days during treatment)	467.61			515.18			466.30		
	(137.34)			(88.92)			(19.75)		
N	32973			15344			160		

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 with p-values in brackets and family-wise p-values in curly braces. Column 1 reports the mean of each employment 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, Cigna coverage status, full-time status, paid hourly status, and job category) and cluster standard errors at the club (for employee-level regressions). Employee-level regressions and control means are weighted by the combination of a weight for exposure to the wellness program and a weight that balances treatment and control samples on demographics. Club-level regressions and control means are unweighted.

Table 4: Impact on Medical Spending

	Employee-level			Stably Employed Subsample			Club-level		
	Mean in Control Group (1)	Reduced Form (Linear) (2)	2SLS (Linear) (3)	Mean in Control Group (1)	Reduced Form (Linear) (2)	2SLS (Linear) (3)	Mean in Control Group (1)	Reduced Form (Linear) (2)	2SLS (Linear) (3)
Spending									
Total Spending	3975.56			3833.69			3989.63		
	(14784.11)			(13035.30)			(2420.31)		
Out-of-pocket Spending	781.87			743.73			780.47		
	(1213.89)			(1057.59)			(226.34)		
By Site of Care:									
Office	2145.43			2178.63			2165.27		
	(7396.97)			(7491.70)			(1271.91)		
Inpatient Hospital	1157.67			1016.85			1152.71		
	(9284.59)			(7329.21)			(1400.39)		
Emergency Room	529.27			497.16			528.43		
	(1759.02)			(1614.33)			(309.13)		
Urgent Care	25.64			25.13			25.68		
	(109.20)			(104.21)			(20.03)		
Other	117.55			115.93			117.55		
	(1343.70)			(1406.94)			(188.61)		
N	7631			6016			160		

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 with p-values in brackets and family-wise p-values in curly braces. Column 1 reports the mean of each medical spending and utilization 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 club (for employee-level regressions). Employee-level regressions and control means are weighted by the combination of a weight for exposure to the wellness program and a weight that balances treatment and control samples on demographics. Club-level regressions and control means are unweighted.

Table 5: Impact on Medical Utilization

	Employee-level			Stably Employed Subsample			Club-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)
Utilization									
By Site of Care:									
Any Physician Visit (%)	71.87			75.15			71.75		
	(44.96)			(43.22)			(7.90)		
Number of Physician Visits	3.23			3.29			3.22		
	(4.13)			(4.06)			(0.75)		
Any Hospitalization (%)	6.70			6.64			6.67		
	(25.00)			(24.89)			(3.74)		
Number of Hospitalizations	0.07			0.06			0.07		
	(0.33)			(0.30)			(0.05)		
Any ER Visit (%)	21.52			22.17			21.49		
	(41.10)			(41.55)			(7.50)		
Number of ER Visits	0.26			0.23			0.26		
	(0.67)			(0.56)			(0.11)		

Any Urgent Care Visit (%)	13.16 (33.81)	13.71 (34.40)	13.36 (7.55)
Number of Urgent Care Visits	0.14 (0.47)	0.14 (0.44)	0.15 (0.09)
Any Preventive Care Visit (%)	36.03 (48.01)	38.86 (48.75)	35.96 (10.78)
Number of Preventive Care Visits	0.36 (0.57)	0.37 (0.53)	0.36 (0.12)
N	7631	6016	160

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 with p-values in brackets and family-wise p-values in curly braces. Column 1 reports the mean of each medical spending and utilization 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 club (for employee-level regressions). Employee-level regressions and control means are weighted by the combination of a weight for exposure to the wellness program and a weight that balances treatment and control samples on demographics. Club-level regressions and control means are unweighted.

Table 6: Impact on Prescription Pharmaceutical Spending and Utilization

	Employee-level			Stably Employed Subsample			Club-level		
	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)
Rx Spending									
Total Spending	1221.15			1147.81			1207.99		
	(7467.09)			(5346.43)			(1230.27)		
Out-of-pocket Spending	93.93			98.42			94.14		
	(170.24)			(172.56)			(28.73)		
Rx Utilization									
Any Medications (%)	58.65			61.62			58.60		
	(49.25)			(48.64)			(8.57)		
No. of Distinct Medications	4.02			4.29			4.01		
	(4.75)			(4.81)			(0.90)		
Total Medication-Months	11.10			11.64			11.07		
	(19.79)			(20.18)			(3.85)		
By Category:									
Any Asthma Medications (%)	11.82			12.63			11.75		
	(32.29)			(33.22)			(5.23)		

No. of Asthma Medication-Months	0.51 (2.50)	0.52 (2.50)	0.50 (0.41)
Any Cardiovascular Medications (%)	22.36 (41.67)	23.92 (42.66)	22.09 (7.76)
No. of Cardiovascular Medication-Months	2.57 (6.54)	2.73 (6.70)	2.53 (1.13)
Any Diabetes Medications (%)	7.09 (25.67)	7.52 (26.37)	6.92 (3.99)
No. of Diabetes Medication-Months	0.96 (4.55)	1.02 (4.74)	0.92 (0.70)
Any Hyperlipidemia Meds (%)	14.00 (34.70)	15.11 (35.81)	13.72 (6.20)
No. of Hyperlipidemia Medication-Months	1.15 (3.50)	1.22 (3.58)	1.13 (0.63)
Any Mental Health Medications (%)	17.51 (38.01)	18.20 (38.59)	17.83 (7.22)
No. of Mental Health Medication-Months	1.66 (5.33)	1.72 (5.40)	1.70 (1.05)

