Chapter 9
Program Evaluation: Defining and Measuring Appropriate Outcomes

Peter A. Hollmann

The major legislation to expand healthcare insurance coverage has failed. Health care expenditures are not sustainable. They are the fastest growing part of the federal budget and threaten the stability of our national economy. Household economies are also threatened with bankruptcy due to medical costs. Our volume based payment system results in a built hospital bed being a filled hospital bed and many unnecessary procedures. Conversely, there are rampant gaps in care with the standards of care being met only half of the time. Medical errors kill thousands. Our educational and training system does not produce the workforce we need. Our technology enamored country diffuses unproven technology as professionals and institutions engage in arms races over the newest devices. As healthcare consumes a growing portion of the gross domestic product, it limits our ability to spend on other worthy areas such as better housing, infrastructure and education – which may actually contribute more to population health than healthcare does. As the cost of health insurance becomes a greater proportion of employee costs, even potentially eclipsing wages, employees are afraid to change jobs, employers drop coverage and America's products become non-competitive in global markets. Our manufacturing base declines, the middle class erodes and there is economic polarization as the American dream slips away from too many. Without a doubt, the landmark legislation signed into law by our President is a failure.

The president is Lyndon Baines Johnson. The year is 1965 and the law is Medicare.

Most of the healthcare professionals, economists and experts of that day have gone on to become Medicare beneficiaries and died. The debate of that era over Medicare was not that different from today's debate over the Affordable Care Act (ACA). Who today really believes Medicare is a failure? If the belief is that it is a

P.A. Hollmann, M.D. (✉)
Division of Geriatrics, Blue Cross & Blue Shield of Rhode Island,
500 Exchange Street, Providence, RI 02903-2699, USA
e-mail: Peter.Hollmann@bcbsri.org

J.S. Powers (ed.), *Healthcare Changes and the Affordable Care Act*,
DOI 10.1007/978-3-319-09510-3_9

success, even if a flawed success, what measures define success or failure? The problems outlined above all exist today, even if some would choose to debate fine points. Undeniably, Medicare has played a major role in shaping the healthcare system and country that we have today. The access to healthcare for seniors and those with conditions such as End Stage Renal Disease would be markedly diminished without Medicare. Healthcare today is much more effective than in 1965 and has played a role in extending the average life expectancy. The addition of Medicaid greatly enhanced access to care for children and those with disabilities and those needing long-term care. In thinking about whether Medicare has been a success or failure or both, we can conceptualize its evaluation because it is a familiar subject. It helps us understand the process and challenges of evaluating the ACA. It also reminds us that regardless of technical accuracy, scholarly research and statistical prowess, in the end public opinion may be the only evaluation that truly matters.

The ACA addresses health insurance, healthcare quality and payment methods designed to promote quality. It builds upon activities already in progress at state and/or federal levels. The goal is to achieve the "Triple Aim" and, in doing so, create a stronger and better America in ways that go beyond health. The Triple Aim has been phrased in slightly different ways at different times and by different users, but is:

- Improving the patient experience of care (including quality and satisfaction);
- Improving the health of populations; and
- Reducing the per capita cost of health care.

What measures will be appropriate to define success in this goal is a complex question. The purpose of this chapter is to explore measures, measurement and use of measurement. The focus is on healthcare quality, but the measures of quality derive from greater goals related to health and economics. Ultimately, it is the intended or eventual use of measurement that matters most. It is the use that will both drive improvement and drive debate. Current measures and measurement are inadequate to the task of providing definitive answers to most meaningful questions ranging from the efficacy of a massive piece of legislation such as the ACA or the quality of care provided by an individual clinician. The process of seeking how best to quantify and promote success will, in itself, be an exercise in quality improvement that must be undertaken if we are to advance our goal of achieving the Triple Aim. The approach used in this chapter is one that is designed to create an overview for clinicians. It is not written for the expert analyst or statistics authority, who will likely recognize some liberties taken for the sake of providing general information.

The Affordable Care Act

The ACA might be boiled down into having two basic goals: increase access to health care by changing health insurance and improve the value of healthcare. Health is a critical attribute of happiness and wellbeing. It is also a critical attribute of a productive society. Health is affected by genetics, habits/lifestyle, medical care, the environment, wealth, education and many interacting factors. Accordingly, in

the vision of Barack Obama, the impact of the ACA is to extend well beyond health, health insurance and quality of care. The President's words of March 5, 2009, place the ambition of the ACA – and therefore, one could argue, the standards by which its success is to be measured – in an expansive social and economic context:

> At the fiscal summit that we held here last week, the one thing on which everyone agreed was that the greatest threat to America's fiscal health is not Social Security, though that's a significant challenge; it's not the investments that we've made to rescue our economy during this crisis. By a wide margin, the biggest threat to our nation's balance sheet is the skyrocketing cost of health care. It's not even close.

