Challenges with measuring savings in shared savings arrangements

Jill Herbold, FSA, MAAA
Anders Larson, ASA, MAAA

Milliman

In recent years, shared savings arrangements have become an increasingly popular way to incentivize providers to improve the efficiency and quality of healthcare. Spurred in part by provisions in the Patient Protection and Affordable Care Act (ACA), these arrangements attempt to tie provider reimbursement to performance on quality measures and reductions in the healthcare expenditures for an assigned population of patients. The most common form of these arrangements involves networks of providers that form accountable care organizations (ACOs) to contract with public or private payers.

Although both providers and payers agree on the terms of these arrangements conceptually, the practical task of measuring improvements by providers is another matter. In particular, attempting to measure reductions in expenditure levels that are due to actions by providers is often extremely challenging. Changes in risk profile, selection bias, outlier claims, and underlying medical trends can all influence how expenditure levels change for a population over time. For these arrangements to work for both parties, the measurements must be as accurate, fair, and transparent as possible.

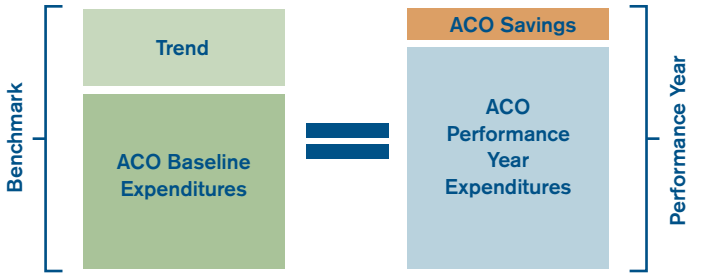
This paper provides an overview of the basics behind measuring savings in shared savings arrangements, discusses common challenges with measuring savings, and explores the importance of the patient assignment methodologies that are used in these programs, such as the Pioneer ACO program and the Medicare Shared Savings Program (MSSP). For simplicity in the remainder of the paper, we will refer to the provider network in these arrangements as the ACO.

Savings formula basics

In most shared savings arrangements, patients are already covered by the payer, but they do not actively choose to be involved in the shared savings arrangement. Instead, the patients are passively assigned to the ACO using an algorithm that attempts to identify which providers are responsible for their primary care. Patients are still free to use any provider covered by their payer; whether or not they are assigned to the ACO only affects whether or not the ACO is responsible for their expenditures in the shared savings arrangement.

Most shared savings arrangements have the same basic formula to estimate savings, which is illustrated in Figure 1. First, expenditures are measured in a historical (baseline) period. Next, an adjustment is made to this expenditure level to "trend" that experience to be comparable to the current (performance) period. This adjusted expenditure level is often referred to as the benchmark. On the opposite side of the equation, expenditures are measured in the performance period. The difference between the benchmark expenditures and the performance period expenditures is referred to as the gross savings.

FIGURE 1: ESTIMATING SAVINGS



In order to determine the final shared savings payment made to the ACO, the gross savings is adjusted based on the sharing percent, quality results, and other terms of the particular shared savings arrangement. In some cases there is no shared savings payment at all if the gross savings does not meet a minimum threshold. Depending on the arrangement, the ACO may also share in losses if performance year expenditures exceed the benchmark.

Not quite that simple

The key is to estimate what the expenditures would have been in the absence of the intervention. Without a random control trial, which is not practical in a shared savings arrangement, measuring savings (or lack thereof) is inherently tricky. There are a variety of ways of attempting to do this, and these options can often produce vastly different results, none of them necessarily "correct." The general concept here is not unlike measuring savings from a disease management program.[1]

The trend component can be calculated in a variety of different ways. For instance, the trend can be based on an agreed-upon market trend, such as the Milliman Medical Index™, or it can be based on the changes in expenditures for a reference population that is comparable to the ACO population. For example, the reference populations in the Pioneer ACO program and MSSP are nationwide groups of beneficiaries that meet the applicable program criteria to be assigned to an ACO. Further, the trend is often adjusted for changes in health status or demographic mix of the ACO population from the baseline to the performance period. The selection of a risk adjustment process is therefore an important component of the trend calculation.[2]

The performance period expenditures are often adjusted in some way in these arrangements. One common adjustment is to remove or truncate large claim costs for individual patients in order to reduce statistical variation of expenditures in the estimation of savings. Setting the threshold for this adjustment should be done carefully because managing high-cost, complex chronic conditions is a key way for providers to reduce overall expenditure levels. For instance, MSSP sets this threshold as the 99th percentile of annual expenditures for all ACO-eligible beneficiaries nationwide within each Medicare entitlement category, such as aged non-dual, aged dual, disabled, or end-stage renal disease (ESRD).

Which patients?

