Paper 3760-2015

Methodological and Statistical Issues in Provider Performance Assessment

Daryl Wansink, PhD, Qualmetrix, Inc.

ABSTRACT

With the move to value-based benefit and reimbursement models, it is essential to quantify the relative cost, quality, and outcome of a service. Accurately measuring the cost and quality of doctors, practices, and health systems is critical when you are developing a tiered network, a shared savings program, or a pay-for-performance incentive. Limitations in claims payment systems require developing methodological and statistical techniques to improve the validity and reliability of provider's scores on cost and quality of care. This talk discusses several key concepts in the development of a measurement system for provider performance, including measure selection, risk adjustment methods, and peer group benchmark development.

INTRODUCTION

Value based care is an emerging strategy in the US healthcare system. The premise of value based care is simple in concept – high quality and low cost provides the greatest value to patients and the various parties that fund their coverage (mostly the government and employers). However, value can be hard to measure empirically. Measuring cost of care has challenges but is fairly straightforward to measure in the current fee for service environment. Quality of care and the outcomes of the care (health and well-being) are far harder to measure today in a comprehensive way that can drive policy decisions and inform patient decisions around seeking care.

A key element to value based care in all of its forms is that you must know what the quality of care is (whether the doctor did the right things or whether the procedure provides the best outcome) and the cost of care. While it is conceptually fairly straightforward to compute the cost of care for services received, it is far more challenging to measure the quality of care and outcomes of care. Furthermore, comparing cost of care and quality of care across doctors is far more challenging because the comparison needs to be fair and accurate. As a result, measurement theory plays an important role in building the methods and scoring systems to quantify cost and quality, especially in assessment of doctors and facilities.

VALUE BASED HEALTHCARE

Value based care is an emerging strategy in the US healthcare system. The premise of value based care is simple in concept – high quality and low cost provides the greatest value to patients and the various parties that fund their coverage (mostly the government and employers). However, value can be hard to measure empirically. Measuring cost of care has challenges but is fairly straightforward to measure in the current fee for service environment. Quality of care and the outcomes of the care (health and well-being) are far harder to measure today in a comprehensive way that can drive policy decisions and inform patient decisions around seeking care.

There are many ways in which a value-based health care payment systems are being implemented and these methods are not mutually exclusive:

1. Bundled payment for services is a way that doctors are paid for the total cost of treating a patient's condition but the budget for the bundle does not pay anymore for services due to medical errors or the choice of less cost effective treatment options.
2. Value based benefits for patients are insurance products designed to reduce financial barriers to receiving care for services that are cost effective. These benefits can range from reducing co-pays for highly effective drugs to only covering services provided at the best facilities in their region.

3. Limited or tiered networks only allow patient to choose from a sub-set of all doctors in their region who are considered to be of good quality and are less expensive at treating health conditions. Patients often face a financial penalty for seeing a doctor outside of their high value network.
4. Centers of excellence are facilities that are designated as being the best at providing high quality and cost effective care. Sometimes they are designated only for a specific procedure, such as a knee replacement or coronary artery bypass graft. A patient would only have a procedure paid for if they go to one of these centers of excellence.
5. Accountable care organizations are groups of doctors organized in a way to provide care to patients and accept some or all financial responsibility for those patients care.
6. Cost and quality transparency tools are common in other industries (think of Yelp or urban spoon for restaurants; consumer reports for many consumer goods), but are not yet common in healthcare. By making cost and quality transparent to the patient, they can make an informed decision to seek care from the best doctors in their network.

A key element to value based care in all of its forms is that you must know what the quality of care is (whether the doctor did the right things or whether the procedure provides the best outcome) and the cost of care. While it is conceptually fairly straightforward to compute the cost of care for services received, it is far more challenging to measure the quality of care and outcomes of care. Furthermore, comparing cost of care and quality of care across doctors is far more challenging because the comparison needs to be fair and accurate. As a result, measurement theory plays an important role in building the methods and scoring systems to quantify cost and quality, especially in assessment of doctors and facilities.

MEASUREMENT THEORY APPLIED TO PROVIDER PERFORMANCE

The two most important and fundamental aspects of measurement theory are reliability and validity. Any measurement process must be shown to be reliable and valid when being implemented for something as important as measuring provider performance on cost and quality.

