

Determining the Value of Wellness Programs

Recent advances can give companies a solid set of return on investment (ROI) measurements on their health improvement programs, provided they are willing to invest in both wellness programs and measurement efforts that effectively gauge those programs' merit. As this article explains, choosing the right methodology will depend on the health improvement programs being evaluated, data and resources available, and the degree of precision desired by management. The authors discuss the different measurement methodologies and various measurement considerations. They conclude that using several methods and multiple iterations under varying sets of assumptions is often useful, not only for calculating ROI but also for providing companies a framework for continual program tracking and improvement.

by **Ronald G. Barlow** | *PricewaterhouseCoopers* and **Don Weber** | *PricewaterhouseCoopers*

Over the last decade, employers, led by those with more than 5,000 employees, have been instituting a variety of health management programs designed to improve employees' (and their dependents') health and productivity. These health improvement programs, ranging from worksite wellness, on-site clinics, health coaching and advocacy, disease management and clinical management, have vastly expanded during the last three to five years, even though their value is still often not fully understood.

Among large employers, 88% responding to the 2011 *PwC Health and Well-Being Touchstone Survey* said they offer some type of wellness program, and 86% offer disease management programs to health care benefit program participants.¹ While respondents tended to doubt program effectiveness, most planned to double down on their investment: 55% believed their wellness programs were minimally or not effective, but 66% were planning to increase their investments in this area.

This raises a crucial question: How, exactly, is program effectiveness being measured? According to a 2010 joint National Business Group on Health and Fidelity Investments survey, only one-third of employers have measurable goals/targets for their health improvement programs, and 59% of employers don't know their return on investment (ROI).² Determining a "CFO-credible" methodology of measuring the cost savings, health improvement or an overall ROI in these programs has been complex, confusing and, at times, simply nonexistent. Many CFOs and other business leaders question the value of these investments and are asking for proof of a positive ROI and clear, understandable and defensible analyses that demonstrate such proof.

The problem is that most human resources (HR) departments either have simply not measured their savings or ROI, or they're depending on the health care vendors—which have a vested interest in the programs—to do so. Often,

these vendor-provided ROI demonstrations use either overly simplified methodologies or extremely complex methodologies, both of which produce questionable results. In addition, most of these demonstrations rely on a single measurement or a very limited set of measurements, which increases the risk of producing inconclusive results.

The good news? Recent advances in data/information availability, together with advanced data analytic methods, can result in a solid set of ROI measurements, provided companies are willing to invest not only in the wellness programs, but in measurement efforts that effectively gauge their merit. These approaches can be as simple as measuring the reduction in negative medical events, such as hospital admissions or emergency room (ER) visits for those enrolled in a condition-specific disease management program. Or they can be more complex, populationwide comparisons. Choosing the right methodology will depend on the health improvement programs being evaluated, data and resources available, and the degree of precision desired by management.

Measurement Methodologies

At their core, ROI measurements hinge on the determination of savings. For health improvement programs, savings can be based on medical claims cost only or more holistically estimated to include items such as productivity or other

business results. Simply put, if it can be demonstrated that the effects of a particular health care program reduced the company's costs over what they would have been without the program, we have measurable savings and the basis for a valid ROI measurement. However, the real challenges lie in knowing what the costs would have been without the program and whether observed changes represent a true cause-and-effect impact.

There are many methodologies in use today to measure the effectiveness of health improvement programs, with varying degrees of applicability. Generally, they can be grouped into the following four categories.

