Robert Wood Johnson Foundation

APRIL 2016

Tricky Problems with Small Numbers: *Methodological Challenges and Possible Solutions for Measuring PCMH and ACO Performance*

Prepared by Nancy McCall and Deborah Peikes, Mathematica

Introduction

Accurately measuring the performance of patient-centered medical homes (PCMHs) and accountable care organizations (ACOs) is difficult when the number of patients they serve is small, due to the large, natural fluctuations in medical expenditures and service use. Providers are increasingly being rewarded through shared savings, performance-related bonus payments, or other financial incentives, but often payments are made without sufficient statistical safeguards in place to ensure the payments fairly reflect provider performance.

The **Centers for Medicare & Medicaid Services (CMS)** 2014 performance year results for its Pioneer and Shared Savings ACOs showed only one-quarter of ACOs slowing health care expenditure growth enough to earn shared savings bonuses. A particular concern expressed by ACOs with fewer than 5,000 attributed patients is that they are held to a higher level of savings than larger ACOs.[1] The primary reason CMS requires these ACOs to reach a higher minimum savings rate (MSR) is because a smaller number of attributed beneficiaries for performance measurement gives less confidence that ACO savings estimates reflect true savings and not random fluctuation that one typically observes in medical expenditures.[2]

Getting ACO and PCMH performance measurement right is a critical issue for purchasers and providers. Purchasers that participate in shared savings programs want to financially reward providers who generate true savings resulting from real improvement in care, and not reward providers who get lucky and show "savings" generated by random variation. Similarly, providers want a high level of confidence that they will be rewarded for true savings if they improve the quality and efficiency of care provided. On the other hand, providers also want protection from being unduly penalized if random variations produce unfavorable outcomes.

ABOUT STATE HEALTH AND VALUE STRATEGIES

State Health and Value Strategies, a program funded by the Robert Wood Johnson Foundation, provides technical assistance to support state efforts to enhance the value of health care by improving population health and reforming the delivery of health care services. The program is directed by Heather Howard at the Woodrow Wilson School of Public and International Affairs at Princeton University. For more informations, visit *statenetwork.org*.

ABOUT THE ROBERT WOOD JOHNSON FOUNDATION

For more than 40 years the Robert Wood Johnson Foundation has worked to improve the health and health care of all Americans. We are striving to build a national Culture of Health that will enable all Americans to live longer, healthier lives now and for generations to come. For more information visit *www.rwjf.org*. Follow the Foundation on Twitter at *www.rwjf.org/twitter* or on Facebook at *www.rwjf.org/facebook*.

ABOUT MATHEMATICA

Mathematica Policy Research seeks to improve public well-being by conducting studies and assisting clients with program evaluation and policy research, survey design and data collection, research assessment and interpretation, and program performance/data analytics and management. Its clients include foundations, federal and state governments, and private-sector and international organizations. To learn more, visit *http://www.mathematica-mpr.com/*.

This Issue Brief provides information to guide state purchasers when designing risk-based provider programs, including shared savings arrangements. The Brief first identifies and describes methodological challenges that arise when measuring performance within an organization with small numbers of patients, including:

1. small panel sizes and the variation in medical expenditures and utilization, and
2. small panel sizes requiring implausibly large performance achievements in order to demonstrate statistical significance

Next, this Brief notes an additional methodological challenge that arises when purchasers attempt to measure performance across a small number of organizations. This statistical challenge, related to small numbers of organizations in a clustered intervention, is relevant for states evaluating performance improvement across multiple PCMHs or ACOs.

Finally, the Brief provides concrete strategies and resources for state purchasers to address these methodological challenges when evaluating PCMH and ACO performance and applying financial incentives and disincentives. The following figures and table include definitions and resources relevant to these topics.

The problem with small numbers and variation in medical expenditures

Commercial insurers and state Medicaid programs are increasingly developing performance rewards, with one common example being shared savings programs. These programs have features similar to the Medicare Shared Savings Program (MSSP) that offers providers an opportunity to share a certain percentage of savings resulting from more efficient provision of care, or "gainsharing."[3] These programs can be one-sided, offering providers a reward if costs for care are below targets, or two-sided, with a requirement that providers also assume some percentage of financial risk, or "downside risk," should medical expenditures be higher than expected.[4]