RELIABILITY

Reliability reflects the precision with which a measure quantifies a concept. A perfectly reliable measure will quantify the measure exactly each time. Typically, though, there is error or ‘noise’ in a measure. This can be quantified in several ways. The first way is to repeatedly measure the same construct over time and see if there is any variation over time. This test, re-test reliability assumes that if the construct hasn’t changed during that time then any differences in the measure are due to noise in the measurement process. The correlation between the test and re-test is a useful indicator of reliability. Another test for reliability involves measuring the correlation among a series of measures intended to measure the same construct. If you can measure a construct several ways then you would expect the measures to correlate with each other. Cronbach’s alpha is a metric commonly used to measure the internal reliability of several measures and is available as an option in PROC CORR. Inter-rater reliability is a similar concept to test, re-test reliability except is based upon measures that involve judgement by two or more individuals reviewing the same observations. For example, the reliability of a medical chart audit can be assessed by how closely two or more people score the same charts.

VALIDITY

Validity assesses whether a measure accurately measures the construct of interest. If a measure does not measure the actual construct in the real world then it cannot be valid – even if it may be precise/reliable. There are a great many ways to measure validity but only a few are relevant for provider performance measurement. Content validity is the most common form of validity assessment used in healthcare analytics. Content validity is a non-statistical test, usually done by experts in the field of study, as to whether the critical elements of the construct being measured arte included in the measure. For example, a measure for quality for orthopedic surgeons probably would not be considered valid if it did not include surgical complications for procedures. Content validity is often established by organizations

such as medical societies, government agencies or associations like the National Quality Forum. Criterion validity is a measure of whether the current measure being used correlates with other measures that are purported to measure the same construct. Typically, you will want to see a measure correlate with other similar measures of the construct but also NOT correlate with measures that are unrelated to the construct. A challenge with criterion validity in provider performance assessment is that there often is not a 'gold-standard' or benchmark to measure against. However, by examining the correlation between related measures you would hope to see a relationship. For example, quality of care is a complex concept but one would hope that a quality measure based on claims data would correlate with other quality measures such as medical society accreditation, patient reviews or audits by quality agencies.

The concepts of reliability and validity are easily understood by clinicians even if they haven't had formal training in measurement theory. And it is important to show how the measures used to assess their performance are fair (i.e., valid) and accurate (i.e., reliable). One of the biggest challenges faced by healthcare analysts is the availability of the data to do checks on reliability and validity. In the next several sections, statistical methods will be reviewed to help show how your measures are fair and accurate given the lack of perfect data.

MEASUREMENT ISSUES IN COST OF CARE

Reliability

Because administrative claims data are built for billing purposes, it would be reasonable to assume that claims are a reliable measure of cost of care. While there are many issues that need to be addressed in cleaning claims data, they will not be addressed here. Examining provider cost of care over time as a form of test, re-test reliability makes sense. Unless the provider has recently gone through contract negotiations, cost of care should be stable over time. Any wide variation in costs could indicate a measurement issue.

Validity

The concept of cost of care as applied to physicians has more to do with the efficiency of care than just measuring the cost. Simply adding up all of the claims paid to a physician does not reflect how efficiently they provide care relative to peers. The cost of care for a physician is determined by a number of factors, some of which may or may not be valid to include in a measure of provider performance. Contracted rates for services, the types of services, the volume of services and the illness severity of patients all will vary across providers.

Approaching this from the perspective of a payer (or a self-insured employer) the types of services, volume of services and contracted costs for services are valid factors that should be included in measuring cost of care. If a provider is willing to receive lower payment for a service, appropriately uses less of a service or uses lower cost services that are equally effective then that provider should receive a more favorable score. The biggest issue with measuring provider cost is the illness burden of the patients. Illness burden will increase the volume and complexity of the services and leads to higher cost of care. Because some providers are part of highly specialized practices or work in academic centers that receive more complex cases, it is critical to adjust for illness severity to have a valid measure of cost of care.

Measuring illness burden is complicated by the fact that administrative claims data are not designed to collect specific and complete clinical information. First and foremost, they are a transactional billing system, and the information required to pay a claim is far less than that needed to fully document illness burden. Even some electronic medical record systems do a poor job of documenting illness severity in a manner that can be easily consumed for analytic purposes. Because the healthcare industry is not going to wait for a perfect measurement system to develop provider efficiency scores, it is necessary to find ways to improve measurement of illness burden with the data at hand.

