Received: 13 February 2018 | Revised: 24 September 2018 | Accepted: 17 October 2018
DOI: 10.1002/hec.3840

RESEARCH ARTICLE

WILEY Health Economics

The impact of financial incentives on health and health care: Evidence from a large wellness program

Liran Einav[1] | Stephanie Lee[2] | Jonathan Levin[3]

[1]Department of Economics, Stanford University, Stanford, California
[2]Foster School of Business, University of Washington, Seattle, Washington
[3]Department of Economics and Graduate School of Business, Stanford University, Stanford, California

Correspondence
Liran Einav, Department of Economics, Stanford University, Stanford, CA, USA.
Email: Leinav@stanford.edu

Abstract
Workplace wellness programs have become increasingly common in the United States, although there is not yet consensus regarding the ability of such programs to improve employees' health and reduce health care costs. In this paper, we study a program offered by a large U.S. employer that provides substantial financial incentives directly tied to employees' health. The program has a high participation rate among eligible employees, around 80%, and we analyze the data on the first 4 years of the program, linked to health care claims. We document robust improvements in employee health and a correlation between certain health improvements and reductions in health care cost. Despite the latter association, we cannot find direct evidence causally linking program participation to reduced health care costs, although it seems plausible that such a relationship will arise over longer horizons.

KEYWORDS
health behavior, health care cost, wellness programs

1 | INTRODUCTION

Workplace wellness programs have become an important tool for employers, insurers, and policymakers to combat rising health care costs. These programs stem from the idea that encouraging working individuals to adopt a healthier lifestyle is a "win-win" strategy. Wellness programs have the potential to lead to better health for the individuals involved, translate to improved labor productivity, and reduce health care costs (Alderman, 2009; Baicker, Cutler, & Song, 2010).

Although some form of wellness program is offered by the majority of large employers in the United States (see Mattke et al., 2013, for a comprehensive review), there is not yet clear consensus as to the impact of such programs or as to which program design features are most effective.[1] Although the potential for cost saving appears large (Bolnick, Millard, & Dugas, 2013) and some studies suggest that wellness programs are associated with cost savings (Baicker et al., 2010; Ozminkowski, Ling, Goetzel, Bruno, Rutter, Isaac, & Wang, 2002), others do not find a significant effect on cost (Aldana, Merrill, Price, Hardy, & Hager, 2005; Caloyeras, Liu, Exum, Broderick, & Mattke, 2014; Jones, Molitor, & Reif, 2018) or worry that such programs shift costs onto less healthy employees (Horwitz, Kelly, & DiNardo, 2013). On March 23, 2010,

[1]Seventy-one percent of large firms with 200 or more employees with health benefits offer programs to help employees stop smoking, 68% offer lifestyle or behavioral coaching, and 61% offer programs to help employees lose weight. Eighty-one percent of large firms report offering at least one of these three programs (Claxton, Rae, Panchal, Whitmore, Damico, Kenward, & Long, 2015).

President Obama endorsed a provision created to encourage employers to implement more substantial standard-based wellness programs.

In this paper, we add to the existing evidence by exploiting rich administrative data from a large U.S. employer that has been a pioneer in designing and implementing a workplace wellness program. The employer introduced its wellness program in late 2008. The program is innovative in its design and offers large financial incentives. A central feature of the program is that financial incentives are directly linked to employees' health. Through its workplace wellness program, the employer provides its employees the strong financial incentives to pass specific health standards.

We describe the employer's workplace wellness program and the financial incentives, as well as our data, in Section 2. The program tracks five common health measures once a year—body mass index (BMI), blood pressure, cholesterol, glucose, and nicotine. Our data contain administrative records for all eligible employees (and their program eligible dependents), as well as annual health measures for all program participants, and health insurance claim-level data for most individuals. We use data from the first 4 years of the program (2009–2012), covering more than 20,000 participants per year. Participation rates among eligible individuals are higher than in most other programs. Perhaps the most unique feature of the program is its incentive structure. Although most employers offer participation payments for employees, the employer we examine ties financial incentives directly to biometric measurements, offering up to $1,000 per year for individuals who do well on all metrics. The financial rewards (paid in the form of insurance premium discounts) can be eliminated entirely for individuals who fail all screenings.[2]

In Section 3, we use a variety of empirical approaches to document a robust year-to-year improvement in each measured biometric for program participants. This complements findings reported in Fu, Bradley, Viswanathan, Chan, and Stampfer (2016) for the same employer.[3] There are two empirical challenges in attempting to attribute causal interpretation to the findings described above. First, participation in the program is voluntary, although internal advertising, benefits sessions, and the financial incentives lead to relatively high participation rates of 75–80% among eligible individuals. Second, health measures are only available for program participants during their participating years so cannot be assessed against nonparticipants. Although there is no perfect way to circumvent these challenges, we use external data to provide some benchmarks for comparison, which generally supports a causal interpretation.

An important part of the analysis in Section 3 is our ability to observe participating individuals for up to 4 years. This means that not only we can assess the binary and immediate impact of program participation, which was the focus of all the studies mentioned earlier but also we can focus on the possibility that the impact of the program may not be immediate and could be linked to consistent, year-after-year participation that may inhibit health-related habits for instance (Charness & Gneezy, 2009; Mochon, Schwartz, Maroba, Patel, & Ariely, 2016). In one specification, we compare the passing rate of program participants between those new participants and individuals who have been participating in the program for several years, and we find that longer program participation is associated with significant and meaningful improvements in screening results, especially for men.

Section 4 relates these health improvements to health care cost and utilization. With a simple difference-in-differences specification, we do not find clear evidence that program participation is associated with lower health care cost, at least over the first 4 years of the program, and this finding is consistent with the results of Jones et al. (2018), who find similar results in a randomized control trial in a similar context. In fact, we find some evidence for higher health care cost over the first 2 years. However, we do present evidence that health improvements in terms of BMI and blood pressure (but not in terms of cholesterol and glucose) are associated with lower health care utilization and cost, as well as empirical evidence that the program triggers individuals to start treating (or adhere better) to their blood pressure and cholesterol-reducing drugs.

In the final section, we summarize our findings and conclude. Overall, although we do not find clear evidence that health care cost has declined as a result of the wellness program, our overall conclusion is quite positive. Participating employees appear healthier, and this will likely make them more productive and perhaps even cheaper in the longer run, regardless of the shorter run cost-benefit analysis. Of course, our analysis and statistical inference rely on data from a single program implemented in a single firm, and further studies with similar programs and other firms are needed to assess external validity.

[2]It is common for employer wellness programs to pay financial rewards in the form of premium discounts, as it is the case here. An interesting question is whether the effect of financial incentives would be different if financial rewards were paid in cash or in other forms.

[3]A related finding is reported by Cawley and Price (2013), who document a response by employees to financial incentives associated with weight loss.

2 | SETTING AND DATA

2.1 | The employer's workplace wellness program

The employer's workplace wellness program in its current form was first implemented for the calendar year 2009 and has continued throughout our sample period with small changes from year to year. All nonunion employees are eligible for the program, as well as their spouses if they are covered (as dependents) by the employer's employer-provided health insurance. Participation in the program is voluntary and has been around 75–80% among all eligible individuals.

Program participation requires individuals to take an annual confidential health screening and have its results reported to the program administration.[4] The screening can either be taken in a doctor's office, with the results transmitted to the company, or more commonly, the program organizes and advertises prescheduled on-site events in which individuals could participate in the health screening session in their job location. Employees are informed about the wellness program through the company intranet and home mailer. Furthermore, human resources teams across the company are trained to ensure that program information is accurately and effectively shared with employees. Employees have access to an online benefits portal where screening details, scheduling dates, and deadlines are accessible.

The health screening session takes about 15–30 min and involves measuring five distinct health metrics and a subsequent optional consultation with a health professional. The five health metrics are BMI, blood pressure, cholesterol, glucose, and smoking. Each is associated with a pass/fail outcome. Passing standards are based on the standard health guidelines, with some leeway relative to the National Institutes of Health (NIH) recommendations.

BMI is the ratio of the individual's weight (in kilograms) to the square of her height (in meters), with a BMI of 30 or below considered a passing result.[5] Blood pressure is measured using the systolic and diastolic millimeter of mercury readings, with passing result requiring that the reading is both below 140 (systolic) and below 90 (diastolic). Cholesterol is measured by total cholesterol, with values below 220 mg/dl considered a passing result.[6] To pass the glucose test (which started in 2010), glucose must be below 116 mg/dl, and to check for smoking, individuals had to obtain a negative result on a nicotine test.[7]

To the extent that financial incentives exist, the typical wellness programs provide such incentives by rewarding program participation. One of the unique features of the workplace wellness program at the employer we study is that the financial incentives to participants are large and are directly linked to successful test results. Each passing result on a biometric comes with a financial reward: a 3.00–5.50 dollar premium reduction in the *weekly* health insurance premium. These incentives can add up across five tests and a full calendar year to approximately 1,000 dollars per individual (or 2,000 dollars per household with two participating adults). Moreover, the program provides even stronger incentives to individuals who are less healthy. If an individual improves her health and meets the passing standard 1 year after having missed it in the prior year, she can receive a rebate for the measure retrospectively, so in such situations, financial incentives are doubled. Appendix A provides complete details of the program rules.

2.2 | Data

Our data include annual information on all the eligible employees and spouses over 5 years (2008–2012), starting in the year before the start of the wellness program and covering the first 4 years in which the program has been in place. In addition to the health measures that are available for all program participants, we also obtained administrative data on the employees' (and dependents) health insurance and pharmaceutical drug insurance claims for employees who are enrolled in the employer's preferred provider organization (PPO) health insurance plan.