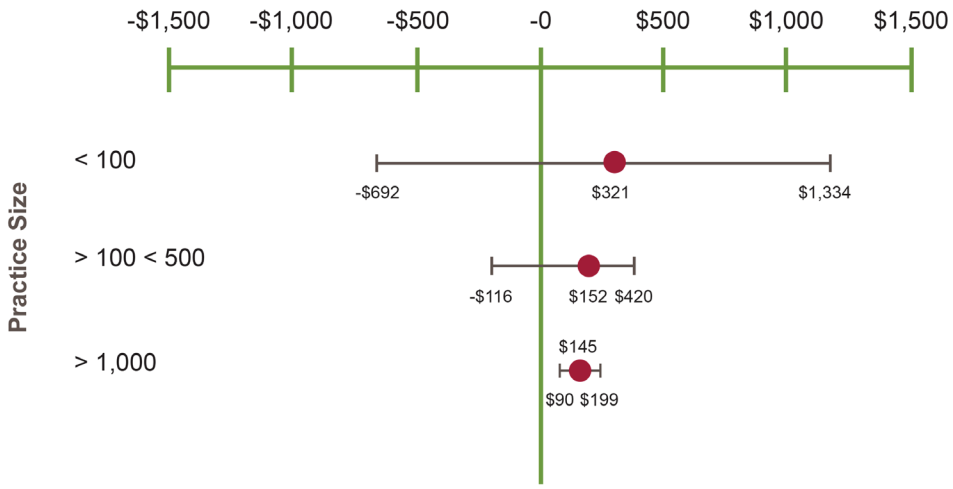
To estimate average savings per person, the majority of shared savings programs assess providers' performance relative to an expected target or a comparison group's performance,[5] but the existence of random variation in medical expenditures leads to statistical uncertainty in measuring the savings rate. This uncertainty is greater for provider practices that have a smaller number of attributed patients. The probability of an incorrect performance reward or penalty is heavily dependent on ACO panel size.[6] The issue isn't just that the provider entity is small (i.e., has few patients), but also that an individual payer or health plan might have few patients attributed to the provider entity.

Uncertainty in measuring the savings rate is also greater for ACOs or practices that serve patients with more diverse medical needs. Larger variation in medical expenditures leads to less precision in estimating savings.[7]

Figure 1 (see following page) provides a visual example of how the degree of precision varies across provider organizations with different numbers of attributed lives. The brackets indicate 80 percent confidence intervals[8] around the mean estimate of no savings, or cost neutrality for a Medicare demonstration testing care coordination.[9] In this figure, all three provider organizations appear to have increased per patient per month Medicare expenditures (red dots) ranging from $145 to $321. However, the confidence intervals drawn around the point estimates of program effects show varying degrees of uncertainty around these estimates; the amount of uncertainty corresponds with the sample size and variance in medical expenditures of each provider's patients. Not surprisingly, for the first provider organization with fewer than 100 patients, the confidence interval of estimated savings is extremely wide, ranging from -$692 to +$1,334. In contrast, the degree of uncertainty is much less for the third provider organization with greater than 1,000 patients, leading one to have a high degree of confidence that this third provider's approach was not cost neutral but rather cost increasing.

Variation in expenditures and other outcomes can lead to two types of measurement errors: (1) false positives, whereby no savings actually occurred but there is observed savings and providers erroneously receive shared savings; and (2) false negatives, whereby there are true savings but there is no observed savings and providers erroneously receive no shared savings, or, if there is two-sided risk, they face a financial penalty. Purchasers often protect themselves from random variation by setting a threshold of required savings before making shared savings payments. The higher the threshold, the greater the protection against inappropriate payments for purchasers, but the bigger the risk that smaller provider organizations may not be paid for saving money, particularly at lower rates of true savings. In the MSSP, CMS dealt with uncertainty by setting of a minimum savings rate (MSR) threshold that must be surpassed before savings are considered "real." For the one-sided MSSP model, the MSR varied inversely by ACO size reflecting greater statistical uncertainty with smaller patient panels. For the two-sided MSSP model, the MSR and a minimum loss rate (MLR) threshold were set at two percent. This MSSP approach

Figure 1: 80% Confidence Interval Around Estimates of Cost Neutrality by Practice Size



means that ACOs will pay a financial penalty if actual expenditures exceed the threshold by two percent or more, and they will share in savings over two percent.

Figure 2 (see following page) displays probabilities, by the number of lives attributed to Medicare MSSP ACOs, that an ACO would pay a financial penalty when true savings are zero (a false positive).[10] **Figure 3** (see following page) displays probabilities, by the number of lives attributed to Medicare MSSP ACOs, that an ACO would not receive shared savings for true savings of three percent (a false negative). In each instance, smaller ACOs have a higher probability than larger-sized ACOs of erroneously paying a penalty or not receiving true savings. ACOs with only 5,000 attributed patients have a 26 percent probability of paying a financial penalty when true savings are zero, in contrast with a two percent probability of paying a penalty for ACOs with 50,000 patients. The risk of false positives declines as the actual savings rate increases, but the pattern of small ACOs facing larger risk of a penalty despite savings persists. When true savings are three percent, all ACOs face a substantial risk of not being rewarded for true savings, but ACOs with 5,000 patients have a 37 percent probability of not sharing in the actual savings, in contrast to a 15 percent probability for ACOs with 50,000 patients.

Two factors drive the higher probability of false positives and false negatives for smaller providers: large variation in medical expenditures and a small number of attributed patients lead to less precision in the estimate of savings.