Any Pain Medications (%)	17.62 (38.10)	18.68 (38.98)	17.60 (7.20)
No. of Pain Medication-Months	0.75 (2.74)	0.76 (2.78)	0.77 (0.54)
Any Antibiotic Medications (%)	12.85 (33.46)	14.16 (34.86)	12.74 (5.54)
No. of Antibiotic Medication-Months	0.39 (1.61)	0.40 (1.65)	0.39 (0.23)
Any Other Medications (%)	34.36 (47.49)	36.43 (48.13)	34.30 (8.03)
No. of Other Medication-Months	3.12 (7.11)	3.27 (7.26)	3.12 (1.14)
N	7631	6016	160

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 with p-values in brackets and family-wise p-values in curly braces. Column 1 reports the mean of each prescription drug spending and utilization 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 club (for employee-level regressions). Employee-level regressions and control means are weighted by the combination of a weight for exposure to the wellness program and a weight that balances treatment and control samples on demographics. Club-level regressions and control means are unweighted.

Table 7: 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)
Continuous Variables						
Cholesterol (mg/dl)	177.60 (41.45)			178.89 (41.37)		
HDL (mg/dl)	52.98 (16.37)			53.55 (16.40)		
Glucose (mg/dl)	101.96 (33.50)			101.76 (32.06)		
Systolic BP (mmHg)	124.29 (16.88)			124.84 (16.81)		
Diastolic BP (mmHg)	79.70 (10.56)			80.09 (10.49)		
BMI	29.70 (7.09)			29.61 (7.00)		
Binary Indicator Variables						
High Cholesterol (≥200)	29.37 (45.57)			30.73 (46.17)		
Low HDL (HDL <40)	22.29 (41.64)			20.78 (40.60)		
Hypertensive (SBP ≥140 or DBP ≥90)	23.10 (42.17)			24.18 (42.85)		
Obese (BMI ≥30)	43.04 (49.54)			42.73 (49.51)		

Standardized treatment effect		
N	2168	1353

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 with p-values in brackets and family-wise p-values in curly braces. Column 1 reports the mean of each biometric 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, Cigna coverage status, full-time status, paid hourly status, and job category) and cluster standard errors at the club. Employee-level regressions and control means are weighted by the combination of a weight for exposure to the wellness program and a weight that balances treatment and control samples on demographics. Standardized treatment effect is calculated using the continuous variables.

Table 8: 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)
Screenings and Exams						
Annual exam (%)	65.49			65.08		
	(47.57)			(47.71)		
Flu shot (%)	35.22			33.48		
	(47.79)			(47.23)		
Percent of other recommended tests received	55.97			57.15		
	(31.03)			(30.59)		
Mental Health and Well-being						
PHQ-2 score of 3 or above (%)	8.57			8.43		
	(28.01)			(27.80)		
SF-8 score – physical summary score	50.79			50.92		
	(7.72)			(7.72)		

SF-8 score – mental summary score	51.17 (9.09)	51.22 (9.09)
Stress at work (%)	55.60 (49.71)	58.12 (49.38)
Unmanaged stress (%)	41.77 (49.35)	41.38 (49.30)
Sleep		
Good quality, adequate amount of sleep (%)	54.16 (49.85)	54.54 (49.83)
Physical Activity		
Regular exercise (%)	61.88 (48.59)	63.20 (48.27)
Three or more days per week of moderate exercise (%)	63.95 (48.04)	64.10 (48.01)
Number of days per week intentionally increase activity	3.05 (2.37)	3.07 (2.36)
Number of hours sitting per day	3.49 (1.73)	3.49 (1.73)

Nutrition		
Number of meals eaten out	1.85 (1.56)	1.82 (1.54)
Number of naturally or artificially sweetened drinks per day	1.84 (1.86)	1.80 (1.84)
Read the Nutrition Facts panel (%)	58.74 (49.26)	58.48 (49.32)
Consume at least 2 cups of fruit and 2.5 cups of vegetables per day (%)	57.55 (49.45)	57.31 (49.50)
Choose whole grain foods and reduced fat foods more often than the regular variety (%)	33.23 (47.13)	34.61 (47.61)
Weight Management		
Considering losing weight in the next 6 months (%)	56.26 (49.63)	56.45 (49.63)

Actively managing weight (%)	54.68 (49.81)	54.72 (49.82)
Tobacco Use		
Smoker (%)	24.68 (43.14)	24.63 (43.12)
Alcohol Use		
Number of drinks per week	4.65 (7.41)	4.72 (7.38)
Medical Utilization		
Number doctor visits in last 12 months	1.52 (1.12)	1.53 (1.11)
Any doctor visit in last 12 months (%)	75.56 (43.00)	76.08 (42.69)
Any ER visit in last 12 months (%)	25.84 (43.80)	25.32 (43.52)
Ever hospital patient in the last 12 months (%)	17.54 (38.05)	17.55 (38.07)

Days spent in hospital	0.43 (1.37)	0.43 (1.40)
Number different prescriptions last 12 months	1.32 (1.64)	1.31 (1.64)
Any prescriptions in last 12 months (%)	52.91 (49.94)	52.39 (49.98)
Standardized treatment effect (mental health and well-being)		
Standardized treatment effect (health behaviors)		
N	2168	1353