Table 1 presents summary statistics. Our initial, full sample—summarized in column 1—consists of all individuals who were eligible for the workplace wellness program, that is, all nonunion employees and their spouses who were covered by the employer's employer-provided health insurance. This sample consists of 115,805 individual years that represent 41,590 unique individuals, of which 30,724 are employees and the rest are covered spouses. Forty-six percent of the observations are male, the average age is 45, and the majority have been working at the company for more than 10 years, with only

[4] Fifty percent of large firms with 200 or more employees with health benefits offer or require employees to complete an annual biometric screening (Claxton et al., 2015).

[5] Pregnant women automatically pass. Individuals with high BMI results could also pass the BMI metric by showing a 10% reduction in BMI in the subsequent year, even if the BMI is still above 30. Alternatively, individuals whose weight is high due to muscles buildup can elect to pass the BMI metric with waist circumference measure that is less than 40 (for males) or 35 (for females) in.

[6] This is effective on 2011 and later. Cholesterol measurement did not take place in 2009, and in 2010, the requirements for a passing result were more stringent (high-density lipoprotein has to be greater than 40 mg/dl, low-density lipoprotein to be less than 130 mg/dl, and triglyceride less than 200 mg/dl).

[7] The nicotine test result is binary, and the passing rate for those who take it is close to 100%, so we do not use this test throughout the analysis.

TABLE 1 Summary statistics

	Full sample (1)	Health only (2)	Utilization only (3)	Complete data (4)
Observations				
Individual-years	115,805	85,673	91,153	61,021
Unique individuals	41,590	34,212	32,939	24,640
Unique employees	30,724	26,059	24,030	18,667
Demographic information[a]				
Male	0.46	0.45	0.46	0.45
Age[b] (average)	45.17	45.39	45.86	46.44
Age,[b] (standard deviation)	11.80	11.80	11.53	11.23
Salary (average, in $000)	58.43	61.95	58.87	61.95
Salary (standard deviation)	37.32	38.20	37.14	38.20
Year of hire, average	1997.6	1996.8	1997.3	1996.8
Share hired in 2009 or later	0.078	0.056	0.064	0.056
State of employment				
California	0.23	0.23	0.19	0.19
Texas	0.17	0.17	0.20	0.20
Oregon	0.16	0.16	0.16	0.16
Other	0.44	0.44	0.45	0.45
Health expenditure				
Annual health spending			5,117	4,865
Standard deviation			19,668	17,911
Percent zero			0.17	0.14
Annual drug (RX) spending			789	947
Standard deviation			3,159	3,079
Percent zero			0.26	0.21
Healthy measures passing rates				
BMI		0.76		0.75
BP		0.85		0.85
Cholesterol (2010–2012)		0.76		0.76
Glucose (2010–2012)		0.86		0.86

Note. BMI: body mass index; BP: blood pressure. Observations are from 2009 to 2012. “Health only” sample includes all individuals who participated in the healthy measures program that year. ”Utilization only” sample includes all individuals enrolled in one of the safeway preferred provider organization options that year. “Complete data” sample includes individual years who satisfy both criteria above.
[a]Age and gender information is available for everyone. Other information is missing for some individuals. Salary and hire information is based on 82,116, 52,788, 75,712, and 52,788 in each sample, respectively. Location (state) information is based on 68,534, 68,534, 48,854, and 48,854 in each sample, respectively. [b]There are 121 individuals whose age is 65+, but the exact age is unknown. We use 70 for these individuals to calculate the summary statistics.

7.8% who were hired in 2009 or after. The employer has a national presence and operates across the country, but more than half of the sample contain individuals in California, Texas, and Oregon.

This initial sample has various elements of missing data for two primary reasons. First, health information is missing for program nonparticipants as their health measures are not being recorded. These account for 26% of the individual years in the sample. Second, eligible individuals may opt to enroll in a health insurance plan that is different from the employer's PPO coverage. They are covered by Kaiser Permanente, which is a vertically integrated health care provider. It receives capitated payments from the employer and thus does not report back to the employer claim-level utilization information. These individuals account for 21% of the individual years in the original sample.

We therefore end up working with different forms of samples, based on the specific analyses. The “health only” sample—summarized in column 2 of Table 1—includes all participating individuals for whom we have health measures, the “utilization only” sample—summarized in column 3—includes all PPO enrollees for whom we have utilization data, and the “complete data” sample (in column 4) is the intersection of the two. As Table 1 shows, the samples are reasonably similar in terms of their observables (gender, age, salary, tenure at the company, and job location). However, selection on unobservables (into the wellness program or into Kaiser coverage) is an important concern, and we will address it below using the different research designs and identification strategies.

3 | HEALTH CHANGES OVER TIME

3.1 | Health trends for program participants

In this section, we use the health only sample to explore how health changes for program participants. We start by looking at all individuals whose health is observed over consecutive years. That is, we drop individuals who are only in the sample for a single year or for nonconsecutive years, and we treat each pair of consecutive years as a separate observation. Thus, for example, individuals who are observed for all 4 years are used to generate three different observations.

Table 2 shows the pass fail transition matrix for all consecutive year observations. As expected, health is highly correlated over time within an individual. Individuals who pass a given test in 1 year are much more likely to pass it in the subsequent year, with passing rates ranging from 91 to 94% conditional on passing the same test in the previous year, whereas passing rates are much lower—18% for BMI and 45–65 for the other metrics—for those individuals who did not pass in the previous year.

Table 2 additionally points to likely improvements in health metrics over time, as the transition matrix is not symmetric: The rate by which individuals' transition from pass to fail is significantly lower than the rate by which they transitioned from fail to pass. To see this more granularly, Figure 1 presents the relationship between the health measure in a given year against the corresponding health measure of the same individual in the previous year. Across all measures, one can observe a clear pattern: Individuals with high (worse) health metrics tend to have lower ones the year after, and individuals with low health metric tend to have higher ones. It is important to note that high measures are bad, but conditional on being low enough, lower measures are not necessarily better. In this sense, Figure 1 shows a clear pattern of improvements in health. To see this, note that we also plot the "passing threshold" in each panel of the figure, and in all panels, the pattern crosses the 45-degree line below the passing threshold (and often even below the more stringent NIH recommendation), implying that individuals whose health metrics are too high tend to improve on average.

The pattern described above suggests that individuals who participate in the program tend to get healthier, at least on average. Of course, this pattern does not mean necessarily that this improved health is caused by the program participation. The improvement in health may be viewed as even less trivial once one realizes that aging alone should make many health metrics deteriorate from year to year, rather than improve. Medical literature has found that blood pressure rises with age, a phenomenon which is associated with structural changes in the arteries (Pinto, 2007). Cholesterol and glucose levels also tend to rise with age (Kannel, 1987; O'Sullivan, 1974). An important concern, however, is that the pattern observed in Figure 1 merely reflects mean reversion, either in true health or in the measurement of health. For example, a similar qualitative pattern could be generated if health is measured with error or if one's day-to-day health fluctuates (so a measurement in a particular day is a noisy signal of one's average daily health over the year).

To explore the potential importance of mean reversion and to more generally compare the pattern observed by the employer we study to some other benchmarks, we have searched for other data sets that follow individuals' health

TABLE 2 Year-to-year passing rates

BMI	Pass in year t	Fail in year t	Blood pressure	Pass in year t	Fail in year t
Pass in year $t-1$	35,360	2,236	Pass in year $t-1$	37,972	3,849
	0.94	0.06		0.91	0.09
Fail in year $t-1$	1,983	8,777	Fail in year $t-1$	4,185	2,350
	0.18	0.82		0.64	0.36
Cholesterol	**Pass in year t**	**Fail in year t**	**Glucose**	**Pass in year t**	**Fail in year t**
Pass in year $t-1$	19,229	1,927	Pass in year $t-1$	21,403	1,229
(actual standards)	0.91	0.09	(actual standards)	0.95	0.05
Fail in year $t-1$	6,141	4,287	Fail in year $t-1$	6,739	2,213
(actual standards)	0.59	0.41	(actual standards)	0.75	0.25
Pass in year $t-1$	22,034	2,317	Pass in year $t-1$	26,152	1,571
(2011 standards)	0.90	0.10	(2011 standards)	0.94	0.06
Fail in year $t-1$	3,336	3,897	Fail in year $t-1$	1,990	1,871
(2011 standards)	0.46	0.54	(2011 standards)	0.52	0.48

Note. Table presents observation counts in each cell (shares below the counts). An observation is an individual observed over two consecutive years. An individual who is observed over three or four consecutive years is included as two or three, respectively, multiple observations. Cholesterol and glucose measurement started in 2010 only, and passing standards were relaxed in 2011, so we report above passing rates based on either actual standard or the relaxed 2011 standards.