1. Test/Control Group Methods (Participant vs. Nonparticipant)

One general category of measurements is based on the comparison of the change in costs (i.e., trend rate) or other statistics between a test group and a control group. The test group represents those members who participated in the particular health improvement program under review, and the control group usually comes from those who did not. As described in the figure, the control group can be adjusted in various ways to arrive at a better comparison. Correct determination of the control group is one key to arriving at a successful measurement under the test/control group analysis

FIGURE

Control Group Determination Methods

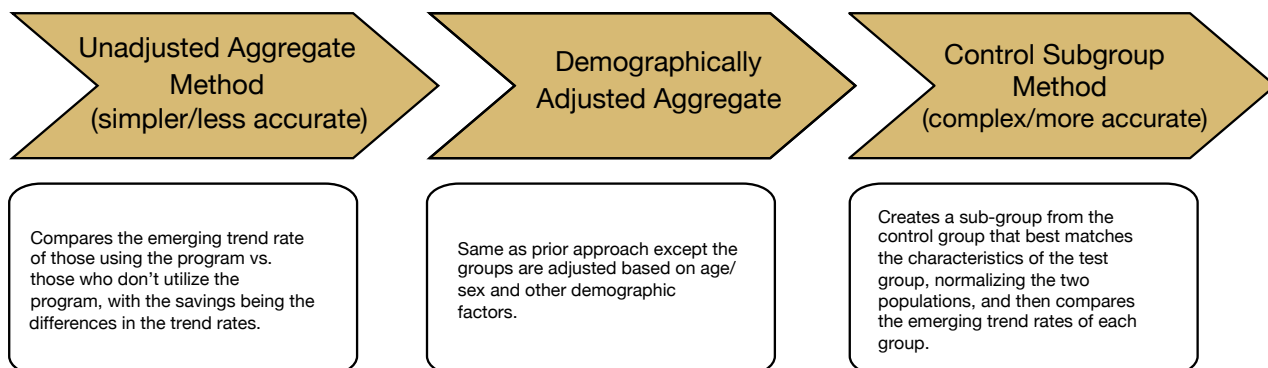


TABLE I**Example of Unadjusted Aggregate Method**

	Test Group (2010 Program Participants)	Control Group (2010 Program Nonparticipants)
Number of employees	1,000	9,000
2009 costs per employee per year	\$12,000	\$8,000
2010 costs per employee per year	\$12,600	\$8,680
Actual trend rate	5.0%	8.5%
Estimate of what the 2010 costs would have been absent the program = $\$12,000 \times 1.085$	\$13,020	
Estimated savings per employee per year: $\$13,020 - \$12,600$	\$420	
Estimated total savings: $\$420 \times 1,000$	\$420,000	

(Assume a 1/1/2010 effective date for the employee-only health improvement program and a one-year before and after look.)

method. Techniques for determining the control group have evolved in recent years as additional data elements and more advanced analytic methods are deployed.

The most basic and common method is what we call the *unadjusted aggregate* method. This method uses the observed trend in the control group to adjust the test group results, which are then compared to the test group actual results to determine savings. Table I demonstrates how this works conceptually.

Essentially, this method assumes that the change for the control group (those who did not participate) is an accurate predictor of what would have happened to the test group (those who did participate) without the program in place.

This approach requires the least amount of information, but ignores what could be significant differences

between the two groups in terms of age, gender and other demographic factors. For example, if only young, single employees choose the program, and we see a large cost trend difference when compared to older employees who have families and who did not participate, then the measured savings may not be an accurate reflection of program effectiveness alone.

Accounting for differences in demographics between the test and control groups adds a degree of precision in the aggregate method. This is accomplished by adjusting the change in costs being measured based on the average demographic indicators. Demographic differences are usually accounted for by using standard actuarial values to reflect differences in age, gender, family status and sometimes geography. Table II shows how this works conceptually.

Even when the control group is adjusted for differences in demographics,

differences can be present due to other factors, most notably health status and selection. Two people with exactly the same demographic profile can have markedly different health statuses and, therefore, significantly different cost and trend levels. Also, voluntary participation introduces an element of selection. Often, someone who chooses to participate in a wellness program might be predisposed to making lifestyle changes based on a desire to improve health status; this can make it seem as if the program is more effective than it really is.