Three approaches can help with measuring illness burden:

1. Disease-specific risk adjustment models.

There are several commercially available risk adjustment models for health conditions. Typically these models use prior diagnosis codes to predict the total cost of treating the condition and are part of episodic

claims groupers such as Episode Treatment Grouper (Optum Insight, Inc.), Medical Episode Grouper (Truven, Inc.) or Prometheus episodes (HCI3). Typically, these models are retrospective or concurrent predictors of cost, rather than prospective. They reflect how much the illness burden of the patient affected the total cost of care. It is important to have a good understanding of the risk adjustment model you are using because some of the models can lead to less accurate results or actually increase measurement error. Some common mistakes are to use total cost of care models to predict disease specific costs. Total cost of care models predict costs associated with ALL health conditions, not just the condition being treated by a given specialist. These models are appropriate for measuring primary care providers, but often specialists only treat one of a patients many health conditions and the model will adjust for health conditions unrelated to the specialist. For some specialties like cardiology or oncology the total cost model might be adequate since care for that condition will often dominate other costs of care.

Another common error is to use a predictive model that includes prior utilization history and costs to predict future costs. These models are mainly used for care management to predict things like hospital admissions or high cost members who may need additional help. Because these models consider how much resources they used in the past, it can unfairly give a provider 'credit' for poorly managing the patient in the past or having a chronic history of over-utilizing services. Models used for risk-adjusted reimbursement by governments never include prior utilization history for this reason.

The issue of whether to use a retrospective or prospective model in risk adjustment is a little more complicated. A retrospective model will capture more acute events that affect cost of care in the current time period but not in future time periods. For example, a serious infection could increase costs in the short term but should ultimately not lead to high costs in the future. Using a prospective model will not adjust much at all for the infection, but a retrospective model will. The issue then becomes whether risk adjusting for the infection is the right thing to do. An infection that is caused by the provider should not be adjusted for. An infection due to some external agent that is independent of the provider's treatment should be adjusted for. Some episode groupers handle this issue of complications of care better than others. For example, HCI3's Prometheus grouper explicitly codes for complications of care and includes them as relevant costs that are not removed in risk adjustment.

2. Provider Peer Groups.

The performance of disease specific risk models can be fairly poor depending upon the condition. Part of the issue is that diagnosis codes often aren't specific enough to show the level of illness. An easy example of that is staging for cancer. ICD-9 codes do not reflect staging unless the cancer has metastasized. Because of the models low performance, it is often good to consider other possible ways to risk adjust. One way is to examine the types of services offered by a provider and then adjusting for the relative cost of patients who need these services. Certain providers may offer a limited number of services that lead to them treating sicker patients. A good example of this is cardiologists. Cardiologists can vary from non-interventionalists (who do NO invasive procedures at all) to highly specialized cardiac electrophysiologists who treat only a handful of conditions with highly specialized procedures. As a result, comparing the cost of care for non-interventionalists to cardiac electrophysiologists on the cost of treating arrhythmias will be invalid because the non-interventionalists would refer their sicker patients to the cardiac electrophysiologist. By comparing a provider only to his or her peers in certain specialties will ensure some elements of disease burden are adjusted for. Determining the types of sub-specialists can be done several ways. One way is to have an expert define the sub-specialties that should be used for risk adjustment purposes. The other way is to use cluster analysis methods (e.g., k-means) on a few important types of procedures and/or conditions to group similar providers with each other.

3. Referral Patterns.

Another form of indirect risk adjustment is to look at referral patterns between providers for treating a patient. If one provider is often the second specialist to treat the patient, then that would indicate the provider is more specialized at treating complex conditions. Because a second specialist needs to get involved, the cost of care is much higher for these patients, but it doesn't reflect the efficiency of treating the patient, only the complexity of their care. Defining who are appropriate 'second' specialists is complicated and needs to be assessed for each health condition. An easy rule is to see if two of the same type of specialty are involved in the care. For example, two cardiologists treating an arrhythmia patient

probably reflects complexity of the arrhythmia. But a PCP and a cardiologist treating an arrhythmia patient might simply be a normal and inexpensive progression of care that may not need to be adjusted for.