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 with p-values in brackets and family-wise p-values in curly braces. Column 1 reports the mean of each self-reported health 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, Cigna coverage status, full-time status, paid hourly status, and job category) and cluster standard errors at the club. Employee-level regressions and control means are weighted by the combination of a weight for exposure to the wellness program and a weight that balances treatment and control samples on demographics. Standardized treatment effect for mental health and well-being is calculated using the outcomes under *Mental Health and Well-being* and the standardized treatment effect for health behaviors uses the outcomes under *Screenings and Exams*, *Sleep*, *Physical Activity*, *Nutrition*, *Weight Management*, *Tobacco Use*, and *Alcohol Use*.

Table 9. Summary of Demographic Characteristics for Employees in Control Clubs

	Employee-level			Stably Employed Subsample			Club-level	
	All-in	**PHA**	**Cigna**	**All-in**	**PHA**	**Cigna**	**All-in**	**Cigna**
Age (yrs)	39.2	40.6	45.0	41.3	41.3	45.5	39.3	40.3
Female (%)	45.8	56.3	46.6	46.1	56.9	46.6	45.8	44.8
Race (%)								
White	54.1	60.8	60.7	55.8	61.0	60.7	57.8	66.6
Black	21.2	18.7	17.5	19.6	18.3	17.3	20.0	14.7
Hispanic	18.5	16.7	16.8	18.6	16.7	17.1	16.6	14.9
Other race	6.2	3.8	5.1	6.0	3.9	5.0	5.6	3.8
Employment (%)								
Full-time salary	13.0	17.3	24.4	15.9	20.0	25.5	13.7	42.7
Full-time hourly	47.5	49.0	65.4	46.1	47.7	68.2	46.9	33.9
Part-time hourly	39.6	33.7	10.2	38.0	32.3	6.3	39.4	23.5
Worker Type (%)								
Sales worker	35.6	36.0	20.8	32.2	33.2	19.4	35.8	24.3
Nonsales worker	47.6	41.8	51.6	48.5	41.9	51.8	46.8	30.4
Other worker	16.8	22.3	27.6	19.3	24.9	28.8	17.4	45.3
N	32973	2168	7631	15344	1353	6016	160	160

Notes: Table lists demographic characteristics for the sample covered by Cigna weighted by months of Cigna coverage. About 35% of the total sample has Cigna coverage. Age is defined as age at the mid-point of the treatment period (October 2015). 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. The PHA subgroup includes all employees who answered at least one question on the PHA survey, including individuals who moved into a primary club during their employment and were eligible to take the PHA but have their treatment status marked by the first club they were in (a secondary control).

Table 10: Balance Between Treatment and Control—Employee Level

Panel A: All Employees – Exposure Weights Only

	(1) Treatment (n=4037)	(2) Primary Control (n=4106)	(3) Primary + Secondary Control (n=28936)	(4) (1) vs (2) P value	(5) (1) vs (3) P value
Demographics					
Age (yrs)	39.5	38.1	38.4	0.057	0.040
Female (%)	48.0	46.3	45.8	0.380	0.168
Race (%)				0.047	0.181
Black	16.1	20.4	21.2		
White	67.8	58.5	54.1		
Hispanic	10.6	17.5	18.5		
Other	5.4	3.6	6.2		

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 club an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between employees at treatment clubs and employees at primary control clubs. Column 5 reports the p-value for the comparison between employees at treatment clubs and all employees at control clubs. All regressions are weighted by individual exposure to the treatment.

Panel B: All Employees

	(1) Treatment (n=4037)	(2) Primary Control (n=4106)	(3) Primary + Secondary Control (n=28936)	(4) (1) vs (2) P value	(5) (1) vs (3) P value
Demographics					
Age (yrs)	38.8	38.3	38.7	0.539	0.839
Female (%)	46.4	45.6	46.0	0.639	0.762
Race (%)				0.749	0.992
Black	19.9	20.1	20.7		
White	56.3	57.9	55.3		
Hispanic	17.8	17.1	17.8		
Other	6.0	5.0	6.2		

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 club an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between employees at treatment clubs and employees at primary control clubs. Column 5 reports the p-value for the comparison between employees at treatment clubs and all employees at control clubs. All regressions are weighted by the combination of a weight for individual exposure to the wellness program and a weight that balances treatment and control samples on demographics.

Panel C: PHA Sub-sample

	(1) Treatment (n=1080)	(2) Primary Control (n=1020)	(3) (1) vs (2) P value
Demographics			
Age (yrs)	41.2	40.1	0.251
Female (%)	57.7	57.3	0.879
Race (%)			0.996
Black	19.1	19.1	
White	57.6	59.0	
Hispanic	18.4	16.7	
Other	5.0	5.2	

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 club an employee appears in during the treatment period. Column 3 reports the p-value for the comparison between employees at treatment clubs and employees at primary control clubs. All regressions are weighted by the combination of a weight for individual exposure to the wellness program and a weight that balances treatment and control samples on demographics.