Consider the breadth of metrics that could be used to define the success of the ACA in light of this broad vision. Not only will the health of populations be used to define quality and efficiency of care by providers and the payment the providers receive, the health of the population will define the successes and failures of the ACA itself. Here are some provisions of the law followed by some potential or current measures by which the success of those provisions might be assessed:

Increase the number of individuals with health insurance by providing access to coverage, financial support and a personal mandate for coverage: the percent of the population with insurance coverage that includes essential benefits; the percent of those at specific income levels with coverage; the percent of younger individuals who purchase coverage through an exchange; the number of businesses that increase or drop employer sponsored coverage; percent of family income going to healthcare; the rate of medical cost driven personal bankruptcy.

Expand eligibility for Medicaid through federal support of state initiatives: The number of newly insured; the number of conversions of private coverage to Medicaid; the financial stability of providers as Medicaid expands; the number of providers accepting Medicaid; state budget surplus or deficit.

Create a competitive marketplace with specific ground rules such as essential benefits, ending lifetime or annual caps and pre-existing condition exclusions and using risk adjusted payments to health plans: market choice; premium stability over time; customer satisfaction and plan stability in enrollments; less "cherry picking" (i.e., tactics to avoid adverse selection such as excluding providers with complex patients from the network).

Improve value by paying for quality or penalizing undesirable outcomes, such as the Medicare Advantage 5 Star Program or Hospital Acquired Conditions penalties: improvements in the quality measures that are used in these programs; improvements in quality measures that are not used in payment; market share of higher performing organizations; beneficiary choice; stability of safety net organizations; reduction in the growth of the rate of the portion of the gross domestic product (GDP) and federal budget spent on health care.

Require first dollar coverage for preventive services: the percent of the population that receives the recommended service; reductions in the target illness morbidity and mortality; reductions in the cost of care for the targeted conditions; fewer days of disability or missed work; increased average life expectancy; increased average life expectancy and decreased disability for lower socioeconomic status populations; fewer unintended pregnancies.

Promote rapid diffusion of cost effective care and system redesign through the creation of programs related to comparative effectiveness and innovation: lower total cost of care trends; optimal care is better defined and new measures of care are defined; population health status improves; hospital readmissions are reduced.

Each listed measure could also be joined by broader measures that reflect the social context of healthcare. For example, better physical and mental health could improve employee productivity and the growth of the GDP. Less money spent on healthcare and fewer medical bankruptcies could mean that affordable housing receives greater attention, more people have rent money and homelessness decreases. Innovative care models may even use healthcare funds for transportation or housing, if that is what it takes to manage the health care costs of certain individuals, further reducing homelessness. Fewer unintended pregnancies may reduce the crime rate.

Principles of Quality Measurement

The almost limitless expansion of evaluation and measurement of the ACA provided above may seem foolish. But it makes a point about keeping an end goal in mind. Diabetes is a condition familiar to most everyone and certainly all healthcare professionals. We measure whether hemoglobin A1c is performed. Do we care if a hemoglobin A1c is performed? No, we really care about the result being optimized. We measure whether the hemoglobin A1c is within a target range. Do we care about the hemoglobin A1c result? No, we really care about avoiding end organ complications of diabetes such a stroke, heart attack, amputation, blindness and kidney failure. Do we care about end organ complications in people with advanced dementia who have diabetes? Probably not, but we care about their comfort. Is there evidence that measuring and controlling the hemoglobin A1c in a person with advanced dementia improves comfort? It is unlikely there is. The converse is just as probable. We care about access to affordable health insurance because lack of health insurance is associated with death, disability and lost productivity, not because we really care about insurance.

Measurement of quality and the outcomes of healthcare is an exercise in compromise: guidelines do not apply to every patient; only major exceptions can be included in measures; data collection must be efficient and therefore may rely upon information primarily submitted for payment purposes; and risk adjustment is impossible or imperfect. For this reason, the intended use of the measure is critical. The intended use should define the selection criteria and measurement methodology. For example, an internally defined measure may be just what is needed to assess the impact of a rapid cycle quality improvement process. However, such a measure would be inappropriate to compare two providers in different regions. Some measures may effectively be used in comparing certain provider types, but not others. For example, a surgical infection rate is much more likely to be related to the facility and its team of providers than an individual surgeon. It is generally true that the broad intent of measurement is to improve health by improving healthcare.

It is impossible to assess interventions unless there is measurement, and the adage is that one cannot improve what one cannot measure.