One major challenge in shared savings arrangements is determining which patients should be reflected in the baseline period expenditures and which patients should be reflected in the performance period expenditures. Because patients are passively assigned to the ACO in most shared savings arrangements, the details of the assignment algorithm will have a large effect on which patients are assigned. Additionally, some arrangements use the same assigned patients in the baseline period as in the performance period (cohort approach), while other arrangements use the same algorithm to assign patients in each period (cross-sectional approach). These two approaches can produce very different results because the baseline period assigned population differs between the two approaches.

Patients also may be attributed prospectively or concurrently. In prospective assignment, the patients are assigned to the ACO for the performance year on the basis of historical claims and generally cannot lose assignment during the performance year, regardless of whether they see ACO physicians during that time. In concurrent assignment, patients are assigned to the ACO for the performance year based on claims during the performance year. In this case, it cannot be known with certainty which patients will be assigned until the performance year is over.

We'll explore these approaches to the patient assignment methodology in more detail over the remainder of this report, focusing on how they impact savings in shared savings arrangements.

Bias in using cohort approach

As discussed earlier, the cohort approach for measuring savings uses the same set of patients in the performance period and the baseline period. The most prominent shared savings arrangement to use this approach is the Pioneer ACO program, in which the Centers for Medicare and Medicaid Services (CMS) contracts with ACOs to manage expenditures for traditional fee-for-service Medicare beneficiaries.[3] The basic goal for ACOs in the Pioneer program is to have a lower trend than the national reference population. The program began in 2012 and used the cohort approach for the first three years before switching to a cross-sectional approach for 2015 and 2016. The use of the cohort approach necessitated additional complexities, such as a decedent adjustment to account for the fact that the baseline period would not include anyone who died during that time period.

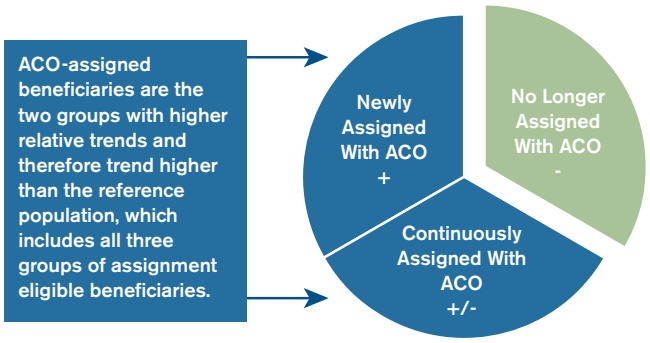
This approach also introduced a bias into the savings calculations that was problematic for many ACOs. The bias is due to differences in the composition of an ACO population compared to the national Medicare reference population. An ACO population, in any given year, contains two types of patients: those who were assigned to the ACO in the prior year and patients who are newly assigned this year. However, the reference population includes a third type: patients who were assigned to an ACO in the prior year but are not assigned in the current year. This is because the reference population includes all *ACO-eligible* beneficiaries, regardless of which (if any) ACO they are assigned to.

1 For more background on this, Ian Duncan's *Managing and Evaluating Healthcare Intervention Programs* looks at measuring the effectiveness of disease management programs under several different methodologies.

2 Leida, H.K. & Wachenheim, L.M. (January 2015). Risk Adjustment and Shared Savings Agreements. Milliman Healthcare Reform Briefing Paper. Retrieved February 18, 2015, from http://us.milliman.com/uploadedFiles/insight/2015/shared-savings-agreements.pdf.

3 Boyarsky, V. & Parke, R. (May 2012). The Medicare Shared Savings Program and the Pioneer Accountable Care Organizations. Milliman Healthcare Reform Briefing Paper. Retrieved February 18, 2015, from http://publications.milliman.com/publications/healthreform/pdfs/medicare-shared-savings-program.pdf.

FIGURE 2: COMPOSITION OF ACO POPULATION VS. REFERENCE POPULATION



We have found that newly assigned patients tend to have above-average trends from the baseline period to the performance period, whereas patients who lose assignment tend to have below-average trends between these periods. The chart in Figure 3, which is based on data from a Pioneer ACO in its third performance year, shows the expenditures over time for the newly assigned patients. Note the spike in expenditures toward the end of the assignment period, indicated by the blue vertical lines. Expenditures do not fall back to earlier levels after this spike, leading to a 94% trend from the baseline to the performance year.

The reason for these excessively high trends is that patients are often assigned to the ACO because of an acute medical event that caused them to visit ACO physicians. However, these patients were generally healthier in earlier time periods, which is precisely why they were not visiting ACO physicians at that time. With the cohort approach, the ACO is now responsible for all historical expenditures for these patients, inflating the ACO trend. The reference population also includes these type of patients, but the high-trend patients are offset by the low-trend patients who recently lost assignment.

This situation may be exacerbated if the ACO includes providers such as cardiologists or skilled nursing facility physicians who deal with patients during high-cost episodes. Trends also might not be as extreme as the example above in shared savings arrangements where the baseline period is the same as the assignment period.