Accounting for such factors can be complicated. Some of the latest techniques involve using multivariate regression analysis to determine which factors are most closely correlated with the cost levels or other items being measured. Then, using the results, a new control group is created from the nonparticipating population that best

TABLE II**Example of Demographically Adjusted Aggregate Method**

	Test Group (2010 Program Participants)	Control Group (2010 Program Nonparticipants)
Number of employees	1,000	9,000
2009 costs per employee per year	\$12,000	\$8,000
2010 costs per employee per year	\$12,600	\$8,680
Actual trend rate	5.0%	8.5%
Demographic factors:		
2009 demographic factor	1.214	0.976
2010 demographic factor	1.262	1.011
Demographic impact on trend	4.0%	3.6%
Adjusted trend to apply to test group = (1.085 × 1.040 ÷ 1.036) – 1	8.9%	
Estimate of what the 2010 costs would have been, absent the program = \$12,000 × 1.089	\$13,066	
Estimated savings per employee per year: \$13,066-\$12,600	\$466	
Estimated total savings: \$466 × 1,000 = \$466,000	\$466,230	

(Assume a 1/1/2010 effective date for the employee-only health improvement program and a one-year before and after look.)

replicates how the test group would have performed if the program had not been in place. Factors that can be considered in the multivariate regression analysis are:

- Age
- Gender
- Family status
- Level/salary
- Risk score
- Total health care costs—medical
- Total health care costs—prescription drugs
- Health risk assessment (HRA) completion and results
- Biometric measurements
- Chronic condition category
- Comorbidities
- Preventive services used.

This new control group is a subset of the initial control

group population that best replicates the underlying tendencies of the test group, apart from the influences of the health care program being evaluated. Some degree of “experimentation” may be necessary to find the best set of factors and the match thresholds (i.e., five-year age buckets instead of exact year match for age) that will be used to create the subgroup. See Table III as an example.

Short of a pure randomized clinical sampling, which is not a viable solution for most employers, no test/control group method is scientifically perfect. But if the data is available, this multivariate regression method appears to do the best job of aligning the control group with the test group for an accurate assessment of the impact of the health improvement program. The control subgroup method can be expected to yield a better measurement of the true effects of the health care program than the unadjusted aggregate method or de-

TABLE III**Example of Control Subgroup Method**

	Test Group (2010 Program Participants)	Initial Control Group (2010 Program Nonparticipants)	Adjusted Control Group (Subgroup of 2010 Program Nonparticipants)
Number of employees	1,000	9,000	1,000
2009 costs per employee per year	\$12,000	\$8,000	\$11,800
2010 costs per employee per year	\$12,600	\$8,680	\$13,100
Actual trend rate	5.0%	8.5%	11.0%
Estimate of what the 2010 costs would have been absent the program = $\$12,000 \times 1.110$	\$13,322		
Estimated savings per employee per year: \$13,322-\$12,600	\$722		
Estimated total savings: $\$722 \times 1,000$	\$722,034		

(Assume a 1/1/2010 effective date for the employee-only health improvement program and a one-year before and after look.)

mographically adjusted aggregate method. However, the control subgroup method is more data- and labor-intensive. In some cases, where the necessary data is not available or the size of the population is not large enough, the control subgroup method may simply not be feasible.

2. Populationwide Analyses (Also Called Historical Control Analyses)

Populationwide analyses usually look at trends in undesirable health utilization statistics across the entire population. Also referred to as a *measurement of negative medical events*, this methodology uses the populationwide analysis of events such as hospital admissions or ER visits to measure the value of specific health improvement programs. It can also include measurement of changes in the number of health risks or occurrences of certain disease states. Care should be taken to adjust for differences in population levels, demographics, plan changes and other factors from year to year that can impact observed trends. Comparison to previous years' events for the identified groups, or normative data, if available, can also be helpful in determining whether the observed trends represent a measurable reduction in cost.

The types of health care statistics that can be included in the populationwide analyses and tracked over time include:

- Medical claims
- Average risk scores
- Number of sick days
- Number of hospital admissions
- Number of hospital readmissions
- Number of ER visits
- Percent of members having a physical exam
- Percent of members getting immunizations
- Number of new diabetic cases
- Number of new renal failure/dialysis cases
- Number of new cardiovascular disease cases
- Number of heart attacks
- Number of strokes
- Number of back surgeries
- Number of knee replacements
- Number of diabetes-related complications, such as amputations
- Number of screening tests completed, such as HbA1c, foot exams, lipid panels or cancer screenings
- Wellness program participation rates

Using multiple analyses reduces the risk of having inconclusive or misleading results.