The problem with small numbers requiring large performance achievements

Evidence from early evaluations showed that PCMHs and ACOs can have an effect on slowing the rate of growth in health care expenditures. This evidence suggests that optimistically, one might expect to find statistically significant savings among PCMHs and ACOs in the range of five percent for a general population, and upwards to 15 percent for a chronically ill population.[11] However, many more recent studies of these types of organizations have shown no positive findings, or positive findings for only a few organizations. Of the 32 Pioneer ACOs, just over 50 percent showed statistically significant lower rates of growth in spending in the first year.[12] The lack of consistent positive findings could be the result of the model not working, organizational or implementation factors rendering the intervention ineffective, or statistical issues related to small samples requiring implausibly large performance achievements to be statistically significant. In designing shared savings programs and setting savings thresholds, estimating the minimum detectable effect (MDE) ensures that the savings thresholds can be detected. **The MDE is defined as the smallest size of the "real" effect that can be detected for a given statistical significance level and power level.** For any intervention, a smaller MDE is more desirable.[13]

Figure 2: Probability Medicare MSSP ACOs Would Pay Financial Penalty When True Savings Are Zero by ACO Size

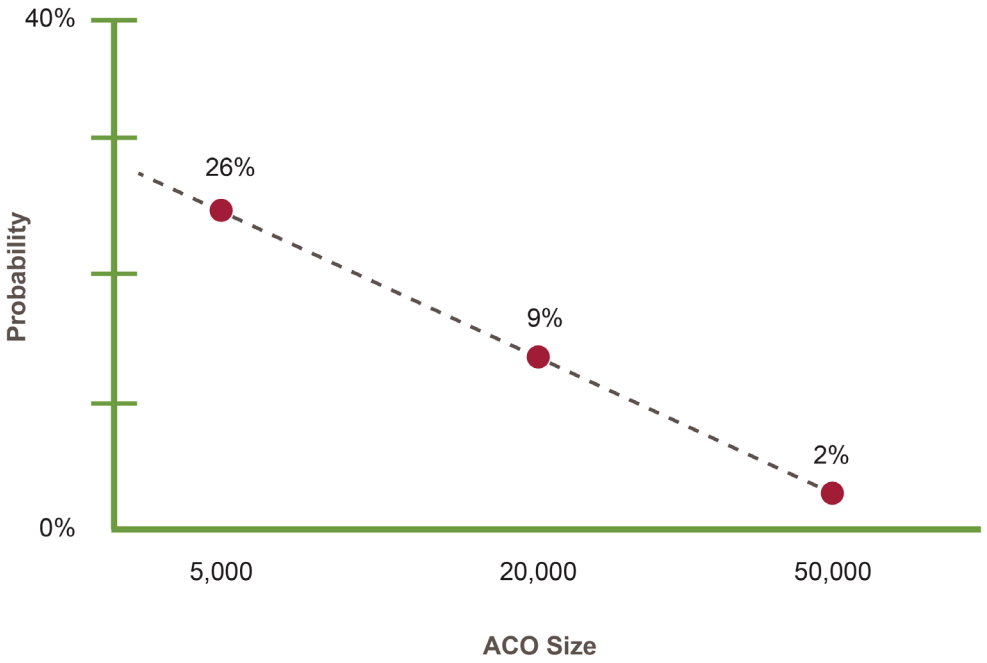


Figure 3: Probability Medicare MSSP ACOs Would Not Receive Savings Rewards for 3% Savings by ACO Size

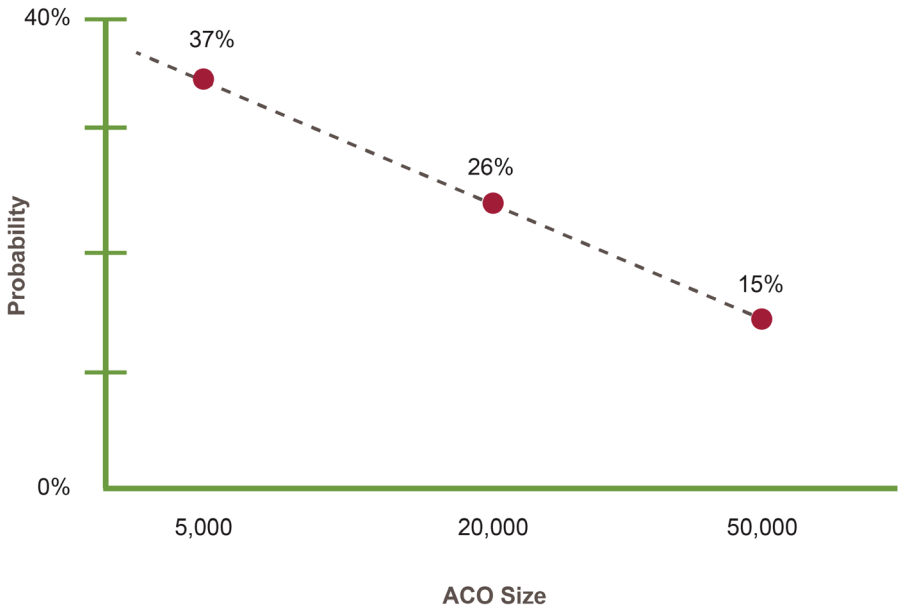


Table 1 shows the importance of considering attributed panel size and the health or disability status of patient populations in PCMH/ACO panels when determining whether or not the resulting savings (or costs) are due to actions by the provider entity. The table displays the Minimum Detectable Effect of savings estimates by number of attributed patients in a provider panel using a desired statistical significance level of 10 percent and 80 percent power. This table is based on sample medical expenditure data for certain commercial, Medicaid, and Medicare populations.