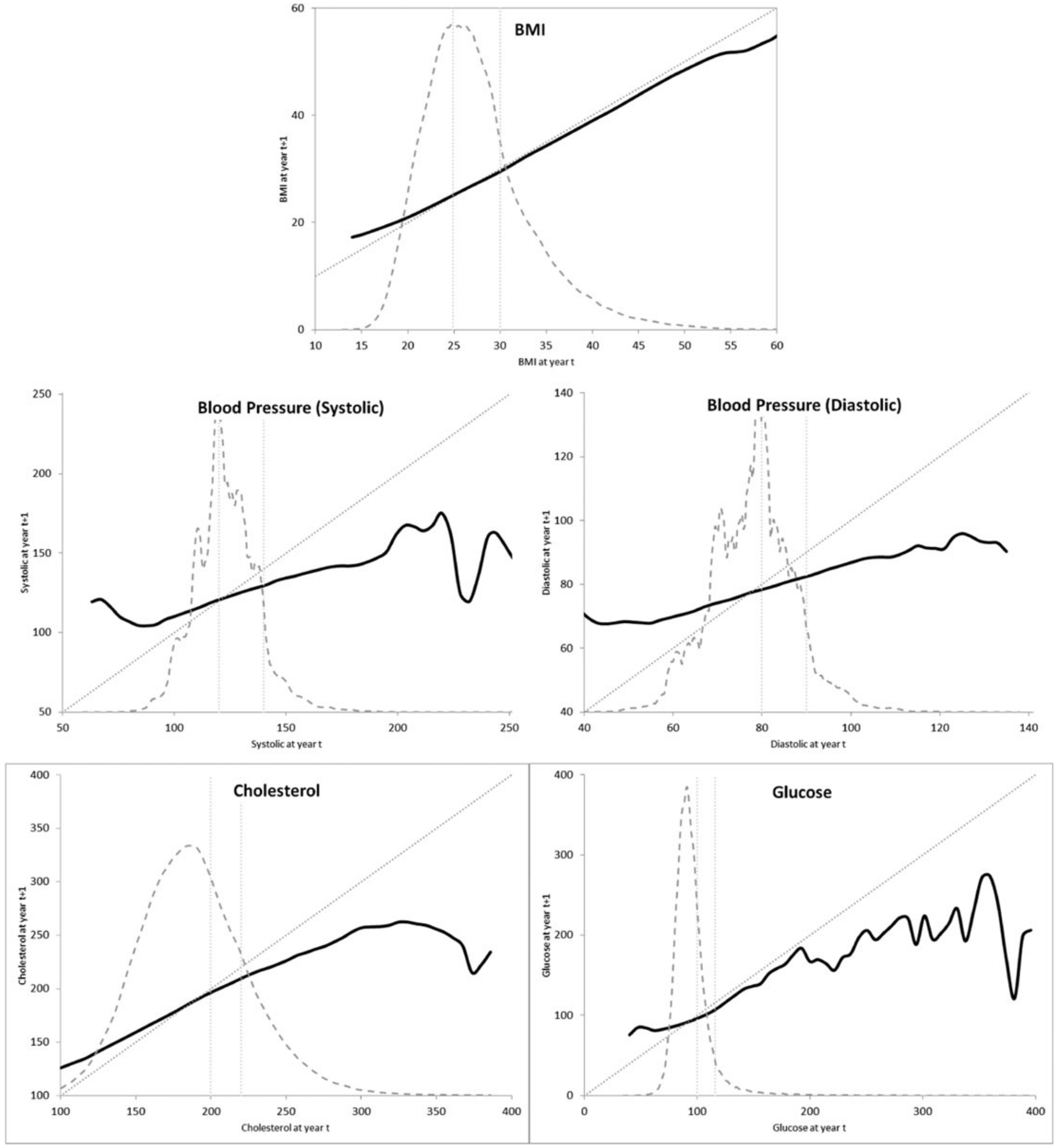

FIGURE 1 The timing of preventive drug purchasing and screening results. Figure plots (thick black lines) predictions from kernel regression of a health indicator in 1 year on the same health indicator the year before, using all available observations on the same person in consecutive years. Each panel presents one type of test. All panels show the 45-degree line, two vertical lines at the employer passing threshold and the (lower) National Institutes of Health recommended level and the underlying distribution of year t-1 values (dashed gray)

over time and found two publicly available data sets that may serve as an imperfect benchmark. One is based on the Framingham Offspring Study, and the other is based on the Coronary Artery Risk Development in Young Adults (CARDIA) study. The Framingham and CARDIA data sets are longitudinal studies designed to examine factors that influence the development of cardiovascular disease. Neither data set follows individuals annually but over a longer time interval. The Framingham study has completed nine clinical examinations with intervals of 4–6 years between consecutive measures, whereas the CARDIA study has completed seven clinical examinations with intervals of 2–5 years between consecutive measurements. Appendix B provides more details about these data sets.

Figure 2 reports changes in BMI for Framingham, CARDIA, and workplace wellness program participants. For Framingham Offspring, changes in BMI are calculated as BMI from clinical examinations held in 1991–1995 minus BMI from clinical examinations held in 1987–1991. For CARDIA, changes in BMI are calculated as BMI from clinical examinations held in 2000–2001 minus BMI from clinical examinations held in 1995–1996. For workplace wellness program participants, changes in BMI are calculated as BMI from program year 2012 minus BMI from program year 2009. Although the populations are not fully comparable and the gap between measurement for the employer is slightly shorter, the pattern that emerges from Figure 2 is quite clear. It shows that the workplace wellness program participants are more likely

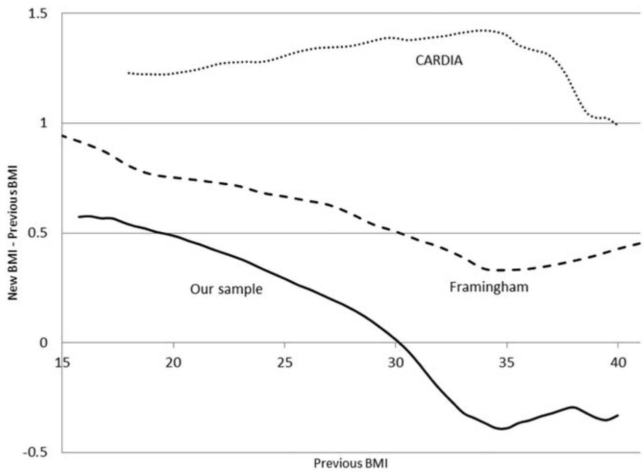

FIGURE 2 Comparison of workplace wellness participants to other populations. Figure shows the estimated change in body mass index (BMI) from one examination to the next examination, estimated using a local polynomial regression plotted against BMI from the previous examination. For Coronary Artery Risk Development in Young Adults (CARDIA), previous BMI indicates BMI from examinations held in 1995–1996, and new BMI indicates BMI from examinations held in 2000–2001. For Framingham, previous BMI indicates BMI from examinations held in 19870–1991, and new BMI indicates BMI from examinations held in 1991–1995. For workplace wellness program participants, previous BMI indicates BMI from examinations held in program year 2009, and new BMI indicates BMI from examinations held in program year 2012. See Appendix B for more details on the CARDIA and Framingham data sets

to reduce their BMI than Framingham Offspring participants or CARDIA participants. CARDIA participants generally increase their BMI by 1–1.5 between two examinations regardless of their BMI from previous examination, presumably due to aging. The Framingham data show some convergence (lower changes to BMI for individuals whose previous measures were higher) but still show (as in the CARDIA data) that BMI on average increases regardless of the previous BMI measure. The employer data show smaller changes overall and, in particular, shows that when BMI from a previous examination (3 years earlier) is greater than 30, the workplace wellness participants tend to reduce weight, whereas Framingham Offspring participants and CARDIA participants gain weight. In Appendix Table D1, we show that these patterns are robust to flexibly controlling for the age and gender of individuals.

3.2 | Differential health trends by participation duration

An important limitation in our analysis is that we only observe health measures for individuals who participate in the program. Therefore, even if program participation was exogenous—and it is unlikely to be the case—we do not have a way to compare health outcomes to a control group of nonparticipants, as for these individuals health measures are not observed. As a way to come closer to a comparable control group, in this section, we compare pass rates of individuals who participated in the program for 4 years to pass rates of those who newly became eligible and participated in the program for the first time in 2012, the last year in our data. The idea behind the empirical strategy is that employees who were eligible for 4 years and employees who newly became eligible are comparable—both represent individuals who elected to participate in the program—except that the former group was exposed to the program for a significantly longer period (4 years relative to only one). Furthermore, we can try to control for employee's decision to participate by looking at employees who have always participated to newly eligible employees who participated, so the comparison is across individuals who always participated in the program when they were eligible.

Table 3 cross tabulates the number of participation years against the number of eligibility years for individuals in our "complete data" sample. Our strategy is to compare health metrics in the "e1p1" cell—that is, individuals who were eligible only in 2012 and participated in the program in that year—to "e4p4" individuals who were eligible to participate in all 4 years (2009–2012) and participated in all years. There are 1,347 in the former group and 6,989 in the latter, and as shown in Table 3, it is notable that the full participation rates are similar: The 1,347 e1p1 individuals constitute 56% of the individuals who were eligible, and the e4p4 individuals constitute 53% of the individuals who were eligible in all years, perhaps suggesting that selection into participation in both groups is similar.

We report two exercises that take advantage of this variation. First, we compare pass rates for each health metric across the two groups, after controlling for year and age fixed effects. That is, we estimate the following regression:

$$\text{Pass}_{it}^{m} = \mathbf{x}_{it}'\gamma + \delta_t + \beta \cdot D_{e4p4} + \varepsilon_{it}, \tag{1}$$

TABLE 3 Participation rates for eligible individuals

Number of years participated		Number of years eligible: 1	2	3	4	Total
	0	1,038	649	1,012	1,668	4,367
		0.44	0.33	0.24	0.13	0.20
	1	1,347	420	478	1,241	3,486
		0.56	0.22	0.11	0.09	0.16
	2		877	619	1,291	2,787
			0.45	0.15	0.10	0.13
	3			2,157	1,978	4,135
				0.51	0.15	0.19
	4				6,989	6,989
					0.53	0.32
	Total	2,385	1,946	4,266	13,167	21,764
		1.00	1.00	1.00	1.00	1.00

Note. Table presents the number of individuals in each cell (shares below the counts). Sample restricted to individuals in the "complete data" sample who were eligible for participation in 2012.