And aggregating results from various analyses into one report or scorecard can be helpful in understanding the big picture, especially when monitoring progress over time.

- HRA completion rates
- Number of health risks identified through the HRA
- Biometric results, such as weight, body mass index (BMI), waist size, blood pressure, cholesterol levels, etc.

3. Longitudinal Studies of Participants (Also Called Pre-Post Cohort Analyses)

In this methodology, members working to manage a given condition, such as diabetes, can be analyzed by use of certain identified statistics or medical events (see the list above) during a base period. The impact of the program is then determined by analyzing the same group after program implementation to determine if there was a reduction in the identified medical events. The average cost of the medical events can be calculated using the employer's own data or normative data.

The methodology does require the employer and the vendor to work together to identify the group to be measured, the diagnostic codes to be used and the events to be measured. For this measurement, care needs to be taken to ensure that the results observed are not significantly influenced by *regression to the mean* (defined later).

4. Qualitative Assessments

In addition to the quantitative analyses described above, it may often make sense to also include a qualitative analysis that assesses such factors as program features, activities, accomplishments, employee feedback surveys and self-reported results. The qualitative analysis is important to complement and explain the numbers and to suggest alternative directions for the quantitative analyses, as well as provide insight to possible program refinements.

A Balanced Scorecard Approach

Unfortunately, there is no “silver bullet.” There are pros and cons for each of the analytic methods described above. As such, it may make sense to include some or all of the methods into a set of measurements that collectively can provide a balanced assessment of the effectiveness of the health improvement programs.

Using multiple analyses reduces the risk of having inconclusive or misleading results. And aggregating results from various analyses into one report or scorecard can be helpful in understanding the big picture, especially when monitoring progress

over time. In this effort, it's important to clearly lay out and understand the strengths and limitations of each method used, which may vary based on the data and the programs in place. Corporate leadership, including the finance department in particular, should be involved when deciding the overall structure of the measurement and ROI reporting.

Measurement Considerations

These areas should be considered when conducting an analysis of the effectiveness of health improvement programs:

- **Data sources.** Depending on the type or scope of the health improvement program and the extent of the measurement analysis, data sources can include medical and prescription drug claims extracts, health risk appraisal information, biometric screening data, risk scores (calculated from the other data), eligibility data, absence and disability claims extracts, and other business results data.
- **Data credibility.** Credibility should be a primary consideration in any analysis of health care claims data. Generally

speaking, the more data that's available and the "cleaner" it is, the more credible will be the results of the analysis and the more precise the savings estimate.

- **Years to be included in the study.** It's preferable to have two years of data prior to the health improvement program implementation date as a sound basis for measurement. This is especially true if the number of participants is not enough to be considered fully credible with one year alone. At least one year of postimplementation data usually is needed to draw any significant conclusions on the effectiveness of the health improvement program, and multiple years are preferred for observing lasting effects.
- **Populations to be included in the study.** Generally speaking, all employees and dependents who have access to the health improvement program and are therefore considered eligible should be included in the study.
- **Outliers.** When analyzing claims data, *outliers* (high-cost claims) may sometimes skew results. Establishing a claims amount threshold, say \$50,000 per year, and then examining the results with and without the claims capped at this threshold can be beneficial in determining the extent to which outliers are affecting results.
- **Savings criteria.** Reductions in medical claims paid is commonly used to measure savings. However, other measures can include change in certain health utilization statistics (e.g., hospital admissions), change in certain health indicators (e.g., risk scores or BMI), reductions in the number of persons with chronic conditions (e.g., diabetics), or changes in productivity or other business measures. Generally speaking, the use of changes in utilization is preferred over changes in costs because health improvement programs are designed primarily to affect health care utilization. Focusing on utilization also minimizes the effects of outliers, changes in provider reimbursements and health care inflation. If we think of costs as being expressed by the simple equation $\text{Total Cost} = \text{Utilization} \times \text{Unit Cost}$, then focusing our measurement analyses on the changes in health care utilization can still be translated into cost savings at the end of the day.