Panel D: Cigna Sub-sample

	(1) Treatment (n=1005)	(2) Primary Control (n=986)	(3) Primary + Secondary Control (n=6626)	(4) (1) vs (2) P value	(5) (1) vs (3) P value
Demographics					
Age (yrs)	44.9	43.8	44.3	0.142	0.293
Female (%)	47.0	44.8	46.6	0.335	0.848
Race (%)				0.831	0.989
Black	16.0	15.3	16.9		
White	60.8	66.1	62.0		
Hispanic	17.9	14.5	16.0		
Other	5.3	4.1	5.1		

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 club an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between employees at treatment clubs and employees at primary control clubs. Column 5 reports the p-value for the comparison between employees at treatment clubs and all employees at control clubs. All regressions are weighted by the combination of a weight for months of Cigna coverage and a weight that balances treatment and control samples on demographics.

Panel E: Stably Employed Sub-sample

	(1) Treatment (n=1892)	(2) Primary Control (n=1930)	(3) Primary + Secondary Control (n=13452)	(4) (1) vs (2) P value	(5) (1) vs (3) P value
Demographics					
Age (yrs)	41.1	40.0	40.6	0.144	0.384
Female (%)	46.8	45.4	46.2	0.506	0.760
Race (%)				0.803	0.973
Black	18.9	18.6	19.1		
White	57.1	59.8	57.0		
Hispanic	18.4	17.0	17.9		
Other	5.6	4.6	6.0		
Employment					
Worker type (%)				0.676	0.158
FT salary	15.3	15.0	16.2		
FT hourly	44.9	47.1	46.2		
PT hourly	39.8	37.9	37.5		
Annual rate ($)					
FT salary	49504	47776	48654	0.139	0.451
FT hourly	25722	24533	25298	0.088	0.439
PT hourly	10302	9982	10034	0.084	0.085
Standard Hours Per Week					
FT salary	40.0	40.0	40.0	-	0.090
FT hourly	35.7	35.9	36.0	0.523	0.173
PT hourly	20.0	20.0	20.0	0.308	0.404
Job Category (%)				0.736	0.312
Sales workers	34.4	32.6	32.0		
Laborers/Helpers	19.9	20.5	20.4		
Operatives	16.2	15.4	15.9		
Service workers	11.6	13.1	12.2		
First/Mid level officials	11.5	11.4	12.5		
Admin Support	4.3	4.4	5.1		
Other	2.0	2.6	2.0		
Health Insurance					
Ever Enrolled in Cigna (2014)	49.9	50.0	48.2	0.933	0.150
Months in Cigna	11.5	11.5	11.6	0.914	0.534
Total medical spending (mean)	5388	3160	4187	0.087	0.344
Total medical spending (median)	996	874	984		

Notes: Employees are included if they were part of the stably employed subsample. All variables are pre-randomization characteristics. Demographics and employment characteristics are taken from the Team Member database supplied by BJ's and based on a December, 2014 snapshot of the employee's position. 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 club an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between employees at treatment clubs and employees at primary control clubs. Column 5 reports the p-value for the comparison between employees at treatment clubs and all employees at control clubs. All regressions are weighted by the combination of a weight for individual exposure to the wellness program and a weight that balances treatment and control samples on demographics.

Table 11: Balance Between Treatment and Control—Club Level

	(1) Treatment (n=20 clubs)	(2) Primary Control (n=20 clubs)	(3) Primary + Secondary Control (n=140 clubs)	(1) vs (2) P value	(1) vs (3) P value
Employee Demographics					
Age (yrs)	39.4	38.1	38.4	0.092	0.065
Female (%)	48.9	46.0	45.8	0.137	0.049
Race (%)				0.006	0.164
Black	18.0	20.3	20.0		
White	68.0	59.4	57.8		
Hispanic	9.0	16.6	16.6		
Other	4.9	3.6	5.6		
ACS Demographics					
Age (yrs)	40.7	39.5	39.6	0.262	0.220
Female (%)	51.2	51.4	51.3	0.410	0.718
Race (%)				0.004	0.013
Black	12.2	13.8	13.3		
White	77.7	76.3	74.5		
Hispanic	9.4	15.4	13.8		
Other	10.1	9.8	12.2		

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 club is located in. Club-level analyses are obtained by first calculating a weighted average for each club (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 club an employee appears in during the treatment period. Column 4 reports the p-value for the comparison between the employees at treatment clubs and the employees at primary control clubs and column 5 reports the p-value for the comparison between the employees at treatment clubs and all employees at control clubs.

Table 12: Summary Statistics for Weights

Panel A: Exposure Weights

	(1) All Employees, Control (n = 28,936)	(2) Stably Employed Sub-Sample, Control (n = 13,452)
Mean	0.38	0.61
Standard Deviation	0.32	0.29
Percentiles:		
Min	0.00008	0.008
1 %	0.002	0.05
5%	0.008	0.10
10%	0.02	0.16
25%	0.08	0.36
50% (Median)	0.28	0.68
75%	0.68	0.86
90%	0.88	0.92
95%	0.92	0.96
99%	1.00	1.00
Max	1.00	1.00

Notes: Exposure weights represent the share of the intervention period in which the individual is employed (and thus exposed to the intervention), calculated as the number of hours worked, divided by the number of hours a full-time worker would have worked during the length of the intervention, bounded between 0 (no exposure and excluded from the sample) and 1 (full-time work for the duration of the intervention). Because exposure is potentially endogenous, we only show the control group here.