MEASUREMENT ISSUES IN QUALITY OF CARE

Quality measures can be grouped into two general categories: process and outcome measures. Process measures examine whether the provider did the right thing for the member (or the member did the right thing) as it relates to evidence based medicine guidelines. Outcome measures have more to do with whether the care provided to the member resulted in a desirable outcome. An example of these types of measures would be for HbA1c tests for patients with diabetes. A process measure will ensure the person had the test done in the correct time frame, an outcome measure will look at whether the HbA1c value is in a 'normal' range for a person with diabetes that indicates good control of the disease.

Reliability

Reliability is a bigger issue with quality of care due to the limited number of measures that are available and the small number of observations for any given provider. As a result, different methods of assessing reliability may be used. One method is to examine the variation in scores across providers compared to the variation in scores within a provider (i.e., across patients). You want to maximize variation across providers and minimize variation within providers with your measures. The greater the variation within a provider, the less likely you will detect a difference between providers. A useful metric to use in assessing reliability is the confidence interval for a quality measure. The confidence interval takes into consideration the variation in provider-level scores and the sample size used to establish their score. This helps assess whether the difference between providers is statistically significant. Confidence intervals are fairly straightforward to compute and are commonly understood by lay persons as to what their meaning is. Technically, they are appropriate to use because the collection of data for quality measurement is typically a sample of patients for a provider and not the total panel seen by the provider over time. As a result, the confidence interval puts a point estimate into the context of the error in a measure and the amount of data available for that provider – which often varies considerably.

Validity

The most common method to establish validity for quality measures is to use measures endorsed by medical societies and agencies like National Committee for Quality Assurance or the National Quality Forum. Generally, this will establish content validity, that the measure is determined to be relevant to the quality of the provider.

One issue raised by providers is the limited number of quality measures available to assess quality based solely on claims data. Often there are only a handful of endorsed measures with sufficient sample sizes to compute a reliable score. This introduces the importance of quality based outcome measures. Outcome measures can be applied to a broader set of patients. For example, a provider's hospital admission rate could be used as a proxy for the quality of care in controlling disease progression. Especially, if you look at emergent or unplanned admissions. This measure can be applied across an entire panel of patients, not just for specific conditions.

Often risk adjustment is not considered relevant to quality measures, especially process measures. The argument is that, according to evidence based medicine guidelines, every patient should receive the same treatment. These measures often involve the patient engaging in a behavior, not just the provider, and an argument can be made for adjusting for psycho-social factors in a providers panel that may lead to non-adherence that is beyond the providers control (such as education level, annual income). However, it is not normative in the industry to do this type of risk adjustment for process measures.

Risk adjustment is far more important in quality based outcomes such as complications of care, unplanned admissions and re-admission rates. Outcomes are influenced by illness burden or disease severity. There are clinical reasons beyond a provider's control that may influence the outcome of care, independent of the provider's skills. While there are nationally accepted risk adjustment methods for some measures like re-admissions, there are far fewer risk adjustment models widely accepted for issues like complications of care. This is a major issue the industry needs to address. In order to have a robust

and valid set of quality measures, incorporation of more risk adjusted outcomes measures is needed in the future.

In addition to risk adjustment, another important concept is the sensitivity of the measure to provider quality. Some of the HEDIS quality measures have extremely high scores for most providers, creating a 'ceiling effect'. The result is that the measure may be valid as representing quality but doesn't discriminate between high and low quality providers. When creating a quality score based on only a few measures, it is possible that you are measuring error more than meaningful differences between providers if most providers score high on a measure. This is similar to the idea of having easy questions on a test. Almost everyone will get the answer right, which means the question does little to help differentiate between students who have learned a lot or not.

CONCLUSION

The wide adoption of value based healthcare policies, benefits and reimbursement strategies makes the importance of measuring cost and quality of care critical to the sustainability of the US Healthcare system. Applying common and well established methods to assess the reliability and validity of cost and quality measures will ensure these models for health care reimbursement are applied fairly and effectively to improve care in the US.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Daryl Wansink
Qualmetrix, Inc.
919.886.2443
dwansink@qualmetrix.com
www.qualmetrix.com