- **Criteria for health improvement program participation.** Health improvement programs often reflect multiple levels of involvement. For some eligible members, participation may go no further than receiving an initial contact; others might have responded with some interest, be fully engaged or have "graduated" from the program. Performing several iterations at various levels of the participation criteria may be warranted.
- **Diagnoses to be excluded, such as maternity and delivery.** As with outliers, the inclusion of members who have certain diagnoses or conditions can sometimes skew results. These conditions, such as maternity and delivery or accidents, usually are unrelated to the health

AUTHORS

Ronald G. Barlow is a managing director in the PricewaterhouseCoopers Human Resource Services practice in Chicago with over 30 years of health care actuarial consulting experience in the employer-sponsored health benefits field. Barlow's primary emphasis is the financial and actuarial analysis of health care programs, including health care strategy development, plan design modeling, wellness and care management programs, health outcomes measurements and the impact of health reform.



Don Weber is a managing director in the Atlanta, Georgia office of PricewaterhouseCoopers' Human Resource Services practice. He specializes in the managed health care field dealing with employers, payers, pharmacy benefit managers and providers.



Weber's 30 years in the health care field includes vendor selection, analysis of pharmacy benefit managers, analysis of employee health care costs, health and productivity strategy including absence management, mental health programs, managed care contracting, strategic planning, hospital operations and financial management.

improvement program. Again, it may be beneficial to examine results with and without the members who have these identified diagnoses.

- **Regression to the mean.** In statistics, *regression to the mean* is the phenomenon that if a variable is extreme on its first measurement, it will tend to be closer to the average on a second measurement. For analysis of health outcomes, this comes into play, for example, when looking at a closed group of members in a particular disease category. In the first year, high claims could have triggered program participation. But a look at these same individuals in the second year will show that some naturally tend to have more average claims. Without an influx of new high-cost members in the second year, it will appear as if improvement in costs occurred. Care must


be taken in any analysis of health improvement programs to ensure that this phenomenon is not skewing results.

Analytics Provide the Path to Tracking Value

The measurement of health improvement program value can be a daunting task for any plan sponsor. Often, a purely scientific approach is simply not possible because what is being evaluated is often something that did not occur. In a postreform era of employer health plans, designing cost-effective programs will remain a top priority, and by applying the analytic approaches described here, employers can begin to produce at least a reasonable estimate of the value of various programs.

In most cases, the use of several methods and multiple iterations under varying sets of assumptions is useful in understanding the range of possible results. The use of an independent advi-

sor to provide a qualified second opinion can often be helpful in determining the correct methodology and deriving assistance with calculations.

But these analytic techniques can deliver more than a calculation of savings and ROI. Thoughtful analytics can also provide a framework for the continual tracking and improvement of the value of an organization's programs and for determining which programs should be enhanced, altered or discontinued. 

Endnotes

1. PwC. May 2011. "Health and Well-Being Touchstone Survey Results," available at www.pwc.com/en_US/us/hr-management/assets/PwC_2011_Health_and_Wellbeing_Touchstone_Survey_Results.pdf.
2. Fidelity Investments. 2010. "Improving Health Outcomes in 2010, Results from the Joint National Business Group on Health/Fidelity Investments Survey," *Fidelity Perspectives*: Winter 2010, published by Fidelity Workplace Investing in conjunction with Fidelity Consulting Services and the National Business Group on Health. Available at http://worldcongress.com/events/HR10000/pdf/thoughtleadership/FINAL%20NBGH_Fidelity_Brief%20Feb%202010.pdf.

Copyright of Benefits Quarterly is the property of International Society of Certified Employee Benefit Specialists and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.