In order to better understand quality measures in health care an overview may be useful. In the 1960s Avis Donabedian described a model of defining quality that looked at three attributes: ***structure, process and outcomes***. This model remains relevant. The definition of "outcomes" may vary depending on whether the use is a clearly relevant patient oriented outcome such as death or whether it is an intermediate, proxy or short term "outcome" such as an LDL level that is truly not an outcome at all, but is a result of a process of care. Each type of these measures or attributes of quality have a role in evaluation and improvement. However, for any of them to be meaningful, the measure must ultimately be linked to an outcome that is meaningful such as death, function or comfort.

An example of a structural measure would be whether a Medicare Accountable Care Organization (ACO) has a governance structure that requires organizational leadership from a person with competencies in geriatric care. This may make sense from a theoretical point of view, but ideally it is bolstered by evidence that such a structure leads to better results clinically or in cost or both. Structural measures are often "standards" and tend to be readily defined and measured. Nursing hours per patient is a structural measure that Medicare has adopted for nursing facility performance measurement (Medicare.gov Nursing Home Compare).

Process measures are those that evaluate the process of care. Whether an appropriate perioperative antibiotic was given at the right time or not for a specified surgical procedure is a process measure used in Medicare (Medicare.gov Hospital Compare). These types of measures are widely used. A major advantage of process measures is that they require much less risk adjustment in use than an outcome measure. If everything in control of the health care team was done properly and the patient died anyway, then it must have been due to uncontrollable factors and the care was good despite the outcome, or so it is theorized. Even process measures may require consideration of the types of factors that might be labelled risk adjustment. For example, obtaining mammograms is a process measure. Breast cancer related morbidity and mortality is the outcome of concern. The rate of obtaining mammograms in the appropriate population in a given practice is dependent on many factors including providers ordering the test, providers explaining the value of the test, the patient's pre-conceived beliefs of the value of the test, the ease of access to the test and the ability to pay for the test. Some of these factors seem almost entirely clinician controlled and others are almost entirely not clinician controlled, yet this is a very widely used process measure without adjustment.

Outcomes measures are likely to be the most meaningful metric. However, they are most likely to require some form of adjustment. A simple example is cancer treatment efficacy being adjusted for stage at presentation. Unfortunately, most adjustment is not so straightforward. Meaningful outcomes may also take years to show separation based upon the quality of care. A wrong site surgery has a fairly instantaneous outcome. The functional, behavioral, vocational and social outcomes related to pediatricians and family physicians screening for developmental disorders has a relatively long time horizon.

There are other ways to categorize quality measures. A very logical method to clinicians is division defined by ***prevention, acute care or chronic disease management***. One would anticipate that a national evaluation would include all these types, but measurement of an individual provider may not include all three depending on the practice type. The Institute of Medicine defined six attribute domains of health care quality: ***safe, timely, effective, efficient, equitable and patient-centered***. This creates an intellectual framework in measure development and selection. It also effectively addresses the need to consider attributes such as efficiency and equity that have not always been considered relevant by professionals focused on the single patient. The National Quality Strategy has translated this into six measurement domains listed as: ***patient and family engagement, patient safety, care coordination, population/public health, efficient use of healthcare resources and clinical process/effectiveness***.

Quality measures may also be defined by the unit of measurement. There are obvious differences in numbers of members, patients or clinical events between a health plan, a hospital and an individual provider. But there is a more fundamental issue regarding ***population*** as compared to ***patient***. Traditionally, clinicians have accepted responsibility for the care the clinician provided to the patient who came to the clinician for that care. As individual clinicians accept greater responsibility for populations, they are more and more measured on performance at the population level. It is not adequate to just do the right thing for the person in front of the clinician. Rather, the clinician or the team the clinician leads must make sure the patient receives the right care, even if that requires outreach and support provided outside of the context of a face to face encounter. The population of concern may vary greatly. It may be all the patients for a single clinician, or all the patients of an integrated healthcare system or even all the persons in a community. But the conceptual difference from a single patient focus is consistent. The transition from single patient focus to population management has many reasons. In some cases it is because clinicians have aggregated into healthcare delivery systems and seek to be evaluated or rewarded based upon efficacy of population management. The transition to value based payment has caused many to recognize that aggregation creates a larger patient population size being measured and thus spreads risk and reduces the potential for adverse effects based upon the randomness of results inherent in small population size. In other cases, it is because clinicians recognize their role in improving access, chronic condition management and other factors that justify population as a unit of measurement. While, population measures may be relatively irrelevant for those who provide time limited specialty acute care, such as an orthopedist who repairs a fractured hip, if that same orthopedist is part of a multi-specialty group that manages a population of persons with osteoarthritis, population based metrics may be valid. Consideration of population metrics also requires consideration of special populations or a range of populations. Measurement of our national healthcare requires a scope sufficient to measure care of different age segments, genders, races and ethnicities, socio-economic status and a host of other population subsets.