TABLE 4a Comparison across short-run and long-run program exposure

	Dependent variable: Pass in …			
	BMI	**BP**	**Cholesterol**	**Glucose**
	(1)	**(2)**	**(3)**	**(4)**
Males				
4-year participation	0.051	0.038	0.111	0.073
	(0.018)	(0.017)	(0.020)	(0.018)
Mean of dependent variable	0.80	0.84	0.70	0.74
R^2	0.008	0.016	0.136	0.1939
N (individual years)	13,253	13,253	10,091	10,091
N (unique individuals)	3,767	3,767	3,767	3,767
Females				
4-year participation	0.042	0.046	0.034	0.0467
	(0.017)	(0.012)	(0.017)	(0.015)
Mean of dependent variable	0.78	0.91	0.76	0.83
R^2	0.006	0.026	0.057	0.1299
N (individual years)	16,050	16,050	12,223	12,223
N (unique individuals)	4,569	4,569	4,569	4,569

Note. Sample restricted to the 1,347 individuals who were eligible in 2012 and participated in the program that year and the 6,989 individuals who were eligible for all 4 years (2009–2012) and participated in all for 4 years. Cholesterol and glucose measurement sarted in 2010 only. In adddition to the reported coefficient, each regression includes year fixed effects and individual age fixed effects. Standard errors are in parentheses.

where δ_t indicates year fixed effects, $\mathbf{x}_{it}$ is a vector of age fixed effects, and D_{e4p4} equals 1 if an individual was part of the group that was eligible for 4 years and participated for 4 years and equals 0 if an individual was eligible for 1 year and participated for 1 year. We estimate this regression separately for each health metric and separately for men and women. The results are reported in Table 4a and suggest that those who participated for 4 years are more likely to pass in 2012 than those who newly became eligible in 2012. The quantitative effects are quite large. Pass rates are approximately 5% higher for women across all metrics, and the differences are even larger among men for some of the metrics (e.g., 15% higher pass rates for cholesterol among men who participated all years).[8]

One concern about this analysis is that as we emphasize throughout, the decision to participate in the program is endogenous, so individuals who participated in the program in all 4 years of their eligibility (the "e4p4" group) may be "more selected" than individuals who participated in the program in their single year of eligibility (the "e1p1" group). We therefore report in Table 4b results from a specification that attempts to address this concern. Specifically, we restrict

[8]We additionally compare health metrics in the "e2p2" cell to those in the "e4p4" cell and find that pass rates are higher for those who were exposed to the program for a longer period.

TABLE 4b Using years of eligibility as an instrument for program participation

	Dependent variable: Pass during 2012 in …			
	BMI (1)	BP (2)	Cholesterol (3)	Glucose (4)
Males				
Years of eligibility (OLS)	0.0093	0.0088	0.0266	0.0158
	(0.0057)	(0.0050)	(0.0055)	(0.0048)
Years of participation (OLS)	0.0421	0.0337	0.0497	0.0388
	(0.0049)	(0.0044)	(0.0048)	(0.0042)
	0.0107	0.0101	0.0308	0.0183
	(0.0065)	(0.0057)	(0.0063)	(0.0055)
Mean of dependent variable	0.76	0.82	0.78	0.84
N (individuals)	6,778	6,778	6,778	6,778
Females				
Years of eligibility (OLS)	0.0005	0.0103	0.0034	0.0055
	(0.0054)	(0.0038)	(0.0047)	(0.0036)
Years of participation (OLS)	0.0471	0.0288	0.0248	0.0251
	(0.0047)	(0.0033)	(0.0042)	(0.0032)
Years of participation (IV)	0.0006	0.0120	0.00394	0.0064
	(0.0062)	(0.0044)	(0.00547)	(0.0042)
Mean of dependent variable	0.72	0.89	0.80	0.90
N (individuals)	8,287	8,287	8,287	8,287

Note. OLS: ordinary least square. Sample restricted to 2012 observations of the 15,065 individuals who were eligible in 2012 and participated in the program that year. Dependent variable is pass in 2012 for each metric. The variable years of eligibility measures the number of years of eligibility. The variable years of participation measures the number of years of participation. The years of participation (IV) row reports coefficient of an IV regression where we instrument with years of eligibility for years of participation (first stage coefficients are 0.865 [0.009] and 0.841 [0.011] for males and females, respectively). In addition to the reported coefficient, each regression includes individual age fixed effects. Standard errors are in parentheses.

attention to only program participants in the last year of the program (2012) and run a similar regression but replace the "e4p4" indicator variable with the number of years of program participation and instrument for it with the number of years of program eligibility for the same individual (which also takes values that range from 1 to 4). That is, we estimate the following regression:

$$\text{Pass}^{m}_{i,2012} = \mathbf{x}'_{i,2012}\gamma + \beta \cdot P_{i,2012} + \varepsilon_i, \tag{2}$$

where $\mathbf{x}_{i,2012}$ is a vector of age fixed effects, and $P_{i,2012}$ is the number of years of program participation as of 2012 (which takes values that range from 1 to 4), and we then instrument for $P_{i,2012}$ with $E_{i,2012}$, which is the number of years of program eligibility for the same individual (which also takes values that range from 1 to 4, with $E_{i,2012} \geq P_{i,2012}$). As before, we estimate this regression separately for each health metric and separately for men and women. The results are reported in Table 4b, which also reports results from the ordinary least sqaure (OLS) regression (without using an instrument) and from the "reduced form" regression, where $E_{i,2012}$ is used instead of $P_{i,2012}$ as the explanatory variable. Consistent with the fact that program participation is endogenous, the OLS coefficients are much greater than the IV coefficients. However, the IV coefficients are not trivial and are statistically significant in most cases, especially for men. For example, the estimated coefficients suggest that for men, one extra year of program participation would increase the passing rate of the cholesterol test by approximately three percentage points, would increase the passing rate of the glucose test by about two percentage points, and would increase the BMI and blood pressure passing rates by one percentage point. The effects for women are smaller (with the exception of blood pressure).

3.3 | Responses to changes in the magnitude of financial incentives

In an alternative approach, we use the variation in incentive amounts to examine how individuals respond to financial incentives. Prior literature has shown that incentives can encourage the development of good health habits (Charness & Gneezy, 2009; Mochon et al., 2016). The hypothesis is that people will respond more to greater incentives, and if they do, it seems likely that the program itself, which relies on financial incentives, improves health. In this analysis, we focus only on BMI and blood pressure because both these metrics were introduced in the first year of its workplace wellness program

(2009) and testing standards remained unchanged throughout the observation period. Appendix Table A2 reports the year-to-year changes in financial incentives. The incentives depend on which PPO plan the individual was enrolled in, which we unfortunately do not observe. For both plans, the weekly incentive amounts have changed from year to year, both up and down, for administrative reasons that are unlikely to be associated with underlying health. Although the exact incentive amounts differ across the two plans, incentive amounts have changed similarly for both plans. To take advantage of this variation, we use the incentives associated with the higher coverage PPO plan ("Choice Fund I"), which covered more individuals. For individuals enrolled in this plan, passing BMI implied a benefit of 6 dollars per week in 2009 and 2010 but only 4 dollars in 2011 and 5.50 dollars in 2011. Similarly, passing blood pressure implied a benefit of only 1.50 dollars in 2009, 4 dollars in 2010 and 2011, and 3 dollars in 2012.

Using the variation in incentive amounts, we estimate the following:

$$Pass_{imt} = \alpha_i + \beta \cdot WI_{mt} + \theta_m + \delta_t + \varepsilon_{imt}, \tag{3}$$

where the dependent variable is equal to one if an individual i passed health measure m in program year t. θ_m and δ_t are, respectively, health measure and year fixed effects, and α_i represents individual fixed effects. WI_{imt} is the weekly incentive amount associated with passing health measure m in year t, and β is the main coefficient of interest. We estimate this regression using all individual-year observations in the "complete data" sample. The average passing rate in the sample is 0.80 (across BMI and blood pressure metrics), and we estimate an effect of 0.012 (standard error of 0.001). That is, every dollar in increased financial benefits (per week) is associated with a statistically significant 1.2 percentage points (approximately 1.5%) higher pass rates.

4 | ASSOCIATED CHANGES IN HEALTH CARE EXPENDITURE

An important motivation for wellness programs across the country is the premise that beyond the obvious and indirect benefits associated with healthier workforce, there are also potential direct benefits in the form of lower medical health care expenditure, which would more directly translate to cost savings for employers. In this section, we use the employer's health care claims to provide evidence that relates health measures to health care expenditure.