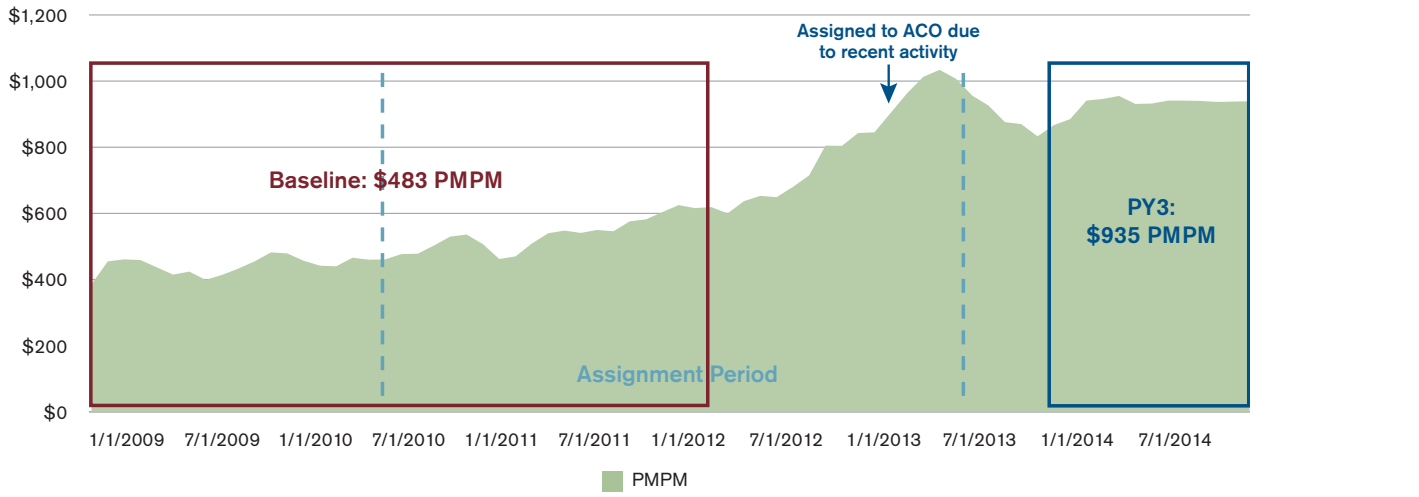
The cross-sectional approach eliminates much of this bias because when patients are newly assigned to an ACO their historical expenditures are ignored. Conversely, when patients lose assignment, their historical expenditures are still used in calculating baseline expenditures. The expenditures in historical periods do not change over time, assuming the ACO does not change its list of participating providers. For one ACO, we estimated that using the cross-sectional approach would have improved the gross savings percentage by approximately 6% and allowed the ACO to share in substantial savings.

The Pioneer ACO program has moved to a cross-sectional approach in 2015, consistent with the approach used by MSSP since its inception in 2013.

Unintended consequences that are due to assignment methodology

Even if the cross-sectional approach is used, there are other aspects of the shared savings arrangement that can adversely affect the measurement of savings. One of these aspects is the assignment methodology. Because these arrangements rely on passive assignment, the arrangement must specify a way to identify which patients should be assigned to the ACO. Some arrangements look back as far as 36 months to determine whether a patient should be assigned to an ACO. Others rely on a much simpler algorithm, such as assigning patients to an ACO for the prior calendar year based solely on their last evaluation and management (E&M) visit during that year. In this case, a patient visiting an ACO physician on December 31, despite having 10 E&M visits to non-ACO physicians earlier in the year, would be assigned to the ACO for the entire year. For obvious reasons, the "recent E&M assignment" methodology can cause patients to move in and out of the ACO often.

FIGURE 3: PROGRESSION OF COSTS FOR BENEFICIARIES ASSIGNED FOR FIRST TIME IN PERFORMANCE YEAR 3 (PY3)



The goal of the assignment methodology should be to assign members to the ACO physicians that are providing the majority of the member's primary care services. Relying on an overly simplistic approach can lead to problems. Consider the following illustrative example of the trouble that the "recent E&M visit" approach can have. A similar dynamic can also occur if the assignment methodology is based on the provider with the most E&M visits during the year, although the results might not be as exaggerated.