Measure Selection

There are several decision points that are undertaken in deciding what to measure. Some are alluded to above with respect to creating an appropriately broad scope of measures that relate to the key attributes of quality. Basics include the following:

1. The condition is meaningful to the population of concern.
2. There is a clearly defined measurable structure, process or outcome.
3. If not an outcome measure, there is an acceptable evidence base for the structure or process of care being related to an outcome.
4. There are existing opportunities for improvement based upon preliminary measurements. This may be due to regional or institutional variation or may be overall suboptimal performance across the population. These are often called gaps in care.
5. Measurement is feasible.
6. The cost and effort devoted to measurement is justified when balanced against the attention and resources that might otherwise be used in improving health.

Each one of these points raises issues. For example, an advocacy group may be justified in believing there should be national quality measures related to the disease that is their reason for existence, but others with a broader perspective may disagree. Those same parties with a broader perspective may conclude that not all measures must be for the most prevalent conditions and that especially vulnerable populations need a measurement focus. There may be controversy regarding the evidence. Should mammography start at 40 or 50? Should it be every year or every other year? What is "feasible" and "efficient" may vary depending upon the level of infrastructure or choices made in measure definitions. A claims based/administrative data based measure of quality may be useful and feasible, whereas chart audit may be superior, but wholly impractical, even if technically feasible. A measure that drives systems of care to change in a way that improves overall care for multiple conditions, not just the target condition, is ideal, whereas a measure that merely results in clinicians playing to the test is less desirable.

Denominators, Population Size and Attribution

Part of relevancy or being meaningful is frequency of the event or prevalence of the illness. But, prevalence also has a direct bearing on whether meaningful measurement can be accomplished. Having a large denominator in a measure has several advantages. The first is that the measure is now a "study" effectively powered to demonstrate real rather than random effects with some high degree of probability. The other advantage is that the probability of skew created by a subpopulation is less. This reduces the need to risk adjust or reduces the error inherent in the imperfections associated with risk adjustment. For example, it is possibly the case that two health plans of 100,000 members each in the same region can be so significantly

different in member characteristics that this difference in characteristics would affect the probability of attaining certain results, but a significant difference in characteristics is substantially more likely when the comparison is between two single clinicians. This phenomenon has relevancy in determining the unit of measurement. It may seem desirable to compare two physicians for their ability to get a Hemoglobin A1c to goal. But that may well not be possible with any validity based upon the denominator of the measure. It is somewhat surprising how few patients with a specific condition many single clinicians have. This small number phenomenon is made worse when the measurement is by payer rather than aggregating the clinician's entire patient panel. An all payer measurement of a practice site may be more valid. Measurement of a collection of practice sites within an integrated healthcare delivery system may be even more legitimate.

Value based payment language requires measurement of an entity. How a patient is assigned to that entity varies. For a Medicare Advantage (Part C) plan, Medicare beneficiaries must enroll in a plan. Some the care they received or did not receive may not have been while a member of that plan, but the measurement year membership is clear. For example, a health plan will get credit for a screening colonoscopy paid for by another prior plan if done within the required look back time period. However, the measured plan must be able to demonstrate with records that it was performed. Likewise, if another plan failed to get the member to such screening for many years past the recommended performance date, the new plan is still responsible to fix the gap in care within the single measurement year.

In many cases attribution is not so simple. Patients often see many doctors, for example. Assume a patient has COPD and hypertension. Annually the patient sees a pulmonologist, who also seems to do a significant amount of primary care for other patients based simply upon billing/procedure codes submitted by the pulmonologist to a payer. Twice a year the patient sees a doctor, who is mutually acknowledged by the patient and that doctor to be the primary care physician, and receives a general assessment and blood pressure measurement. The patient experiences a burn on his arm one holiday weekend and has three visits to an urgent care facility for assessment and dressing changes. The doctor there is a family physician, but does not provide chronic care management or preventive services other than immunizations. The patient then manages the burn on his own. Attribution may assign this patient to the urgent care doctor as this doctor had the plurality of office visits performed on the member during the year. Of course, attribution could be different if the database and logic used was set up so that the urgent care physician could not have a patient attributed to him, except for assessment of the care she or he provided (e.g. a measure of the quality of minor burn care). Diagnoses could theoretically be used to define primary care, but this would be an extraordinary challenge given the breadth and overlap of conditions managed by different clinicians. However, diagnosis may be valid for assignment in the case of the clinician who reported the diagnosis of hypertension for the visit being assigned the responsibility of getting the blood pressure to the goal. The performance of a Medicare Annual Wellness Visit might be used to define the Medicare beneficiary's primary care clinician, but the Annual Wellness Visit may be performed by anyone, not just the primary care

staff by current rules. Where this becomes especially relevant is when a population is to be managed and payment is based upon this. The managing clinicians may effectively manage someone who ultimately is not even attributed to them and potentially fail to manage someone who is ultimately attributed to them, but whom they thought was the responsibility of another entity.