Table 1 shows two alternative statistical approaches commonly used to estimate the average per patient savings rate: an interrupted time series (ITS) model, and a difference-in-differences (D-in-D) model:[14]

- The ITS approach is used to calculate the difference between an observed average per patient medical expenditure estimate during a performance year, and a baseline average expenditure estimate trended forward to the performance year using known changes in plan benefits, patient case-mix, and payment rates.
- The D-in-D approach is used to calculate the difference in the rate of growth from a baseline year to a performance year between the provider's average per patient medical expenditure and a comparison group's average per patient expenditure.

The ITS approach is most common in evaluating ACO performance, where a credible comparison group is difficult to select; while the D-in-D approach is most common in evaluating PCMH performance. It is generally believed that the D-in-D approach is a stronger financial performance measurement technique, and MDEs are considerably smaller using the D-in-D model because the D-in-D sample size is double that of the ITS model. Further, with only two baseline years included in the ITS model, the variance of the trended outcome greatly increases the variance of the impact estimate, and therefore, the MDE.

Based on the illustrative data in Table 1, for savings estimates to be found statistically significant:

- A provider organization with attributed panel sizes of at least 10,000 privately insured patients would need

Table 1: Illustrative Examples of Minimum Detectable Effects (MDE) of Percentage Savings by Payer, Health Status, and Panel Size

	Privately Insured		Medicaid		Medicare	
	All Patients	Chronically Ill	All Adults	Disabled	All Patients	Chronically Ill
Interrupted time-series (ITS) model against a benchmark						
Attributed Patients	(3.73)	(2.5)	(4.2)	(1.09)	(2.46)	(1.64)
1,000	88%	59%	99%	26%	58%	39%
5,000	39%	26%	44%	11%	26%	17%
10,000	28%	19%	31%	8%	18%	12%
20,000	20%	13%	22%	6%	13%	9%
Difference-in-Differences (D-in-D) model with comparison group						
Attributed Patients						
1,000	31%	21%	35%	9%	20%	14%
5,000	14%	9%	16%	4%	9%	6%
10,000	10%	7%	11%	3%	7%	4%
20,000	7%	5%	8%	2%	5%	3%

Notes: (coefficient of variation)

to reduce expenditures by 28 percent (relative to the average expenditures in the baseline period) using an ITS trending approach, and by 10 percent using the D-in-D approach.

- A provider organization with an attributed panel size of at least 20,000 privately insured patients would need to reduce expenditures by 20 percent (ITS), and by seven percent (D-in-D), for savings estimates to be found statistically significant.

However, for savings estimates to be found statistically significant using an ITS trending approach, a provider with only 1,000 privately insured attributed patients would need to reduce expenditures by 88 percent. The large percentage savings is required because of a high coefficient of variation (CV) and a small sample size. Using a D-in-D approach, a provider with 1,000 privately insured attributed patients would need to reduce expenditures by 31 percent for savings estimates to be found statistically significant. Demonstrating either level of savings in medical expenditures (88 or 31 percent) far exceeds plausible levels for PCMHs or ACOs.

Providers with 1,000 attributed adult Medicaid patients (including but not limited to persons with disabilities) would also have to achieve implausibly large savings effects for a high degree of confidence that the effect is real, such as a 35 percent reduction in costs using a D-in-D approach shown in Table 1. Consequently, state purchasers should consider establishing a much larger minimum panel size and take statistical power (that is, the size of the MDE) into account when determining thresholds for PCMHs or ACOs to share in savings or risk. States and providers should know with reasonable certainty whether any savings or losses are likely due to the provider organization's performance rather than random variation in medical expenditures. Without this level of confidence, a provider organization could see huge "savings" one year and huge "losses" the next, both of which may not reflect their actual performance.

In addition, when restricting performance measurement to more homogenous and sicker population panels, reflected in a smaller CV, MDEs are considerably smaller. For example, a PCMH provider with 1,000 Medicaid attributed patients with disabilities would need to reduce expenditures by 26 (ITS) and nine percent (D-in-D), respectively, for savings estimates to be found statistically significant.