The basic fact that in the cross section, healthier individuals spend, on average, less on health care than sicker individuals is widely documented,[9] and in Appendix Figure D1, we present a similar pattern in the context of our data. However, it may be less obvious that relatively short-run improvements in health translate to lower medical spending. To explore this in the context of our data, we estimate the following equation:

$$\log(1 + \text{MedExp}_{it}) = \alpha_i + \delta_t + \beta \cdot \text{Health}_{it} + \varepsilon_{it}, \tag{4}$$

where *Health* represents an individual health measure in a given year, and α_i and δ_t represent individual and year fixed effects, respectively. The dependent variable is the (logarithm of) total medical expenditure of individual i in year t, which we obtain by aggregating all of the individual's (who is covered by one of employer's PPO health plans) claims. We estimate this regression separately for each health measure, and the parameter of interest β is identified off variation in health measures for a given individual over time. Table 5 presents the results. Interestingly, the results suggest that reductions in BMI and blood pressure are associated with nontrivial reductions in health care spending, while changes in glucose and cholesterol measurements do not appear to systematically correlate with health care expenditure. For example, the estimates suggest that a one point reduction in BMI—which is approximately a 3 kg reduction in weight for most individuals—is associated with more than a 1.5% reduction in expected medical costs. Appendix Table D2 reports results from a two part model (Belotti, Deb, Manning, & Norton, 2015), showing that the relationship is driven by both the extensive margin (greater propensity of individuals to have a nonzero expenditure) and the intensive margin. Appendix Table D3 shows that these correlations are larger (in relative terms) for higher BMI and for higher spending individuals.

Moreover, because health care expenditure data are available for all employees (and their dependents) who enrolled in employer's PPO health plan, we can also report a reduced form estimate of the overall effect of the employer's workplace wellness program on health care expenditure. Specifically, we estimate the following regression:

$$\log(1 + \text{MedExp}_{it}) = \alpha_i + \delta_t + \sum_{t=2009}^{2012} \beta_t \cdot \text{Participation}_{it} + \varepsilon_{it}, \tag{5}$$

[9] See, for example, Finkelstein et al. (2009, 2010), Hammond and Levine (2010), Pronk, Goodman, O'connor, and Martinson (1999), Sturm (2002), and Wee, Phillips, Legedza, Davis, Soukup, Colditz, and Hamel (2005).

TABLE 5 The relationship between health measures changes and health care expenditure

	Dependent variable: log(medical expenditure + 1)				
	(1)	(2)	(3)	(4)	(5)
BMI	0.0174				
	(0.0074)				
BP (Systolic)		0.0036			
		(0.0012)			
BP (Diastolic)			0.0043		
			(0.0017)		
Cholesterol				0.0010	
				(0.0007)	
Glucose					0.0003
					(0.0014)
Mean of dependent variable	6.20	6.21	6.21	6.23	6.22
R^2: within	0.0005	0.0007	0.0005	0.0008	0.0006
R^2: between	0.0071	0.0017	0.0028	<0.0001	0.0010
R^2: overall	0.0056	0.0003	0.0006	<0.0001	0.0003
Hausman specification test					
Chi^2	41.45	104.28	143.03	100.39	46.11
p value	0	0	0	0	0
N (individual years)	55,373	55,974	55,909	39,234	39,232
N (unique individuals)	23,562	23,737	23,718	19,667	19,650

Note. Table reports coefficients and standard errors (in parentheses) from regressing log(annual medical expenditure + 1) on each biometric measure separately. Observations are individual years in the "complete data" sample restricted to individuals who got their biometrics measured. Biometric measurements are missing for some individuals who auto-passed by passing all five metrics by satisfying NIH standards in the previous year or received exemptions. For each biometric, we dropped observations below the 0.5th percentile and above 99.5th percentiles. Cholesterol and glucose measurement started in 2010 only. In adddition to the reported coefficient, each regression includes individual fixed effects and year fixed effects. Standard errors are clustered at the individual level.

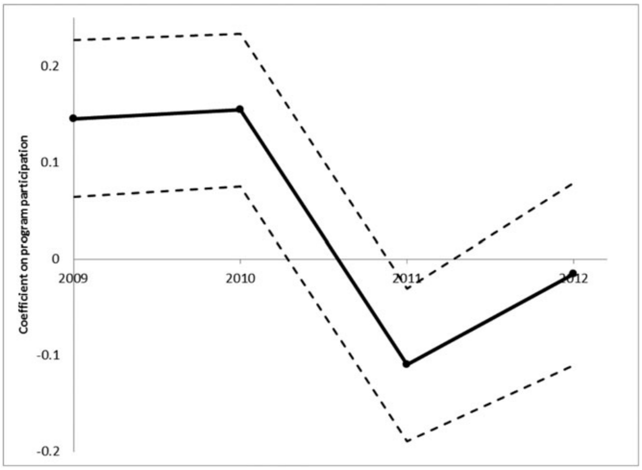

FIGURE 3 Relationship between program participation and health care expenditure. Figure plots estimated coefficients and 95% confidence interval of β_t from estimating Equation 5. Observations not only are individual years in the "utilization only" sample but also include observations from year 2008, the last year before the program started. There is a total of 109,288 individual-years observations (18,135 from 2008 and 91,153 from 2009–2012), covering 32,939 unique individuals

where $Participation_{it}$ is a dummy variable that is equal to 1 if an individual i participated in the program in year t and 0 otherwise. We again use individual and time fixed effects. Identification is primarily driven by incorporating into the data observations on 2008 health care spending, at which point, individuals have yet to participate in the program. We also obtain identification from the less frequent occasion in which eligible employees enroll in the program later or drop off. This variation is imperfect, so any attempt to attribute causal interpretation to the estimated coefficients should be cautioned. With this caveat in mind, the results, presented in Figure 3, suggest that initially (2009 and 2010) program participation is in fact associated with increased health care expenditure, but later, the health care expenditure declines to levels that are at or below the preprogram spending levels. One reasonable interpretation of these results is that initially, the workplace wellness program causes individuals to pay more attention to their health than before and incur certain health care expenditure, such as additional tests or preventive medication, that increases health care cost. In the next subsection, we indeed find that the use of preventive medication increases with the program. However, in the longer run, such increased expenditure improves one's health, and individuals thus reduce their (curative) spending.

4.1 | Preventive medication

There are multiple ways by which an individual could go about improving health measures, ranging from lifestyle and dietary changes to taking prescribed medications. Of the four health measures that are being used in the program, two—blood pressure and cholesterol—have a very easy ramification: preventive medication (see, e.g., Cholesterol Treatment Trialists, 2008; Krousel-Wood, Thomas, Muntner, & Morisky, 2004). Blood-pressure drugs are quite effective in reducing one's blood pressure, and anticholesterol drugs are effective in reducing one's cholesterol level. Figure 4 shows the overall improvement in these two health measures during the course of the sample, with noticeable decline in the higher end of the spectrum.

We use individuals' prescription drug claims to examine the timing of blood pressure and anticholesterol medication claims. To identify blood pressure and anticholesterol medication, we use the list of common medications used to treat high blood pressure and common medications used to treat high cholesterol. We then use label names included in the prescription drug claims to identify blood pressure and anticholesterol medication. We examine when individuals get their drug purchased relative to their biometric examination date.

Consistent with preventive drugs playing a role in these improvements, Figure 5 plots the timing of blood pressure and anticholesterol drug purchases, separately for individuals who passed and failed the screening. The figure clearly shows that the timing of drug purchase is not at all associated with the screening date for those individuals who passed

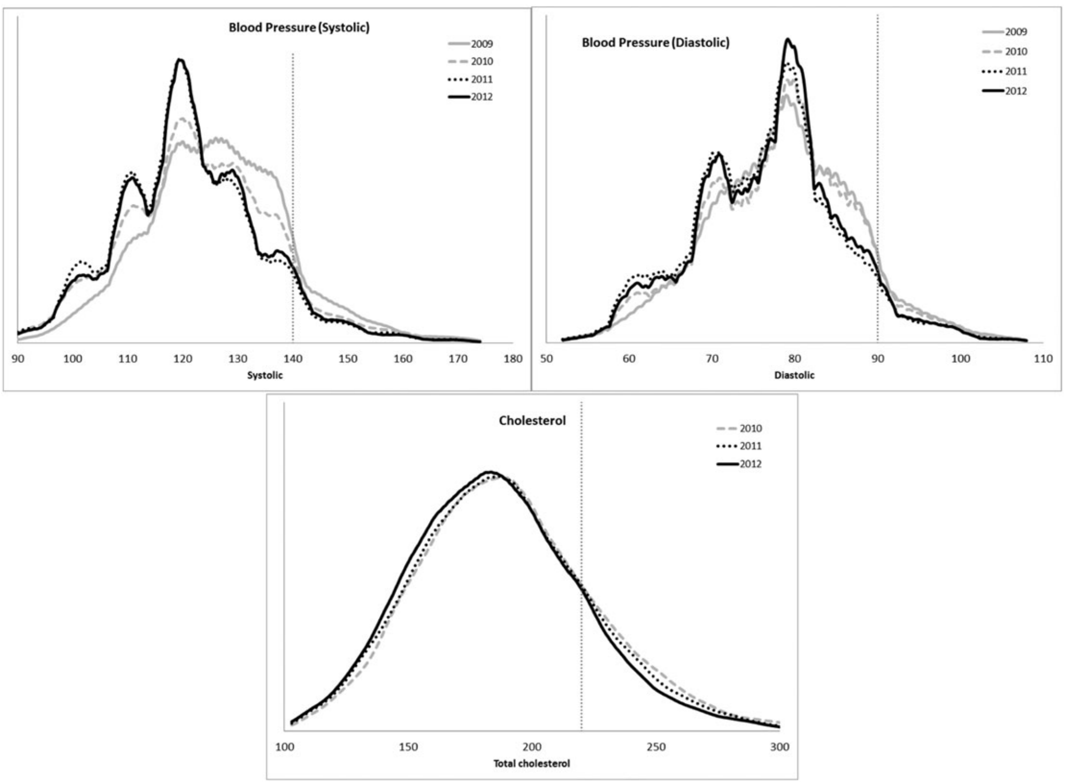

FIGURE 4 Improvements in blood pressure and cholesterol measures over time. Figure plots kernel densities of the cross-sectional distribution of systolic (top left), diastolic (top right), and total cholesterol (bottom), for each year (recall that cholesterol screening began in 2010). Observations are all individual years in the "health only" sample that were measured. Measurements are missing for some individuals who auto-passed by passing all five metrics by satisfying National Institutes of Health standards in the previous year or received an exemption. We dropped observations below the 0.5th percentile and above the 99.5th percentiles. The vertical lines show the passing threshold, above which an individual fails. The sample includes 77,958 individual-years observations (32,142 unique individuals) for systolic, 77,921 (32,116) for diastolic, and 54,892 (26,722) for cholesterol