Panel B: Balance Weights

	(1) Treatment (n = 4037)	(2) Primary Control (n = 4106)	(3) Primary + Secondary Control (n = 28,936)
Mean	1.00	1.00	1.00
Standard Deviation	0.27	0.08	0.04
Percentiles:			
Min	0.73	0.94	0.92
1%	0.76	0.94	0.93
5%	0.77	0.95	0.95
10%	0.78	0.95	0.96
25%	0.82	0.96	0.97
50% (Median)	0.86	0.99	1.00
75%	1.20	1.00	1.03
90%	1.26	1.02	1.05
95%	1.65	1.04	1.06
99%	1.69	1.39	1.07
Max	1.70	1.43	1.07

Notes: Balance weights are constructed to balance age, sex, and race/ethnicity between treatment and control groups, calibrated to the distribution of these variables in the overall study sample.

Table 13: First Stage Estimates

Panel A: Employee-level—All

	All		Surveyed		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 (%)	0.35	56.99	2.55	77.85	0.54	63.56
		(2.48)		(2.02)		(2.66)
		[0.00]		[0.00]		[0.00]
Modules completed (#)	0.01	2.39	0.10	3.83	0.02	2.97
		(0.11)		(0.17)		(0.16)
		[0.00]		[0.00]		[0.00]
3 or more modules (%)	0.21	39.90	2.01	63.27	0.36	49.11
		(1.97)		(2.80)		(2.93)
		[0.00]		[0.00]		[0.00]
Average total incentive payment ($)	0.39	103.69	3.22	168.51	0.73	169.13
		(5.46)		(8.17)		(10.26)
		[0.00]		[0.00]		[0.00]
N		32,973		2,168		7,631

Notes: 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 club 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 the combination of a weight for exposure to the wellness program (FTE weight) and a weight that balances treatment and control samples on demographics. Standard errors shown in parenthesis and p-values in brackets.

Panel B: Employee-level—Stably Employed Subsample

	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 (%)	0.42	61.73	2.72	81.39	0.55	64.36
		(2.61)		(2.12)		(2.80)
		[0.00]		[0.00]		[0.00]
Modules completed (#)	0.01	2.76	0.10	4.24	0.02	3.08
		(0.13)		(0.18)		(0.16)
		[0.00]		[0.00]		[0.00]
3 or more modules (%)	0.25	45.89	2.10	69.55	0.35	50.51
		(2.30)		(2.79)		(2.90)
		[0.00]		[0.00]		[0.00]
Average total incentive payment ($)	0.46	125.77	3.33	195.02	0.73	187.58
		(6.83)		(9.67)		(10.61)
		[0.00]		[0.00]		[0.00]
N		15,344		1,353		6,016

Notes: 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 club 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 the combination of a weight for exposure to the wellness program (FTE weight) and a weight that balances treatment and control samples on demographics. Standard errors shown in parentheses and p-values in brackets.

Table 14: Heterogeneity

	Absenteeism (%)		Total medical spending ($)		Regular exercise (%)		Considering losing weight (%)	
	N	Control Mean	N	Control Mean	N	Control Mean	N	Control Mean
Gender								
Female	13339	2.86	3005	4740.30	517	56.88	526	61.53
Male	15597	2.44	3621	3354.24	394	68.07	371	49.21
Age								
Below 40	18726	2.48	2691	2397.55	504	67.59	480	54.88
40 and above	10210	2.80	3935	5009.16	407	56.47	417	57.46

Notes: N is control group size only. Control means are weighted by the combination of a weight for exposure to the wellness program and a weight that balances treatment and control samples on demographics.

	SF-8 physical summary score		SF-8 mental summary score		Sweetened Drinks (No.)		BMI		Systolic BP (mmHg)	
	N	Control Mean	N	Control Mean	N	Control Mean	N	Control Mean	N	Control Mean
Gender										
Female	539	49.93	539	50.16	566	1.66	609	30.31	612	120.59
Male	408	51.87	408	52.46	427	2.07	461	28.94	462	128.99
Age										
Below 40	515	51.97	515	49.65	537	2.14	578	29.13	581	119.29
40 and above	432	49.71	432	52.57	456	1.58	492	30.23	493	128.82

Notes: N is control group size only. Control means are weighted by the combination of a weight for exposure to the wellness program and a weight that balances treatment and control samples on demographics.

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

Appendices

Appendix 1: Description of Phase 1 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. Employees only received a gift card for completing Take Charge one time.

Module 2

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

- Summary: This six-week program was presented as a series of webinar presentations or in paper form 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 who completed 5 out of 6 webinars and returned the verification form received a $50 BJ's gift card. Employees with Cigna coverage received an additional $150 Weight Management Reimbursement in the form of a BJ's gift card for completion.

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 club 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. Clubs were also in competition with the top club based on % participation and the top club with the highest average weekly minutes of exercise reported each receiving a trophy to display in the club, 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 returned the verification form received a $25 BJ's gift card.