The problem with small numbers of provider organizations in initiatives

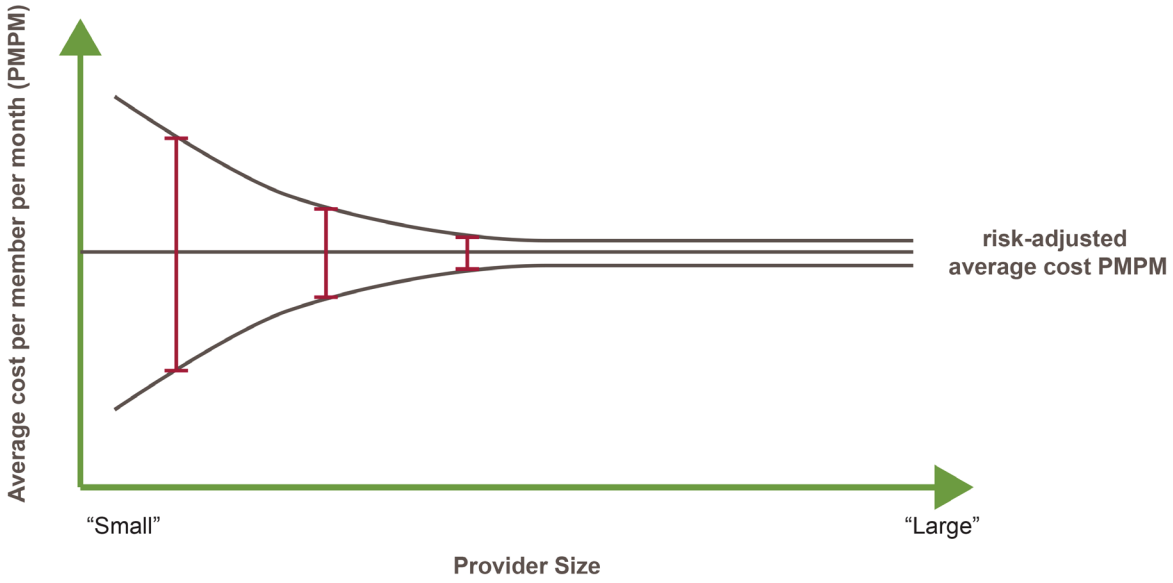
When designing and measuring performance **across** multiple provider organizations, the problem with small numbers becomes one of too few provider organizations included in the assessment, rather than too few patients. Outcomes of patients within the same provider organization are somewhat similar to each other; therefore, each additional patient in a provider organization does not provide completely new information about outcomes. In other words, each patient is not an independent observation. Patient survey ratings of the availability of after-hours care at primary care practices can illustrate this point. Whether 50 or 500 patients from each practice respond to the survey, the average access ratings for each practice are unlikely to change much. In practice, the amount of "clustering" within and across provider organizations varies widely for different provider types, settings, and outcomes, and can only be known by looking at the data.

Accounting for clustering is critical when designing provider-level interventions that transform care within an entire provider organization. The lack of independence of patient outcomes has two effects: (1) it reduces the effective sample size,[15] or the number of unique observations; and (2) it understates the standard errors of estimates. A reduction in the *effective* sample size means there is less confidence in an estimate, and it increases the size of the effect that providers must achieve to be statistically significant (the MDE). Failing to adjust analyses for clustering of outcomes increases the likelihood of a false positive finding, and it understates the statistical uncertainty.

Strategies

When designing risk-based provider contracts, state purchasers should develop a balanced approach to setting and measuring performance targets, in order to: (1) increase the likelihood that the provider's performance is accurately captured; and (2) minimize the likelihood of false savings or losses for both purchasers and providers.[16] This Issue Brief has focused upon challenging problems that arise with small numbers, as displayed in **Figure 4** (see following page). In this diagram, the center line represents the risk-adjusted average cost per member per month (PMPM), and the two curved lines around it represent the upper and lower bounds

Figure 4: Hypothetical Confidence Intervals Around a Risk-adjusted Cost Per Member Per Month by Provider Size



of the 90 percent confidence interval around the average PMPM estimate. As the number of providers' attributed lives increases, the confidence intervals narrow to be closer to the savings estimate, and the random chance of an error declines.

Next, this Brief offers two sets of strategies to ensure accurate performance measurement across providers of different sizes. State purchasers and providers should consider these strategies when designing and accepting risk-based provider contracts. The first set of strategies focuses on reducing provider-level variation in expenditures and savings estimates. The second set focuses on how to account for variation in health care expenditures when determining if true savings occurred across providers with different numbers of attributed lives.

Strategies to reduce variation in expenditures and savings estimates

A. Assess prior variation in expenditures. Actual variation in expenditures is often *unknown* before providers enter into risk-based contracts. Important steps to take before starting an at-risk contract include: (1) analyze two years of baseline claims data (that is, data before the contract begins) to obtain information on the mean and variance of expenditures; and (2) construct a confidence interval around each provider's mean annual expenditures, for all patients and for important subgroups of patients (e.g., those with chronic conditions).