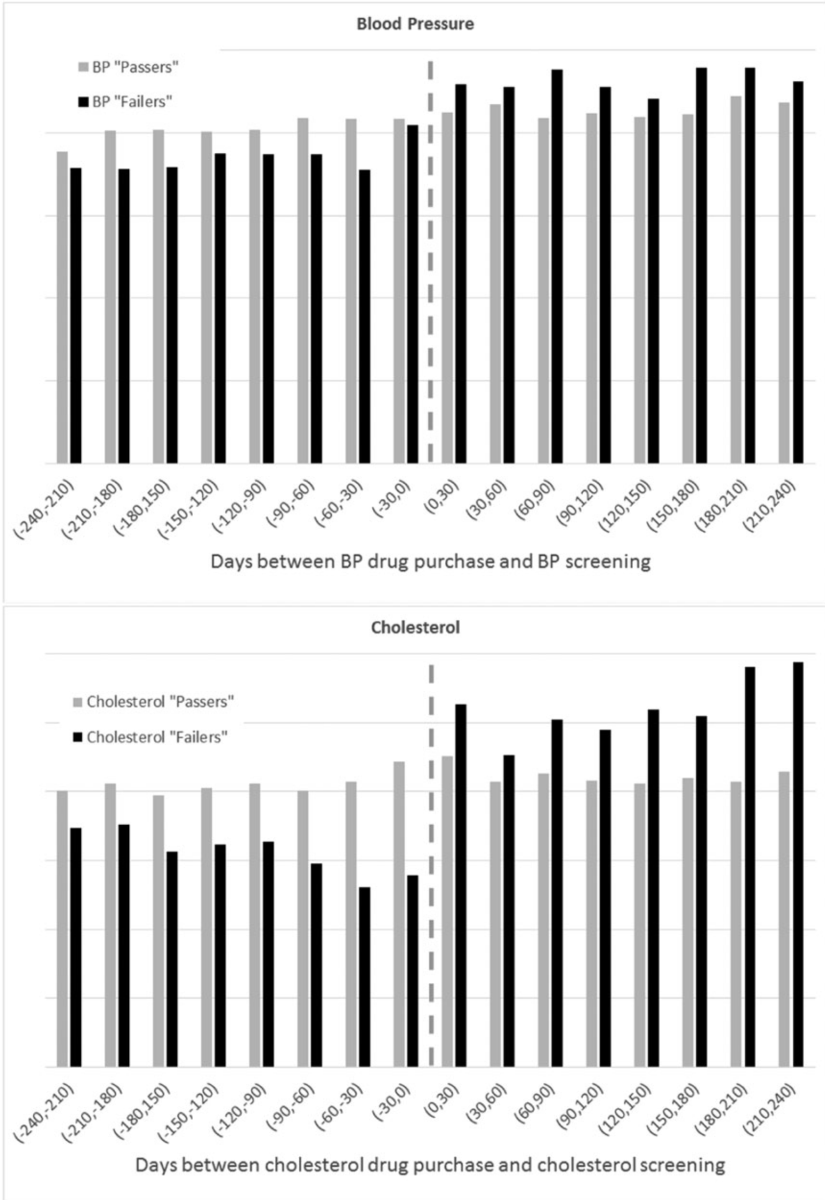

FIGURE 5 The timing of preventive drug purchasing and screening results. Figure presents the distribution of the timing of preventive drug purchasing around the screening date. The top panel shows results for blood pressure screening (and blood pressure medications) in 2009, and the bottom panel shows results for cholesterol screening (and anticholesterol drugs) in 2011. The key variable is the difference between the drug purchase and the screening date, with negative (positive) numbers reflect purchases that occurred before (after) the screening. Appendix C provides more details on the data construction. In both panels, the black bars present (normalized) frequency for those individuals who failed the screening, whereas the gray bars repeat the analysis for those who passed

the screening, presumably because such individuals are taking care of their blood pressure and/or cholesterol condition regularly. Yet for individuals who failed the screening, we see a clear pattern where they are much more likely to take preventive drugs after the screening relative to before, validating that the screening either alerts individuals to their previously undiagnosed condition or nudges them toward adherence.

5 | CONCLUSIONS

In this paper, we evaluated the impact of a novel wellness program of a large U.S. employer. We find robust evidence that health biometrics improved for program participants, especially for individuals who have been participating in the program for several years, and that these improvements—at least for BMI and blood pressure—are associated with reduced health care cost and utilization. Of course, our analysis and statistical inference rely on data from a single program implemented by a single company; further studies with similar programs and other employers are needed to assess external validity.

As wellness programs have spread, there is increasing interest and debate about their efficacy. This paper adds an additional data point in attempting to assess whether these programs do in fact improve employee health and reduce health care costs. We do not find clear evidence for overall cost reduction; however, it is possible that this objective should get less consideration in the short run. As we documented in Section 4, health care costs initially may rise with additional screening, and in fact, we observe an increase in the use of preventive medication with the program. At the same time, preventive care may imply a healthier workforce, improved productivity, and eventually lower costs. For example, Gubler et al. (forthcoming) find that a wellness program improves worker productivity. Assessing the longer run impact of wellness programs is an important avenue for future work.

ACKNOWLEDGMENT

We thank the editor and two anonymous referees for useful comments, Shirley Yarin for terrific research assistance, David Chen for helping us with the data, and Kent Bradley and Patricia Lin for answering our many questions. We acknowledge support from the Stanford Institute of Economics and Policy Research.

ORCID

Liran Einav http://orcid.org/0000-0003-3349-5356

REFERENCES

Aldana, S. G., Merrill, R. M., Price, K., Hardy, A., & Hager, R. (2005). Financial impact of a comprehensive multisite workplace health promotion program. *Preventive Medicine*, *40*(2), 131–137.

Alderman, L. (2009). Getting healthy, with a little help from the boss. *New York Times*, B6. May 22, 2009.

Baicker, K., Cutler, D., & Song, Z. (2010). Workplace wellness programs can generate savings. *Health Affairs*, *29*(2), 304–311.

Belotti, F., Deb, P., Manning, W. G., & Norton, E. C. (2015). twopm: Two-part models. *Stata Journal*, *15*(1), 3–20.

Bolnick, H., Millard, F., & Dugas, J. P. (2013). Medical care savings from workplace wellness programs: What is a realistic savings potential? *Journal of Occupational and Environmental Medicine*, *55*(1), 4–9.

Caloyeras, J. P., Liu, H., Exum, E., Broderick, M., & Mattke, S. (2014). Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. *Health Affairs*, *33*(1), 124–131.

Cawley, J., & Price, J. A. (2013). A case study of a workplace wellness program that offers financial incentives for weight loss. *Journal of Health Economics*, *32*(5), 794–803.

Charness, G, & Gneezy, U (2009). Incentives to exercise. *Econometrica*, *77*(3), 909–931.

Cholesterol Treatment Trialists (2008). Efficacy of cholesterol-lowering therapy in 18 686 people with diabetes in 14 randomised trials of statins: A meta-analysis. *Lancent*, *371*(9607), 117–125.

Claxton, G., Rae, M., Panchal, N., Whitmore, H., Damico, A., Kenward, K., & Long, M. (2015). Health benefits in 2015: Stable trends in the employer market. *Health Affairs*, *34*, 1779–1788.

Finkelstein, E. A., DiBonaventura, M. D., Burgess, S. M., & Hale, B. C. (2010). The costs of obesity in the workplace. *Journal of Occupational and Environmental Medicine*, *52*(10), 971–976.

Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., & Dietz, W. (2009). Annual medical spending attributable to obesity: Payer- and service-specific estimates. *Health Affairs*, *28*(5), w822–w831.

Fu, P. L., Bradley, K., Viswanathan, S., Chan, J. M., & Stampfer, M. (2016). Trends in biometric health indices within an employer-sponsored wellness program with outcome-based incentives. *American Journal of Health Promotion*, *30*(6), 453–457.

Gubler, T., Larkin, I., & Pierce, L. (forthcoming). Doing well by making well: The impact of corporate wellness programs on employee productivity. *Management Science*.

Hammond, R. A., & Levine, R. (2010). The economic impact of obesity in the United States. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, *3*, 285.

Horwitz, J. R., Kelly, B. D., & DiNardo, J. E. (2013). Wellness incentives in the workplace: Cost savings through cost shifting to unhealthy workers. *Health Affairs*, *32*(3), 468–476.

Jones, D., Molitor, D., & Reif, J. (2018). What do workplace wellness programs do? Evidence from the Illinois workplace wellness study. (*NBERWorking Paper No. 22796*) Cambridge, MA.

Kannel, W. B. (1987). Metabolic risk factors for coronary heart disease in women: Perspective from the Framingham study. *American Heart Journal*, *114*(2), 413–419.