As a general rule, assignment of responsibility for a quality metric should consider the locus of control of the party being measured. Control may not be complete. There may be patient factors. There may be system factors. These alone do not make measurement pointless. But performance is unlikely to improve and behavior unlikely to change if the result measured is entirely outside of the control of the provider of care being measured. A good example of this limitation is a measure of the national cost of care trends called the Sustainable Growth Rate (SGR). From a national economic perspective it is logical to assert that the segment of the economy devoted to healthcare expenditures cannot consistently grow faster than the overall economy. However, the SGR is enforced at the individual clinician level and is based upon cost trends at a national level of a subset of Medicare expenditures – those paid on the physician fee schedule. No amount of dedication to stewardship of resources by a single individual will have an impact on the SGR. But payment reductions when the SGR is exceeded fall upon every individual.

Adjustments

The perfectly fair adjustor that makes all comparisons valid is the Holy Grail. This is the domain of the statistical experts. Adjustment can create more valid comparisons. It also introduces an element that clinicians can perceive as invalid or obtuse. The greater the level of sophistication of the adjustment, the more complex it typically becomes and the more likely it will appear to be a "black box" to the party being measured. For most measures there is no accepted adjustor. Some bear mentioning, however. The most significant risk adjustment relates to payments to plans for populations. Medicare Advantage plans have been paid this way for quite some time and new exchange products will use a closely related adjustor to redistribute revenue between plans. The adjustment is the Centers for Medicare and Medicaid Services (CMS) Hierarchical Condition Category (HCC) system. This system is diagnosis based and does require that the diagnosis be managed, evaluated, assessed or treated, if it is to be included in the payment adjustment algorithm. Nonetheless, the huge financial impact of this adjustor and the response/need of plans to maximize revenue using it, has raised concerns that it is not just adjusting for risk related *expenditures*, but has become a major *revenue* center. This is an example of how risk adjustment may generate as much controversy as it resolves. There are methods to estimate probability of all cause readmission that are tested and being used (e.g., by the National Committee on Quality Assurance). The logic and specific mathematics of these models do not translate to other uses, such as adjusting for expected emergency department visit rates or expected rate of blood pressure being at goal.

Another commonly used adjustment is some form of outlier methodology. Outlier patients could be eliminated altogether. For example, one patient uses the emergency department 20 times a year and that one patient drives the emergency visit rate per patient for a practice. Another practice with the same number of patients has 20 patients who visit the emergency room once each. They have the same rate. However, it may be that the first practice has expanded hours, always immediately responds to pages and manages a wide range of conditions in the office, while the other practice has done little to reduce use of the emergency room as a site of the type of care that could be provided in the office. The outlier patient results in incorrect conclusions about the first office. A more typical method involves truncating outlier costs. A large group involved in a risk sharing arrangement will have costs of up to $100,000 per year for a given patient attributed to the group. Costs over this amount are not attributed. This reduces the effect of a single patient on per capita or per member per month expenses, but does not eliminate any recognition of the costs. Therefore, under this methodology, there is no chance that a $99,000 patient would appear more expensive than a $200,000 patient.

Episode treatment groups may be used to compare total costs of care for a specific condition. This method defines a condition and has rules as to when the condition starts and ends, i.e. when it is an episode. It also includes rules as to which expenses are condition related and which expenses are not condition related. Some models also divide related expenses into those that are expected and those that are complication related. An example of an episode treatment group would be the cost of care over a year for a patient with heart failure. It would start at the beginning of the year, even if the heart failure was not diagnosed until mid-year. All office visits to certain specialties would be included, even if the diagnosis on the claim was not heart failure. This would account for other potentially related conditions being included. Certain procedures such as echocardiograms, electrolyte and renal chemistries, cardiac catheterizations and cardiac rehabilitation would be included whereas care for a fracture of the radius would not be included. Certain inpatient diagnosis related groups would be included, whereas others would not. While this is simply intended as an example, it becomes obvious that a host of decisions must be made about what is or is not part of the episode. The radius fracture could be due to a fall caused by debility related to heart failure. The visit to the cardiologist may have been for dyspnea that was actually caused by anemia from a gastrointestinal blood loss and not remotely heart failure related. If costs are used for comparison purposes, there needs to be a decision as to how to handle price variation. This is especially important outside of Medicare where allowed payment amounts may vary considerably. If one seeks to measure real costs, then price variation may be relevant. For example, a group that accepts risk for the cost of care may save money without adversely affecting quality by simply using a lower cost provider such as a free standing radiology facility rather than a hospital based facility. On the other hand, if the goal is to look at efficiency related to utilization patterns, price may not be relevant and could actually obfuscate the analysis.