B. Perform test calculations to determine if savings level is plausible. Use baseline data to estimate the minimal detectable savings rate for each provider to determine if the number of attributed lives is sufficiently large to detect a savings rate they might plausibly achieve. The Maryland Multi-payer PCMH program performed test calculations of cost savings before launching its program.[17]

C. Reduce variance in expenditures. Reducing variance in expenditures increases precision of estimates and reduces the risk of false positives and false negatives. A number of strategies should be considered to reduce variance:

a. Truncate extreme values of per capita expenditures (e.g., at $100,000) in the baseline and performance year periods, as is done in the MSSP, or exclude catastrophic cases (e.g.,

organ transplants) as another way to reduce the variation that arises from the inclusion of extreme values.

b. Use the same patients to measure outcomes in baseline and performance periods, which will reduce variance in expenditures. One strategy preferred by providers is prospective attribution, whereby providers are at-risk during the performance year for patients they had previously treated. This allows them to know in advance the patients with whom they must intervene to yield a meaningful change in performance.

D. Consider covering all patients served by the provider but measuring effects among higher-risk patients. While larger sample sizes will generally increase the statistical precision around the savings estimate, including all of a provider's patients in performance measurement can often have the opposite effect of increasing expenditure variance. Restricting performance measurement or risk-based contracting to a smaller, more costly group of patients will likely reduce the variance. It will also increase the likelihood of detecting a true effect.[18] This approach is most feasible to implement with a prospective attribution of lives, and the use of a predictive risk-score algorithm such as the Chronic Illness and Disability Payment System (CDPS), Diagnostic Cost Groups (DCGs), or Adjusted Clinical Groups (ACGs),[19] to identify those at highest risk of future expenditures.

E. Use Risk adjustment. Risk-adjustors should be used to ensure adequate adjustment of expenditures in the performance period, even in a prospectively assigned set of patients. Case-mix can vary substantially across providers. In addition, risk-adjustors help account for changes in attributed patients' health status over time that may increase expenditure variance in a practice or ACO's attributed panel.

F. Group small providers together for purpose of performance measurement. As shown in Figure 4, the variance in expenditures decreases as sample size increases. One strategy for state purchasers to consider is pooling across small providers to increase the sample size available to assess performance. Before doing so, it is important to use baseline data to determine if pooling across providers reduces expenditure variance. State purchasers structuring such an arrangement should also: (1) determine which providers to pool; (2) obtain provider buy-in on how and when to pool performance of different organizations; and (3) determine the methodology for how to share savings or financial penalties across practices and solicit input from involved provider organizations.

G. Adjust standard errors for clustering. Accounting for clustering in federal evaluations of PCMH and ACO initiatives that involve multiple providers has become standard practice.[20] However, many studies of initiatives tested by states and private payers have not adjusted for clustering, perhaps because of resource constraints. These initiatives sometimes have too few providers involved, and adjusting for clustering would make it harder to demonstrate statistically significant findings. The degree of clustering should be directly calculated from baseline data with common statistical packages[21] during the design phase, and outcomes should be calculated accounting for clustering to the extent that purchasers are measuring performance among multiple providers. Doing so will help minimize erroneously calculating savings when there are none, or no savings when there are actual savings.

Strategies to account for random variation in health care expenditures in determining if true savings occurred

A. Set a minimum threshold for the number of attributed lives before including small providers in a shared savings arrangement. If test results using baseline data show that the number of attributed lives is too small to detect an effect size that is plausible to produce, the arrangement should either exclude small practices, require them to band together to achieve larger samples, or allow smaller organizations to remain in one-sided risk contracts.

B. Directly account for random variation in determining if the savings rate is achieved and the percentage of savings shared. Two strategies are worthy of consideration:

a. Remove variation in baseline spending growth. DeLia[22] recommends making the growth in baseline expenditures a fixed number (i.e., setting a three percent growth in expenditures), rather than basing growth on an estimate from a sample that contains random variation (i.e., measuring expenditure growth in a contemporaneous comparison group, or adjusting baseline spending by the national growth rate).

b. Adjust shared savings proportion by probability of true savings. Weissman and colleagues

developed a shared savings approach for the Massachusetts Patient-Centered Medical Home Initiative[23] whereby the proportion of savings shared was adjusted to reflect the probability that providers achieved true savings. This was calculated by multiplying the proportion of savings that was to be shared by the provider-specific probability of achieving true savings (1 minus the probability of achieving shared savings as a result of chance) for a fixed threshold of savings, (e.g., one percent, two percent). This allows for a "sliding-scale" of the proportion of savings to be shared.

Conclusion

It is important for states to incentivize provider organizations by fairly rewarding their performance. This Brief offers purchasers and providers concrete strategies that can help when designing and accepting risk-based provider contracts for smaller populations. With the goal of improving health outcomes and cost-efficiency of health services, these risk-based contracting strategies are designed to (1) offer clear financial rewards for true savings and improved performance; and (2) protect providers from penalties related to random variations in medical expenditures and service use.

Endnotes

[1] McClelland M et al. *Health Policy Issue Brief: How to Improve the Medicare Accountable Care Organization (ACO) Program*. Engelberg Center for Health Care Reform at Brookings, June 2014. http://www.brookings.edu/research/papers/2014/06/16-medicare-aco-program-changes.

[2] Medicare Share Savings Program: Shared Savings and Losses and Assignment Methodology Specifications. Centers for Medicare and Medicaid. Version 3. December 2014. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/Shared-Savings-Losses-Assignment-Spec-v2.pdf.