Module 8

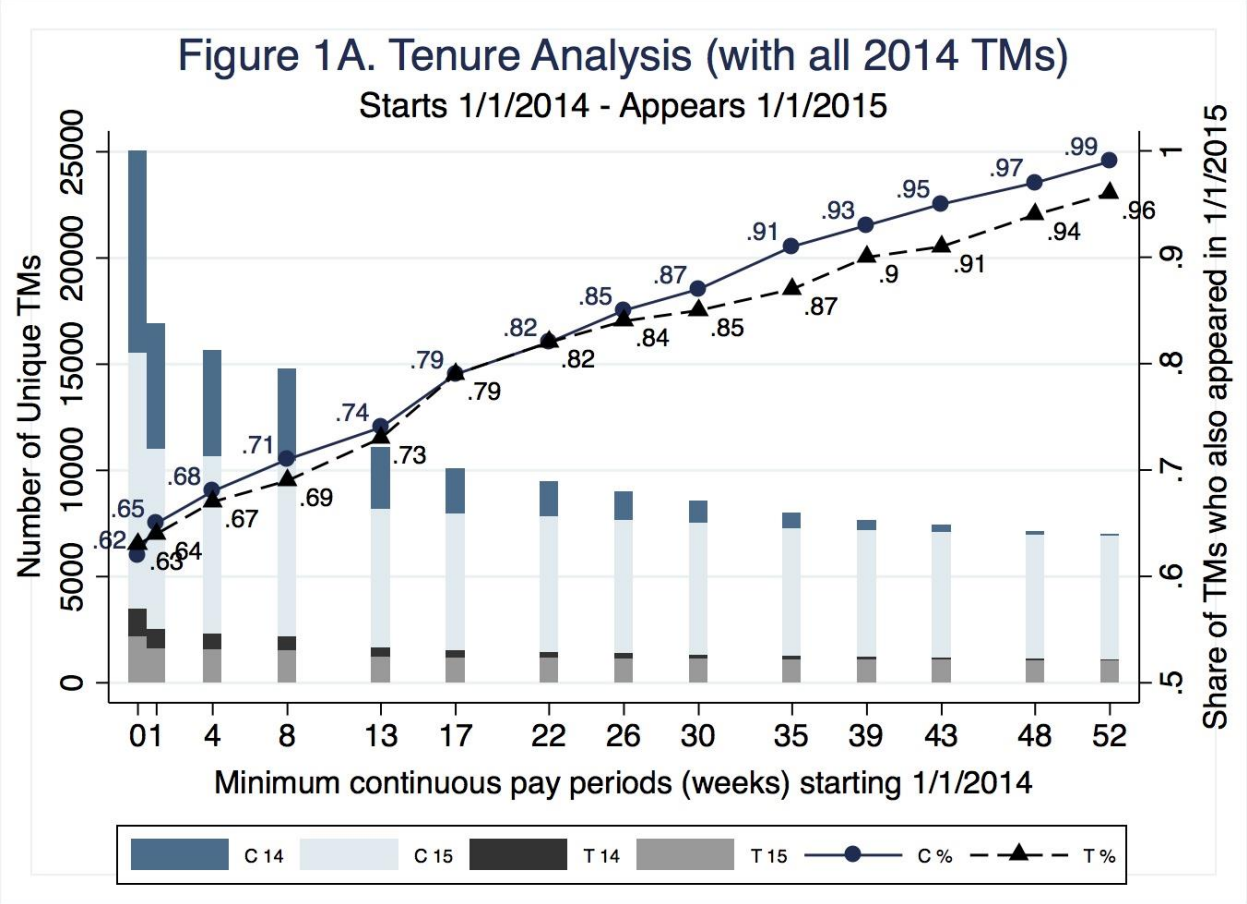
Movin' in May (5/1/2016-5/31/2016)

- 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 club.