Krousel-Wood, M., Thomas, S., Muntner, P., & Morisky, D. (2004). Medication adherence: A key factor in achieving blood pressure control and good clinical outcomes in hypertensive patients. *Current Opinion in Cardiology*, *19*(4), 357–362.

Mattke, S., Liu, H., Caloyeras, J. P., Huang, C. Y., Van Busum, K. R., Khodyakov, D., & Shier, V. (2013). Workplace wellness programs study: Final report. RANDCorporation. Available at http://www.rand.org/pubs/research_reports/RR254.html

Mochon, D., Schwartz, J., Maroba, J., Patel, D., & Ariely, D. (2016). Gain without pain: The extended effects of a behavioral health intervention. *Management Science*, *63*(1), 58–72.

O'Sullivan, J. B. (1974). Age gradient in blood glucose levels: Magnitude and clinical implications. *Diabetes*, *23*(8), 713–715.

Ozminkowski, R. J., Ling, D., Goetzel, R. Z., Bruno, J. A., Rutter, K. R., Isaac, F., & Wang, S. (2002). Long-term impact of Johnson & Johnson's Health & wellness program on health care utilization and expenditures. *Journal of Occupational and Environmental Medicine*, *44*(1), 21–29.

Pinto, E (2007). Blood pressure and ageing. *Postgraduate Medical Journal*, *83*(976), 109–114.

Pronk, N. P., Goodman, M. J., O'connor, P. J., & Martinson, B. C. (1999). Relationship between modifiable health risks and short-term health care charges. *Journal of the American Medical Association*, *282*(23), 2235–2239.

Sturm, R. (2002). The effects of obesity, smoking, and drinking on medical problems and costs. *Health Affairs*, *21*(2), 245–253.

Wee, C. C., Phillips, R. S., Legedza, A. T. R., Davis, R. B., Soukup, J. R., Colditz, G. A., & Hamel, M. B. (2005). Health care expenditures associated with overweight and obesity among US adults: Importance of age and race. *American Journal of Public Health*, *95*(1), 159.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Einav L, Lee S, Levin J. The impact of financial incentives on health and health care: Evidence from a large wellness program. *Health Economics*. 2019;28:261–279. https://doi.org/10.1002/hec.3840

APPENDIX A: COMPLETE DETAILS OF THE EMPLOYER'S WORKPLACE WELLNESS PROGRAM

Eligible participants can participate in any or all of the health measures they choose. Receiving incentive discounts for one passing measure is not contingent upon participation or meeting standards for other measures. The program's details regarding pass standards and incentive structure have slightly changed during each program year.

Appendix Table A1 summarizes passing standards for each measure and year (2009–2012). BMI is the ratio of the individual's weight (in kilograms) to the square of her height (in meters), with a BMI of 30 or below considered a passing result. Alternatively, individuals can pass the BMI metric with waist circumference measure that is less than 40 (for

TABLE A1 Passing standards in the health measure program, 2009–2012

Program year	BMI	BMI alternative (waist circumference)	Blood pressure	Cholesterol	Glucose	Nicotine
2009	<30	<40 in (men) < 35 in women)	<140/90 mmHg	–	–	Negative result
2010	<30	<40 in (men) < 35 in (women)	<140/90 mmHg	HDL > 40 mg/dl & LDL < 130 mg/dl & Triglyceride < 200 mg/dL[a]	<116 mg/dL[a]	Negative result
2011	<30	<40 in (men) < 35 in (women)	<140/90 mmHg	Total cholesterol[b]< 220 mg/dL	<116 mg/dL	Negative result
2011	<30	<40 in (men) < 35 in (women)	<140/90 mmHg	Total cholesterol[b]< 220 mg/dL	<116 mg/dL	Negative result

Note. BMI: body mass index. Table shows passing standards for each health measure. BMI is the ratio of the individual's weight (in kilograms) to the square of her height (in meters). Alternatively, individuals can pass the BMI metric with waist circumference measure. Blood pressure is measured using the systolic and diastolic millimeter of mercury (mmHg) readings. In 2009, cholesterol testing was only required of those in a "high risk" demographic group. Those at "low risk" received the cholesterol incentive by filling out a documentation form. Cholesterol and glucose measurement started in 2010 only, and passing standards were relaxed in 2011.

[a] In 2010, individuals had to pass both cholesterol and glucose to receive incentives associated with either of the two. [b] Total cholesterol is equal to low density lipoprotein + high density lipoprotein + triglycerides/5.

TABLE A2 Financial incentives associated with passing

	BMI	BP	Cholesterol	Glucose	Nicotine
PPO enrollees in Choice Fund I					
2009	6	1.5	–	–	6
2010	6	4	4	4	6
2011	4	4	4	4	4
2012	5.5	3	3	3	5.5
PPO enrollees in Choice Fund II					
2009	3	0.75	–	–	3
2010	5	2.5	2.5	2.5	5
2011	3	3	3	3	3
2012	4.5	2	2	2	4.5

Note. BMI: bosy mass index; BP: blood pressure; PPO: preferred provider organization. Table shows incentive amounts (in current $, per week). The first panel shows the weekly incentives received (if passed) by individuals enrolled in "Choice Fund I" (the higher coverage PPO health plan), and second panel shows the weekly incentives received by individuals enrolled in "Choice Fund II" (the lower coverage PPO health plan). We do not observe that PPO enrollees are in, but our understanding is that the majority of individuals were covered by Choice Fund I.

males) or 35 (for females) in. Blood pressure is measured using systolic and diastolic millimeter of mercury readings, with passing result requiring that the reading is both below 140 (systolic) and 90 (diastolic). In 2009, cholesterol testing was only required of those in a "high-risk" demographic group. Those at "low risk" received the cholesterol incentive by filling out a documentation form. Cholesterol and glucose measurements started in 2010, and passing standards for cholesterol and glucose have been relaxed by the employer in 2011. In 2010, individuals had to pass both cholesterol and glucose standards to receive financial incentives.

For each health measure participants passes, individuals receive financial incentives as shown in Appendix Table A2. Because the law regulates incentive amounts based on medical plan's costs, there are two different incentive schemes that depend on the individual's choice of PPO health plan. If an employee does not pass a test, they can receive incentives retroactively by retesting and meeting the standard the following year. For those who do not meet the BMI standard, they can also reduce their body weight by 10% or more in the subsequent year and pass the screening tests (even if the new weight does not meet the standard).

In 2010, an auto-pass system was introduced where if individuals passed all five biometrics by satisfying the NIH recommended levels, which are more stringent than the employer's (see Figure 1), individuals will not have to test again in the subsequent year and will automatically receive financial incentives for the subsequent year. The auto-pass requirements were as follows: BMI between 18.5 and 24.9, blood pressure less than 120/80, total cholesterol less than 200, glucose less than 100, and negative nicotine test.

Health screenings take place during the fall open enrollment period, and the financial incentive is applied to the following year's health insurance premium. Employees are informed about screening opportunities via a home mailer and postings on the company intranet. Employees also have access to an online benefits portal where screening details, scheduling dates, and deadlines are accessible. Health screenings can be done through on-site workplace screening, submission of results through independent labs or personal physicians or a testing kit mailed to the employee's home. A waiver can be submitted by individuals with special medical conditions in which it would be medically inadvisable or unreasonably difficult to satisfy the standard, such as pregnancy and type 1 diabetes.

APPENDIX B : ADDITIONAL DETAILS ABOUT THE CORONARY ARTERY RISK DEVELOPMENT IN YOUNG ADULTS (CARDIA) AND FRAMINGHAM DATA SETS

CARDIA. The CARDIA is a study examining the development and determinants of clinical and subclinical cardiovascular disease and its risk factors. It began in 1985–1986 with a group of 5,115 Black and White men and women aged 18–30 years. The participants were selected so that there would be approximately the same number of people in subgroups (of race, gender, education, and age) in each of 4 centers: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. These same participants were asked to participate in follow-up examinations. The study has completed seven clinical examinations with intervals of 2–5 years between consecutive measurements. Exam 5 held in 1995–1996 and included 3,883 individuals. Exam 6 held in 2000–2001 and included 3,627 individuals. The 3,322 individuals aged 32–49 (mean age 40) participated in both Exams 5 and 6 to get their BMI measured in both exams. The changes in BMI from Exams 5 to 6 provide benchmarks to compare BMI patterns observed at the employer. Additional details are available at https://www.nhlbi.nih.gov/research/resources/obesity/population/cardia.htm.

Framingham. The objectives of the Framingham Offspring Study are to study the incidence and prevalence of cardiovascular disease and its risk factors. The original Framingham study began in 1948 with 5,209 adult subjects from Framingham, MA. With the aging of the Framingham cohort and with interest in familiar aggregation and heritability, a new cohort consisting of the offspring of the original cohort was sampled. Spouses of offspring were also included. This new sample, begun in 1971, constituted a second generation of participants, permitted new assessment of risk factors and outcomes, and provided a resource for the genetic analyses, which were yet to come. The Framingham Offspring Study include 5,124 men and women, ages 5–70 years at entry consisting of offspring of the original Framingham cohort. By 2014, the study has completed nine clinical examinations with intervals of 4–6 years between consecutive measures. Exam 4 held in 1987–1991 and included 3,903 individuals. Exam 5 held in 1991–1995 and included 3,683 individuals. The 3,506 individuals aged 28–76 (mean age 51) participated in both Exams 4 and 5 to get their BMI measured in both exams. The changes in BMI from Exams 4 to 5 provide benchmarks to compare BMI patterns observed at the employer. Additional details are available at https://biolincc.nhlbi.nih.gov/studies/framoffspring.