Propensity matching is used at times. This methodology looks at matching two populations through weighting methods. Then comparison is made. Again decisions

need to be made about how weighting is made and what models are used for weighting. Would a historical average over the last 3 years be used to create prospective weights or would the activity of the measurement year be used to retroactively create weights?

Adjusting for socio-economic status (SES) is controversial. Few would dispute the social determinants of health such as wealth and education. However, adjusting for these might mean that it is acceptable to have lower quality of care for those in a lower SES. The debate about test performance in schools and the quality of the school is just this type of debate. It is not just a healthcare quality issue, but a broader social issue. What may determine the need to adjust or not is how the measurement is used. If safety net facilities are generally acknowledged as doing incredible work with challenging populations, yet a pay for performance system drives them into the red financially, there is a problem. The solutions to that problem may be less obvious, but may include a factor related to SES or comparison to peer entities at least. If the use is simply to create information for facilities to use over time, SES adjustment may be irrelevant.

Setting Goals and Thresholds

Various terms are used for a result that is desired. It could be a goal, i.e., something that is sought to be achieved. It could be a benchmark, which usually means a result that is excellent, possible and has been attained by some entity. It could be a threshold, meaning that attainment triggers something, such as additional payment. Each measure may have all of these and there may be multiple tiers or thresholds. The distinctions may be irrelevant if the goal of an organization is to hit the threshold.

A measure usually must be tested to determine if there are variations or gaps and if it can be reliably collected/performed. This testing process also allows an historically-based definition of a goal, median, threshold or top benchmark performance. It may be that the ideal is 100 % of the time XX will occur. But, as clinicians know, there are usually valid reasons for performance at a level of less than 100 %. This is why practice guidelines are called guidelines. There are reasons such as patient rights or other conditions that are too rare or diverse to list as exclusions that affect results. Therefore, historical norms and relative rates are typically used. The goal of measurement is first and foremost to drive *improvement*. So a practice or a hospital may focus on pushing the numbers in the desired direction. The public or a value conscious payer may be equally or more interested in identifying and/or rewarding higher *performance*. This potential dichotomy is characterized as pay for performance contrasted with pay for improvement, when payment is involved. The arguments for both methods are strong. Failing to recognize improvement can create hopelessness and disadvantage those who care for the most challenging populations. Rewarding improvement alone fails to recognize those who may have heavily invested in improvement long ago and now are sustaining those results. They may be improving, but in areas for which there are not yet measures used by the performance program. If improvement alone is recognized, they would be deemed failures

because of their very success, whereas a perennial poor performer without legitimate explanation for past results finally improves a little and is now deemed the successful party. A hybrid method recognizing performance, but also recognizing improvement may be used to address both positions.

An example of the real world challenges of setting thresholds is seen in the Medicare Pioneer ACO program. The program uses a comparison of the specific ACO cost trend to trends in a national reference population of beneficiaries who are not in the ACO. Accordingly, a high performing ACO in a cost efficient region can fail as there are no savings, because they are efficient historically, even when their absolute costs (not cost trends) are well below national medians of other ACOs or non ACO aligned beneficiaries. This is true, even if the ACO performs better than its regional non ACO providers. Presumably, such better performance is ACO related and not related to regional variation. On the other hand, an ACO that has historically high costs in an historically high cost region shows some improvement, and while still relatively costly, is rewarded. This would be true even if the ACO did no better than the regional providers. Another analysis that relies upon comparison to the local community or a nearby community may show different results. The Pioneer ACO method compares trends. Therefore, if the two populations being compared do not dramatically change over the time periods from baseline to measurement, risk adjustment is less of an issue. So this method has some appeal. However, this method may cause one organization that is doing good work, to move away from an alternative payment method that is in theory designed to pay for value, because the organization is not being paid for the value it brings and is not recovering the investment costs necessary to obtain those results.

The intended use of the measure also defines how the thresholds should be set. The goal could be to reward only the top performers. In this case, the threshold is either purely performance relative to a pre-defined percentile (e.g. the top quartile) or attaining a result that based on history represents top performance. The latter may be selected so that a specific target can be announced in advance. However, the threshold would be different if all but low performers were to be recognized for investing in improvement. The threshold could also be a gate. For example, it may be that the structure of the program is to allow shared savings in cost of care for a population. However, the payer wants to be certain that quality did not deteriorate while savings were achieved. In this case, a floor quality performance rate might be the ticket to sharing savings. It could be that quality metrics must be maintained, but need not improve or be higher than the norm to pass through the gate for sharing in savings.