Appendix 2. Determination of the Stably Employed Subsample

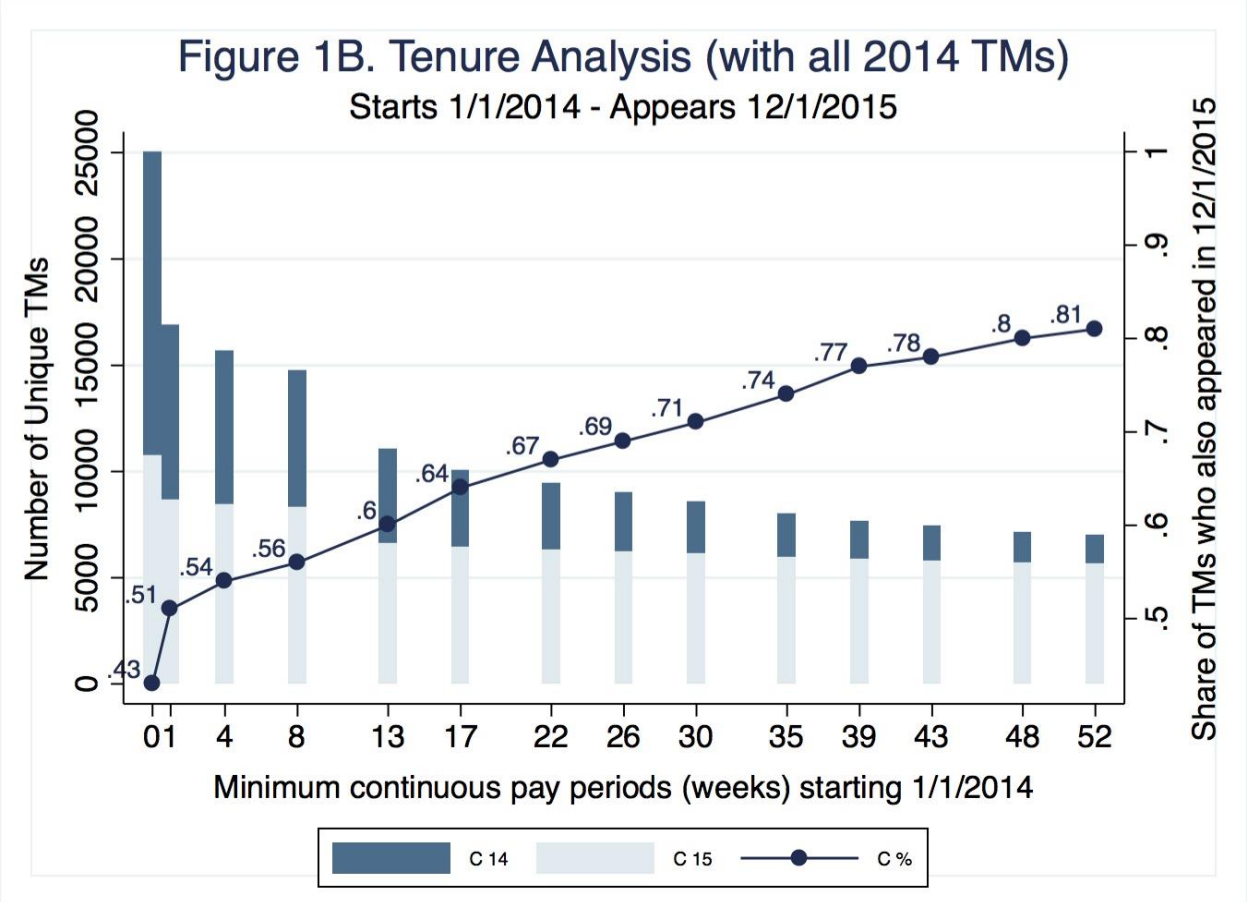
We conducted an analysis of the duration of employment (tenure), which informed our definition of the pre-specified stably employed subsample. In this analysis of tenure is described below, the nomenclature "TM" stands for "team members," the way that BJ's refers to employees.

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).



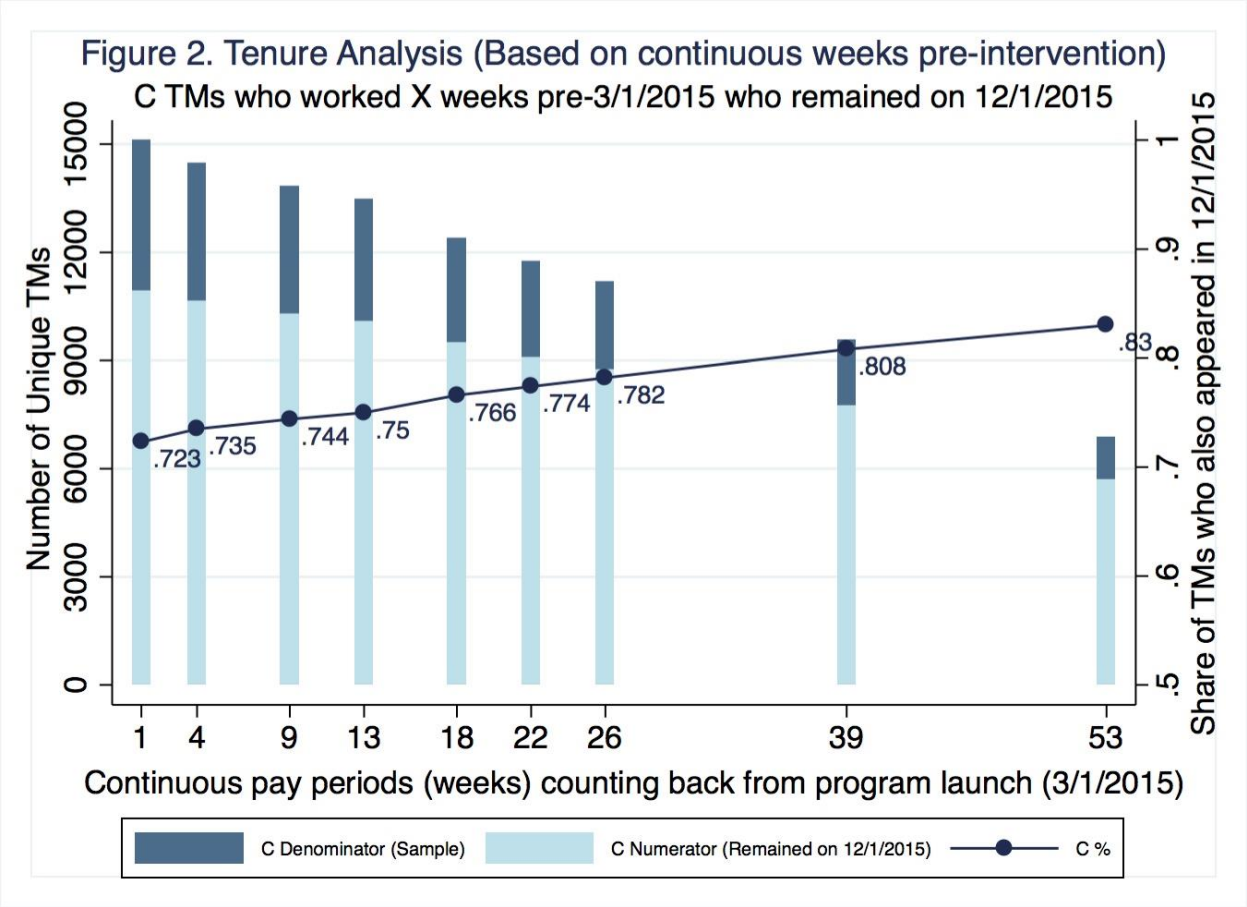
Thus, for example, the interpretation of the bar when X=17 is as follows. There were about 10,000 employees in control clubs who were employed at BJ's on 1/1/2014 and who worked through the first 17 weeks of 2014 (dark blue bar). Among these employees, about 79% (or about 7,900) were still working on 1/1/2015 (light blue bar). The same retention of 79% was found among employees in treatment clubs—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 club 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 clubs 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 clubs.