[3] ibid.

[4] DeLia D, Hoover D, and Cantor J. "Statistical Uncertainty in the Medicare Shared Savings Program. *Medicare and Medicaid Research Review*. 2(4), 2012. Technical Appendix: http://dx.doi.org/10.5600/mmrr.002.04.s04.

Weissman J, Bailit M, D'Andrea G, and Rosenthal M. The Design and Application of Shared Savings Programs: Lessons from Early Adopters. *Health Affairs*. 31(9):1959-1968, 2012.

Bailit M, Hughes C. *Key Design Elements of Shared Savings Payment Arrangements*. The Commonwealth Fund, August 2011.

[5] There are several approaches commonly used to estimate the average per patient savings rate. One common approach is to calculate the difference between an observed average per patient medical expenditure estimate during a performance year and a baseline average expenditure estimate trended forward to the performance year using known changes in plan benefits, patient case-mix, and payment rates. A second common approach is to calculate the difference between an observed average per patient medical expenditure estimate during the performance year and a comparison group's average per patient expenditure.

[6] DeLia D, Hoover D, and Cantor J. "Statistical Uncertainty in the Medicare Shared Savings Program. *Medicare and Medicaid Research Review*. 2(4), 2012.

[7] The degree of natural variation observed in medical expenditures is often reported using the coefficient of variation (CV). CVs will be largest when the patient panels include all patients, and smallest when the panels include only the chronically ill, who have more similar patterns of health care expenditures. CVs also tend to be largest for small panels, and smallest for large panels. While the amount of variation will differ across providers depending on patient case-mix and practice patterns, a review of findings from early PCMH initiatives reported average medical expenditure CVs ranging from 1.64 for chronically ill Medicare beneficiaries to 3.73 for all privately insured Medicare beneficiaries, [Peikes D, Dale S, Lundquist E, Genevro J, and Meyers D. *Building the evidence base for the medical home: what sample and sample size do studies need?* White Paper (Prepared by Mathematica Policy Research under Contract No. HHSA2902009000191 TO2, AHRQ Publication No. 11-0100-EF Rockville MD: Agency for Healthcare Research and Quality, October 2011). https://pcmh.ahrq.gov/sites/default/files/attachments/Building%20Evidence%20Base%20PCMH%20White%20Paper.pdf].

[8] An 80% confidence interval means that 80 percent of all possible 80% confidence intervals will contain the "true" estimate of savings in the range displayed, if one were to calculate savings many times. For more precise "true" estimates, a higher level of confidence can be used, e.g., 90% or 95%.

[9] Brown R, Peikes D, Chen A, Ng J, Schore J, and Soh C. *The Evaluation of the Medicare Coordinated Care demonstration: Findings for the First Two Years*. Evaluation Report (Prepared by Mathematica Policy Research under Contract No. 500-95-0047 TO09). Centers for Medicare & Medicaid Services. Baltimore, MD. March 2007. Confidence intervals from the evaluation of the first two years of the Medicare Coordinated Care Demonstration by size of participating organization.

[10] Figures 2 and 3 were developed using data from Exhibit 4, Financial risks and expected income for ACOs with varying levels of true savings, from DeLia, D Hoover D, and Cantor J. (2012) DeLia et al. calculated probabilities for hypothetical ACOs with different numbers of attributed lives participating in the Medicare MSSP two-sided model. An average savings rate is estimated as the difference between a risk-adjusted performance year per capita spending level and a 3-year weighted risk-adjusted baseline per capita spending level trended forward by projected per capita national growth in Medicare expenditures.

[11] Peikes D, Dale S, Lundquist E, Genevro J, and Meyers D. *Building the evidence base for the medical home: what sample and sample size do studies need?* White Paper (Prepared by Mathematica Policy Research under Contract No. HHSA290200900019I TO2). AHRQ Publication No. 11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality. October 2011.

Nyweide D, Wolton L, Cuerdon T, et al. "Association of Pioneer Accountable Care Organizations versus Traditional Medicare Fee for Service with Spending, Utilization, and Patient Experience." *Journal of the American Medical Association*. 2015: 313(2): 2152-2161.

[12] ibid.

[13] Key factors that affect the size of the MDE are 1) the coefficient of variation, 2) the desired level of statistical significance and power, 3) the amount of clustering of patient outcomes in provider organizations, and 4) the number of patients and provider organizations. Reducing the variance of estimates, reducing clustering, and increasing the number of provider organizations (and patients if there is not much clustering) are advantageous because they reduce the MDE.