Where and how the dial is set also relates to other objectives. If the goal is to get providers to seriously think about measuring quality, one might just pay for reporting as was done in the CMS Physician Quality Reporting System (PQRS). If the goal is to drive lower performing providers out of the market and to force them to merge their entity with or lose their patients to an organization that has a formula for success, then targets may be rather aggressive. They may also need to be adjusted in a way that is local market dependent if high targets based on national norms would mean there were no providers left standing. The amount of money (if any) at stake may also determine the threshold of success. High performance reward thresholds may be set at a level of very high performance if the reward is unequivocally a bonus payment.

The same might be true if the target result was highly aspirational and what was at stake was a trophy. However, if the payment is essential for operations, it is unlikely a target that fatally wounds all but a few high performers could be chosen.

A by-product of setting performance goals based upon historical results is that measures should not be perpetually modified. Some stability and consistency is needed. There are other reasons for this such as the added costs of measurement if abstraction software must be constantly modified for ever changing measures.

Finally, thresholds may be selected based upon confidence intervals around a measure. In other words, the threshold is selected because it represents performance that is statistically highly likely not to be a random effect. Assume that the score 0.75 is the threshold result for the top quartile among a group of entities being measured. Assume 0.75 means 75 % of the time the desired process was performed and 1.0 means it occurred 100 % of the time. However, the individual entities being measured really have a result within the band of X plus or minus 0.25 with a 95 % probability based upon their population sizes. Assume the entities are similar in population and this confidence interval is constant. It would probably not be reasonable to conclude that threshold must be 1.0, even though only if the threshold is set at 1.0 can it be certain that the actual result is 0.75 or greater. It might be more reasonable to set the threshold at 0.5 knowing that all actual 0.75 performers and above will be recognized. This decision also means that entities that are actually only at 0.25 may also get recognized. If these alternatives are not acceptable, a minimum denominator that reduces the size of the confidence interval may be selected. However, this may exclude too may entities for the goal of the program. Ultimately, such a calculus and logic could result in abandonment of the measure as being useful or feasible.

Other Challenges

There are a host of other challenges in measurement and evaluation. Most healthcare expenses in Medicare are for beneficiaries with multiple chronic conditions. Most quality measures are single condition oriented. Those with expertise in caring for the multi-morbid recognize the weakness of such measurement. Recognition of weakness rises to serious concern when performance measures affect payment as is the case in the Value Based Purchasing provisions of the ACA. Care for those with multiple chronic illnesses requires clinician and patient to set priorities. The patient's values may direct that a goal that makes sense for other patients is not set as a goal for them. There are patient experience surveys that address whether the patient felt involved and respected in their care and such surveys may provide a mechanism to measure patient centeredness, which may be what matters most for this population. The Consumer Assessment of Healthcare Providers and Systems (CAHPS) program from the Agency for Healthcare Research and Quality (AHRQ) is designed to achieve measurement of patient-centeredness and is also expected to be part of Medicare evaluation programs.

There are not well defined measures for many provider types and population subsets that could be the dominant type of patient for a specific provider. Clinicians who care for a highly atypical patient population may not be appropriately measured

by instruments that work well for those caring for the more typical population mix. At extremes, adjustment is likely to be ineffective in addressing this problem.

Performance measurement is intended to support quality care. It is important to acknowledge that while reducing variation in care using evidence based standards is generally desirable and likely to improve the health of a population, care must be applied at the individual patient level.

Developing Measures Through Consensus

Quality measures used within the ACA must meet certain standards. They are generally developed by using a process of consensus, endorsement and validation. The National Quality Forum (NQF) plays a major role in endorsing measures that have been developed and presently has a formal role in the PQRS process. It also may convene groups to develop measures and endorse measurement processes. Many organizations may develop measures, such as a medical specialty society, AHRQ or the National Committee on Quality Assurance (NCQA). Measures may be used in ways that are not exactly as intended in some programs (e.g. in a private payer program), but the use in ACA Value Based Payment programs is more tightly governed. These programs go through the rule making processes of the federal government with published proposed rules, comment and publication of final rules.

National Quality Strategy: Prioritization and Alignment

One serious concern is the proliferation of measures and measurement. In an ideal world, measurement would be organic in care, not just built into documentation systems. It would contribute to focusing on what really matters. Many clinicians using electronic records are all too familiar with the concern that record structure seems to support payment and reporting programs at the expense of supporting clinical care, patient interaction, clinician focus and critical thinking, even if the same clinician acknowledges the many merits of selected measurement and electronic records. There are legitimate concerns that the cost of measurement diverts resources that could be better used. The Institute of Medicine has labelled the need to combat measure proliferation as “Counting What Counts”. The National Quality Strategy (NQS) is designed to address this as well. The measures and measurement of the ACA will reflect the NQS as amended periodically.