As an example of interpretation, there were about 13,500 employees in control clubs 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 clubs 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 throughout the rest of 2015. The X axis shows the months elapsed since start of the treatment (0 is the end of February 2015, while 10 is the end of December 2015). This graph shows a smooth decline in the sample of employees to reach 75% by the end of December 2015 (consistent with Figure 2).

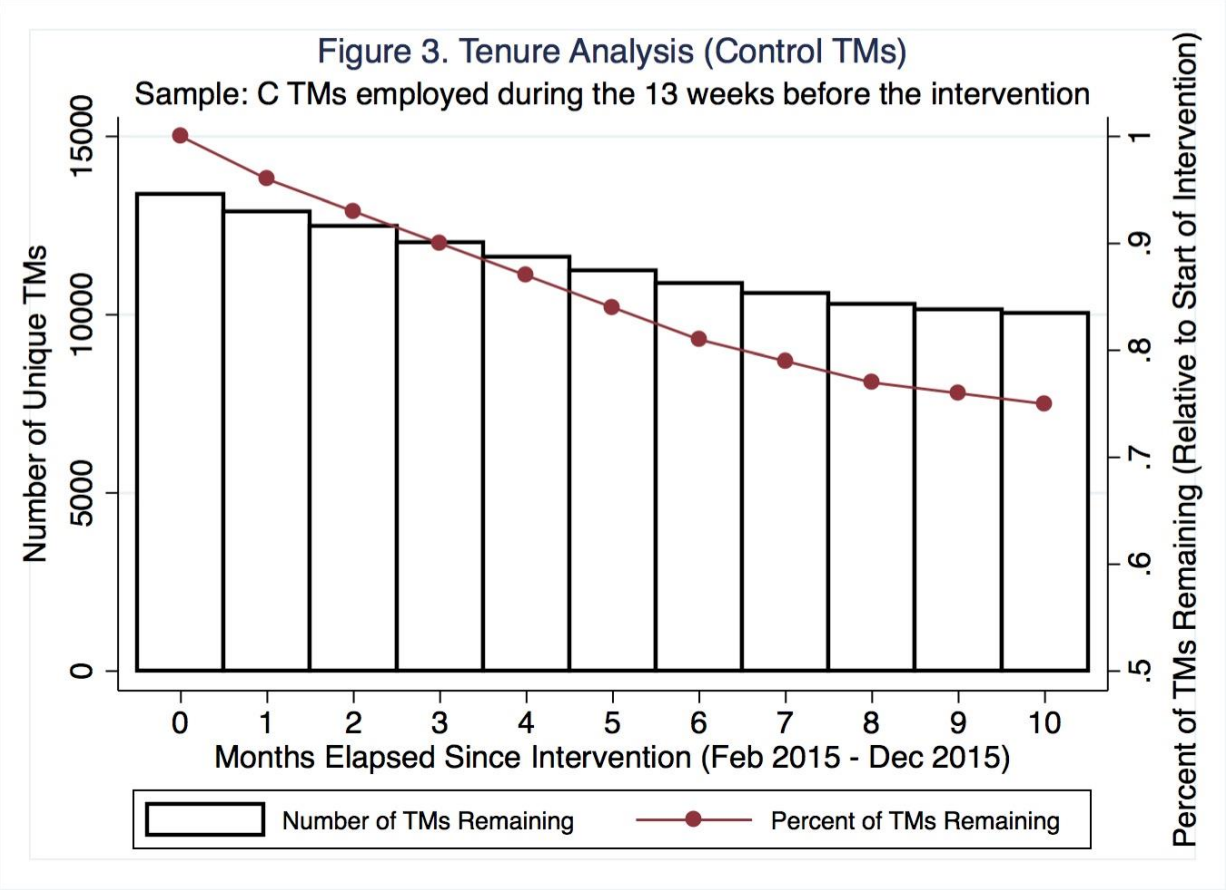


Figure 3. Tenure Analysis (Control TMs)
Sample: C TMs employed during the 13 weeks before the intervention
Number of Unique TMs
0
5000
10000
15000
Percent of TMs Remaining (Relative to Start of Intervention)
.5
.6
.7
.8
.9
1
0
1
2
3
4
5
6
7
8
9
10
Months Elapsed Since Intervention (Feb 2015 - Dec 2015)
Number of TMs Remaining
Percent of TMs Remaining

1218
1219