[14] The minimum detectable effects (MDEs) in Table 1 were calculated assuming two designs for determining changes in expenditures, or savings, during a one-year performance period relative to a two-year baseline period. The top half of Table 1 provides MDEs using an interrupted time-series (ITS) model which compares performance of a single provider to a forecasted trend benchmark. The bottom half of Table 1 provides MDEs using a difference-in-differences (D-in-D) model which compares performance of a single provider to an equally sized contemporaneous comparison group. The MDEs are considerably smaller using the (D-in-D) model because the D-in-D sample size is double the ITS model. Further, with only two baseline years, the variance of the trended baseline outcome greatly increases the variance of the impact estimate and, therefore, the MDE. The following assumptions are the same in both models: (1) intraclass correlation coefficient (ICC) = 0, or no clustering of outcomes within provider; (2) coefficient of variation (CV) for commercial insurance and Medicare patients is taken from ranges observed in the literature evaluating patient-centered medical home initiatives (Peikes et al) and for disabled Medicaid patients from Stacy Dale., and Carol Irvin. "SSDI Beneficiaries Medicaid Expenditures: During the Waiting Period and Beyond." Princeton, NJ: Mathematica Policy Research, October 2006, and from Mathematica internal analyses of a broad cross-section of adult Medicaid patients' 2013 expenditures; (3) the models include person-level covariates and the amount of variation in the outcomes explained by person-level covariates is 0.50; and (4) the models exclude group-level covariates. In the D-in-D design, the correlation between the baseline and performance-year outcomes is 0.50. Further, the correlation between the individual-level covariates in the regression model and the treatment-group indicator variable is 0.10, which corresponds to a high-quality match between the treatment and comparison groups.

[15] As shown in Peikes et al., the effective sample size is the total number of patients reduced by the intraclass correlation coefficient (ICC) and number of patients per provider (n):

Effective Sample Size = Actual sample size/((1+ICC*(n-1))

With an ICC of 0, the number of unique observations is the number of patients. With an ICC of 1, the number of unique observations reduces to the number of provider organizations included in the study. Peikes et al. also showed that with 20 provider organizations (10 in the intervention group and 10 in a comparison group), the minimum detectable effect (MDE) for reducing total cost of care rises from 5% to 66% when there is a moderate degree of clustering. This is an implausibly large reduction in costs required before it is likely to be considered statistically significant.

[16] McClellan M. et al. Issue Brief: How to improve the Medicare Accountable Care Organization (ACO) Program. Engelberg Center for Health Care Reform at Brookings. June 2014.

Weissman J, Bailit M, D'Andrea G, and Rosenthal M. "The Design and Application of Shared Savings Programs: Lessons from Early Adopters." *Health Affairs*: 31(9):1959-1968, 2012.

Bailit M, Hughes C. *Key Design Elements of Shared Savings Payment Arrangements*. The Commonwealth Fund, August 2011.

DeLia D. "Leaving it to chance: the effects of random variation in shared savings arrangements." *Health Services Outcomes Research Methods*. 13:219-240, 2013.

Mann C. Center for Medicaid, CHIP and Survey & Certification, Centers for Medicare & Medicaid Services, Department of Health & Human Services. Letter to State Medicaid Directors. Shared Savings Methodologies. August 30, 2013.

[17] Bailit M, Hughes C, Burns M, and Freedman D. *Shared-Savings Payment Arrangements in Health Care: Six Case Studies*. The Commonwealth Fund, August 2012.

[18] Peikes et al. October 2011.

[19] Weir S, Aweh G, and Clark R. Case Selection for a Medicaid Chronic Care Management Program. *Health Care Financing Review*. 2008. 30(1): 61-74.

[20] McCall N, Haber S, van Hasselt M, Cromwell J, Sorensen A, Farrell K, et al. *Evaluation of the Multi-Payer Advanced Primary Care Practice (MAPCP) demonstration: First Annual Report*. Prepared for the Centers for Medicare & Medicaid Services. RTI International, April 2013.

Taylor E, Dale S, Peikes D, Brown R, Ghosh A, Crosson J, Anglin G, Keith R, Shapiro R, et al. *Evaluation of the Comprehensive Primary Care Initiative: First Annual Report*. Prepared for the U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services. Mathematica Policy Research, January 2015.

Evaluation of the CMMI Accountable Care Organization Initiatives: Pioneer ACO Evaluation Findings from Performance Years One and Two. Prepared for the Centers for Medicare & Medicaid Services. L&M Policy Research LLC, March 2015.

[21] Intraclass correlations (ICCs) can be calculated using several computation packages such as SAS and STATA. An example of how to use each of these statistical packages to calculate ICCs is found in Appendix F of Peikes D, Dale S, Lundquist E, Genevro J, Meyers D. *Building the evidence base for the medical home: what sample and sample size do studies need?* White Paper (Prepared by Mathematica Policy Research under Contract No. HHSA290200900019I TO2). AHRQ Publication No. 11-0100-EF. Agency for Healthcare Research and Quality, October 2011.

[22] DeLia D. "Leaving it to chance: the effects of random variation in shared savings arrangements." *Health Services Outcomes Research Methods*. 13:219-240, 2013.

[23] Weissman J, Bailit M, D'Andrea G, and Rosenthal M. The Design and Application of Shared Savings Programs: Lessons from Early Adopters. *Health Affairs*. 31(9):1959-1968, 2012. http://content.healthaffairs.org/content/31/9/1959.abstract.