The National Quality Strategy was first published in 2011. It is led by the Agency for Healthcare Research and Quality (AHRQ) on behalf of the U.S. Department of Health and Human Services (HHS). It was established as part of the Affordable Care Act in order to facilitate a consistent focus on quality improvement efforts and a nationwide approach to measuring quality. The ACA requires HHS agencies to develop Agency-Specific Plans to achieve the NQS priorities; establish annual benchmarks for success; and regularly report on progress against these benchmarks.

The ACA also established the Interagency Working Group on Health Care Quality. This group includes 24 Federal agencies and has a mission to foster collaboration, cooperation, and consultation on quality-related efforts between Federal departments and agencies, and with the private sector. The NQS is not a federal program despite the essential facilitation role federal agencies play and the requirements of the ACA to have a national strategy. The NQS achieves its goal by working with the NQF. It has two formal partnerships: the National Priorities Partnership and the Measures Application Partnership. The National Priorities Partnership is made up of over 50 national organizations with a shared vision to achieve better health, and a safe, equitable, and value-driven healthcare system. The Measures Application Partnership is a public-private partnership that reviews performance measures for potential use in Federal public reporting and performance-based payment programs. It also seeks to align measures used in public and private payer programs.

Project evaluations, such as those of new activities of the Center for Medicare and Medicaid Innovation (CMMI) will have evaluation metrics relevant to the specific project. However, major programs such as Medicare 5 Star, PQRS and other Value Based Purchasing programs will reflect these activities. Programs such as PQRS, Meaningful Use and others will align as the NQS achieves its goals. The NQS 2013 Progress Report to Congress outlines measures related to the six priorities (Table 9.1).

Table 9.1 NQS - Improving Quality Across Six Priority Areas (2013 Report to Congress)

Measure focus	Measure name/description
Priority 1. Making care safer by reducing harm caused in the delivery of care	
Hospital-acquired conditions	Incidence of measurable hospital-acquired conditions
Hospital readmissions	All-payer 30-day readmission rate
Priority 2. Ensuring that each person and family is engaged as partners in their care	
Timely care	Adults who needed care right away for an illness, injury, or condition in the last 12 months who sometimes or never got care as soon as wanted
Decision making	People with a usual source of care whose health care providers sometimes or never discuss decisions with them
Priority 3. Promoting effective communication and coordination of care	
Patient-centered medical home	Percentage of children needing care coordination who receive effective care coordination
3-Item care transition measure ®	During this hospital stay, staff took my preferences and those of my family or caregiver into account in deciding what my health care needs would be when I left
	When I left the hospital, I had a good understanding of the things I was responsible for in managing my health
	When I left the hospital, I clearly understood the purpose for taking each of my medications

(continued)

Table 9.1 (continued)

Measure focus	Measure name/description
Priority 4. Promoting the most effective prevention and treatment practices for the leading causes of mortality, starting with cardiovascular disease	
Aspirin use	Outpatient visits at which adults with cardiovascular disease are prescribed/maintained on aspirin
Blood pressure control	Adults with hypertension who have adequately controlled blood pressure
Cholesterol management	Adults with high cholesterol who have adequate control
Smoking cessation	Outpatient visits at which current tobacco users received tobacco cessation counseling or cessation medications
Priority 5. Working with communities to promote wide use of best practices to enable healthy living	
Depression	Percentage of adults who reported symptoms of a major depressive episode in the last 12 months who received treatment for depression in the last 12 months
Obesity	Proportion of adults who are obese
Priority 6. Making quality care more affordable for individuals, families, employers, and governments by developing and spreading new health care delivery models	
Out-of-pocket expenses	Percentage of people under 65 with out-of-pocket medical and premium expenses greater than 10 % of income
Health spending per capita	Annual all-payer health care spending per person

Specific measures within specific federal programs are too numerous to list. For example, the 2014 Medicare Part C 5 Star program has 36 measures and the Part D 5 Star program has 15 measures. PQRS has hundreds because it is for use by many different professional disciplines. ACOs must report 33 quality measures in the Medicare Shared Savings Program.

Summary

Society, government and professionals have devoted considerable time and effort to devising methods to improve care and achieve the goals of the Triple Aim. It is a work in progress. There is a mandate in the ACA to measure, improve measurement and use measurement of value in payment. The efficacy of these efforts will be a measure of the success of the ACA itself.