- Sample ACO enters into a shared savings arrangement for a Medicare ACO using the cross-sectional approach with patients assigned each year on the basis of the most recent E&M visit in that year. Without any interventions, Sample ACO expects to have a 1.0% annual trend.
- As Sample ACO starts its first performance year, it decides that one area of focus is to reduce expenditures for skilled nursing facility (SNF) services.
- In order to make an impact, its strategy is to direct patients to preferred SNFs that employ Sample ACO physicians, who are making a concerted effort to reduce length-of-stay and overall expenditures.
- This strategy proves to be successful, as the assigned patients who visit the SNF have 10% lower per capita expenditures than the assigned patients who visited the SNF in the prior year.
    - However, by shifting more patients to SNFs with Sample ACO physicians, patients in need of skilled nursing care are more likely to be assigned to Sample ACO because they are likely to receive E&M visits with Sample ACO physicians while in the SNF.
    - Therefore, in the performance year, Sample ACO winds up with SNF patients making up 3% of its population, compared with 2% in the prior year.
    - Because SNF patients tend to be more expensive overall, this raises the per capita expenditures for Sample ACO, despite the apparent improvements in efficiency that were made.

The chart in Figure 4 illustrates this issue.

FIGURE 4: TRENDS WITH AND WITHOUT SNF INITIATIVE

TIME PERIOD	NON-SNF PMPM	SNF PMPM	PORTION OF ASSIGNED LIVES IN SNF	TOTAL PMPM	TREND
Baseline	$800	$2,000	2%	$824	n/a
Performance (without initiative)	$808	$2,020	2%	$832	1.0%
Performance (with initiative)	$808	$1,800	3%	$838	1.7%

In many shared savings arrangements, there would be a risk adjustment that would increase the ACO's benchmark in this situation due to the fact that the ACO now has a less healthy population. However, it is unlikely that a risk adjustment based on health status would fully account for the known increased prevalence of expenditures for SNF services, unless the risk adjustment model explicitly incorporated variables beyond diagnoses and drug information.

Concurrent or prospective assignment?

The choice of whether to use concurrent or prospective assignment is not clear-cut, and the preferable option may vary based on the circumstances of the particular arrangement.

Prospective assignment, which is used in the Pioneer ACO program, is appealing to ACOs because it allows them to begin managing a particular set of patients at the start of the performance year without the risk of those patients being de-assigned during the year. Additionally, the ACO will not have any patients assigned that it was not aware of during the performance year. With concurrent assignment, which is used in MSSP, the ACO might receive quarterly or monthly lists of patients that are likely to be assigned, but the true list of patients is not final until the year is over.

The trouble with prospective assignment is that the ACO inevitably will be responsible for patients who do not see ACO physicians during the performance year. Patients may see ACO physicians regularly during a historical period, but because of a change in circumstances (perhaps the onset of a new condition), they see non-ACO physicians for their care during the performance year. Under prospective assignment, the ACO is still responsible for the expenditures for these patients, yet the ACO has little opportunity to manage their care. In the Medicare ACOs, it is not uncommon to see year-to-year turnover rates above 20%, meaning that a sizeable portion of patients are not seeing the ACO physicians they are assigned to for care during the performance year.

Ultimately this decision is a matter of preference. Concurrent assignment may produce a more meaningful estimate of the ACO's impact, but prospective assignment removes uncertainty about which patients the ACO is responsible for.

Conclusion and other considerations

There are certainly other considerations in designing a shared savings arrangement. For instance, there should be sufficient incentives to entice both efficient and inefficient providers to participate. Under the current Medicare Shared Savings Program regulations, it may be challenging for provider organizations that are already operating efficiently to succeed financially, because the savings are based on expenditure trends rather than the overall level of expenditures. Efficient providers have less wasteful utilization at the start of the program and therefore have less room to improve, whereas inefficient providers are often able to attain low trends simply by trimming some "low-hanging fruit."

One possible solution would be to adjust the benchmark trend that ACOs must beat based on the overall level of expenditures for the ACO. In this case, the benchmark trend would be decreased (more challenging) for providers that start the program with expenditures above a certain risk-adjusted level, while the benchmark trend would be increased (less challenging) for providers that start the program with expenditures below a certain risk-adjusted level.

With time, we have been able to observe many challenges with the measurement of savings in the shared savings arrangements in use today. We know that any method for attempting to measure savings will be imperfect in some way, but at the same time, we cannot let "perfect be the enemy of good." As changes are proposed to the current arrangements and as new arrangements are established, it is critical to perform simulation and modeling in advance to help avoid unintended consequences. This process provides the best chance for shared savings arrangements to have the intended effects of reducing expenditure levels while maintaining a high quality of care, thereby increasing the long-term viability of this payment model.

Jill Herbold, FSA, MAAA, is a principal and consulting actuary with the Indianapolis office of Milliman. Contact her at jill.herbold@milliman.com.

Anders Larson, ASA, MAAA is an associate actuary with the Indianapolis office of Milliman. Contact him at anders.larson@milliman.com.

FOR MORE ON MILLIMAN'S HEALTHCARE REFORM PERSPECTIVE

Visit our reform library at www.milliman.com/hcr
Visit our blog at www.healthcaretownhall.com
Or follow us on Twitter at www.twitter.com/millimanhealth