APPENDIX C: ADDITIONAL DETAILS ON THE ANTICHOLESTEROL AND BLOOD PRESSURE PRESCRIPTION DRUG CLAIMS ANALYSIS

We use individuals' prescription drug claims to describe the timing of blood pressure and anticholestero medication claims. To identify blood pressure and anticholesterol medication, we used the list of common medications used to treat high blood pressure (http://www.webmd.com/drugs/newlinecondition-1432-High%20Blood% %20Pressure%20%20Hypertension%20.aspx) and common medications used to treat high cholesterol (http://www.webmd.com/drugs/condition-701-High%20cholesterol%20%20 Hypercholesterolemia%20.aspx), respectively. We then use label names included in the prescription drug claims to identify blood pressure and anticholesterol medication.

The prescription drug claims data cover 1,948,794 drug claims made by 45,903 unique individuals (employees and their spouses) in 2008–2012. The drug claims include a total of 302,220 blood pressure drug claims made by 14,041 unique individuals and 139,092 anticholesterol drug claims made by 9,313 unique individuals.

We observe the date of each biometric examination and the date of drug purchase. To study the timing of blood pressure medication claims, we examine when individuals get their blood pressure purchased relative to their blood pressure examination date. We examine blood pressure drug purchases that took place between 240 days before and 240 days after the 2009 blood pressure examination. Observations are 36,897 blood pressure drug purchases made by 3,322 unique individuals who participated in blood pressure examination in 2009. Among these 3,322 individuals, 2,450 unique individuals passed blood pressure examination in 2009 and made 27,051 blood pressure drug purchases (240 days before and after the examination). The remaining 872 individuals failed blood pressure screening and made 9,846 blood pressure drug purchases.

To examine the timing of cholesterol drug medication claims, we examine cholesterol drug purchases that took place between 240 days before and 240 days after the 2011 cholesterol examination. Cholesterol measurements started in 2010, and cholesterol became an independent biometric starting 2012. (In 2010, individuals had to pass both cholesterol and glucose standards to receive financial incentives.) We observe 21,031 anticholesterol drug purchases (made between 240 days before and after the cholesterol examination) made by 2,611 unique individuals who participated in cholesterol screening in 2011. Among these 2,611 individuals, 2,168 individuals passed the screening and made 18,704 anticholesterol drug purchases (240 days before and after the examination), whereas the remaining 443 individuals failed the cholesterol screening and made 2,327 cholesterol corresponding drug purchases.

APPENDIX D: ADDITIONAL TABLES AND FIGURE

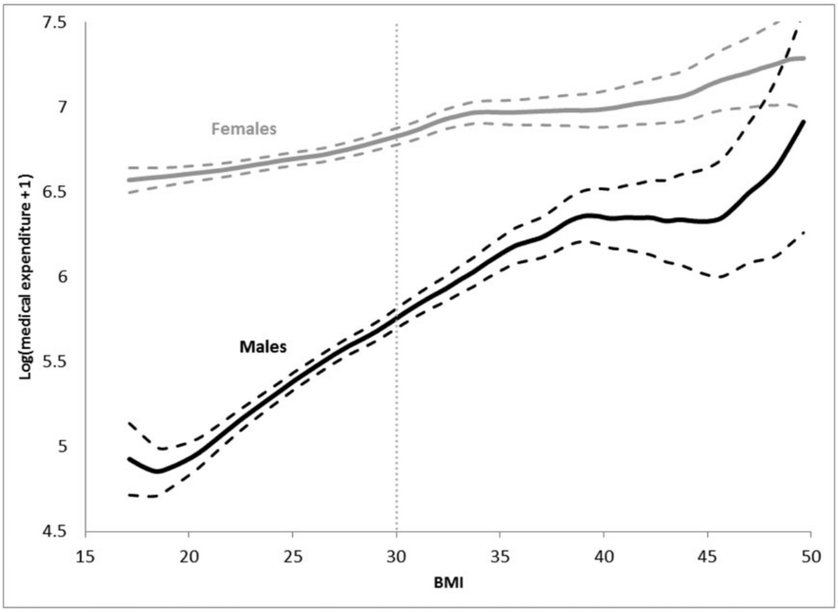

FIGURE D1 Relationship between body mass index (BMI) and health expenditure. Figure plots predictions from kernel regression of log(yearly medical expenditure + 1) on BMI, separately for men and women. The dashed lines show the 95% confidence intervals, and the vertical dotted line shows the employer's passing standard (of 30). Observations are individual years in the complete data sample for individuals who got their BMI measured. BMI measurements are missing for some individuals who auto-passed by passing all five metrics by satisfying National Institutes of Health standards in the previous year or received an exemption (e.g., due to pregnancy). We dropped observations with BMI below the 0.5th percentile and above 99.5th percentiles. The sample includes 55,373 individual-year observations, consisting of 10,892 unique men individuals and 12,670 unique women

TABLE D1 Comparisons with Coronary Artery Risk Development in Young Adults (CARDIA) and Framingham, adjusting for age and gender

	Matching method	Treated	Controls	Difference	Standard error	*t* statistics
Outcome variable: Change in BMI						
CARDIA	Unmatched	0.113	1.272	−1.159	0.05	− 21.83
	Kernel	0.151	0.905	−0.754	0.53	−1.42
	1-Nearest-neighbor	0.113	0.735	−0.622	1.24	−0.50
	5-Nearest-neighbor	0.113	0.841	−0.728	0.65	−1.12
Framingham	Unmatched	0.113	0.620	−0.507	0.05	− 10.37
	Kernel	0.113	0.742	−0.629	0.05	− 12.07
	1-Nearest-neighbor	0.113	0.526	−0.413	0.24	−1.74
	5-Nearest-neighbor	0.113	0.545	−0.432	0.12	−3.49

"Treatment" refers to workplace wellness program participants, and "control" refers to CARDIA or Framingham participants. We use logit regression and regress treatment on age and gender to estimate the propensity scores. For kernel matching, we use the Epanechnikov kernel and 0.06 bandwidth. Once individuals are matched, we can compare change in body mass index between workplace participants and CARDIA/Framingham participants. This is an estimate of the "average treatment effect on the treated." For the top panel, observations include 6,845 workplace wellness program participants and 3,322 CARDIA participants. For the bottom panel, observations include 6,845 workplace wellness program participants and 3,506 Framingham participants.

[a] For workplace wellness program participants, 4,631 are on support, and 2,214 are off support.

TABLE D2 Results form a two-part model

	Depedent variable: Log(medical expenditure + 1)									
	logit	GLM	logit	GLM	logit	GLM	logit	GLM	logit	GLM
A. Males										
BMI	0.0451	0.00434								
	(0.0036)	(0.00033)								
BP (systolic)			−0.0041	−0.00002						
			(0.0014)	(0.00013)						
BP (diastolic)					−0.0037	−0.00039				
					(0.0019)	(0.00019)				
Cholesterol							−0.005	−0.00036		
							(0.0005)	(0.00005)		
Glucose									0.0042	0.00048
									(0.0013)	(0.00010)
N (individual years)	25,768	25,768	25,845	25,845	25,821	25,821	18,630	18,630	17,893	17,893
N (unique individuals)	10,892	10,892	10,914	10,914	10,904	10,904	9,033	9,033	8,869	8,869
B. Females										
BMI	0.0114	0.00239								
	(0.0035)	(0.00020)								
BP (systolic)			−0.0079	−0.00006						
			(0.0017)	(0.00010)						
BP (diastolic)					−0.0047	−0.00017				
					(0.0024)	(0.00014)				
Cholesterol							−0.0021	−0.00007		
							(0.0007)	(0.00004)		
Glucose									0.0003	0.00029
									(0.0017)	(0.00010)
N (individual years)	29,605	29,605	30,129	30,129	30,088	30,088	21,863	21,863	21,352	21,352
N (unique individuals)	12,670	12,670	12,823	12,823	12,814	12,814	10,879	10,879	10,785	10,785

Note. BMI: body mass index; BP: blood pressure; GLM: generalized linear model. Table reports coefficients and standard errors (in parentheses) from a two-part model regressing log(annual medical expenditure + 1) on each biometric measure separately. A logit model is fit for the probability of observing a positive-versus-zero outcome. Then, conditional on a positive outcome, a GLM model is fit for the positive outcome. Observations are individual years in the "complete data" sample restricted to individuals who got their biometrics measured. Biometric measurements are missing for some individuals who auto-passed by passing all five metrics by satisfying National Institutes of Health standards in the previous year or received exemptions. For each biometric, we dropped observations below the 0.5th percentile and above 99.5th percentiles. Cholesterol and glucose measurement started in 2010 only. In addition to the reported coefficient, each regression includes year fixed effects and individual age fixed effects.

TABLE D3 Heterogeneity in the relationship between health changes and spending changes

	Dependent variable: Log(medical expenditure + 1)			
	Split by maximum expenditure		Split by maximum BMI	
	Below median	Above median	BMI ≤ 30	BMI > 30
BMI	0.0157	0.0197**	0.0144	0.0236*
	(0.0121)	(0.00929)	(0.00892)	(0.0129)
N (individual years)	27,637	27,736	44,040	11,333
N (unique individuals)	13,022	10,540	18,204	5,358

Note. BMI: body mass index. Table reports the same specification as in Table 5 of the main text but splits individuals based on their spending levels (left two columns) and BMI levels (right two columns). To create the split, we generate the highest annual spending (or BMI) of an individual that is observed over the entire observation period and use this highest level to define individual as above or below the median highest spending or above and below a highest BMI of 30.