Clinician Level Measurement and Improvement – Improving Reliability, Actionability, and Engagement

by Mark C. Rattray, MD

I. Preface

Healthcare purchasers, health plans, policymakers, and consumer advocates see the world very differently from individual practicing physicians and healthcare providers in general. So differently that one might posit that they are living in different worlds.

In one world there is a crisis of affordability and global competitiveness resulting from increasing costs, serious quality gaps between what is and what could be, and a frantic effort to identify and deploy cost and quality improvement solutions.

In the other there is the threat of declining personal incomes leading to the search for greater productivity and higher margin revenue sources, increased administrative burdens that pull time away from family, and the constant challenges to stay current with rapidly advancing medical knowledge, technology, and state of the art therapies. Litigation constantly lurks over each physician's shoulder.

One world calls for greater accountability and the formation of accountable entities. Supply-sensitive geographic variation is a sore upon the landscape. Costs must be driven out of the system. Clinical performance measurement is a vehicle to provide consumers' tools to make better choices and to help purchasers tie payment to value.

The other world defines healthcare costs as revenue. American entrepreneurial spirit drives what is still largely a cottage industry. In instances where this cottage industry has been replaced, one finds larger entities more formally driven by business plans seeking revenue and margin. Technology is leveraged into better bottom lines. Clinical performance measurement is a burden, and where tolerated should directly support clinician interests in quality improvement.

These worlds *must* intersect at the patient level -- not as the lowest -- but the *highest* and almost singular common denominator. We must, in a patient-centered fashion, establish as a prime directive the provision of the "right concepts" of care – right amount of care, right types of care, right timing of care, right locations of care, right providers of care – with provider revenue aligned with care optimization. Enabling such an intersection is the task before us.

Clinicians managing care are unique within our system of care providers. They, along with their patients, largely control certain critical determinants of optimal care. These clinicians and to a variable extent their patients, are likely to have the greatest influence on the amount, types and timing of care. Physicians may also strongly influence patients' choices of care providers and locations of care.

This pivotal role of physicians within our system and documented wide variation in condition-specific use of resources between individual physician practices has led the purchaser and payer community to widely deploy episode of care-based cost and resource use measurement systems. Further, these stakeholders have hypothesized that these tools, along with process of care quality measurement, may be leveraged to define and deploy "high performing networks" to reduce cost trends and improve care delivery.

II. Acknowledging historical tensions

Purchaser momentum has led to a rush to market of physician measurement systems that often lacked transparency, due process, statistical rigor, and actionability. Too often the needs for tact, sensitivity, basic understandings of human emotions, and humility as to the reliability of measurement tools have been afterthoughts. On the other hand, inertia driven by physician egos, blinding focus on day to day practice productivity, and the view that measurement should reside in the self-regulating domains of medical professions, has impeded the potential real improvement that the identification of physician practice variation can spawn.

The purchaser/payer and provider worlds have increasingly collided over physician practice measurement in the past several years. Figure 1 (separate attachment), contemplates the variables associated with physician performance measurement and predicts the likely intensity of physician response to such measurement.

Best received (but perhaps least effective) is privately shared information regarding discrete process measures of quality performance, with no transparency of measurement beyond the individual provider. Most poorly received are efforts that seek to combine resource use with composite process and outcome measures to tier physicians into preferred networks or exclude them from health plan participation altogether. Additional factors that have frequently given rise to substantial provider resistance to purchaser/payer based measurement efforts were:

- Methodological issues, especially disbelief in the validity of the rankings based on small "n" (numbers of measurement opportunities), episode of care attribution, lack of transparency of measurement methods, and lack of validation studies
- Relative lack of due process, which for physicians means the ability to review results in detail prior to release more broadly, ability to request reconsideration of tiering status, and a mechanism for appeals to a third party arbiter in unusual cases
- Use of measures for pay for performance and public transparency
- Unilateral creation of measures; lack of specialty society endorsement

Responding to similar physician concerns, the New York Attorney General pursued and obtained agreements from local and national health plans as to conditions surrounding individual physician measurement and reporting. Those health plan agreements included provisions to:

- Ensure that physician rankings are not solely cost-based;
- Use established national standards, including those endorsed by the National Quality Forum, to measure quality;
- Incorporate measures to foster more accurate physician comparisons;
- Disclose to physicians how rankings are designed and provide a process to appeal incorrect ratings;
- Disclose to consumers how physicians are ranked and provide a process for consumers to register complaints about the system; and
- Nominate and pay for a ratings examiner -- subject to the attorney general's approval -- to oversee the ranking program's compliance activities

The Robert Wood Johnson Foundation funded a George Washington University study on the legalities of provider measurement and tiering[1]. That study found no legal prohibition from engaging in such activity, but also noted that undertaking such activity in an opaque manner, with "reckless" methods and without due process could result in a backlash that may have legal merit.

---

[1] Rosenbaum S, Kornblet S, Borzi P. An Assessment of Legal Issues Raised in "High Performing" Health Plan Quality and Efficiency Tiering Arrangements. http://www.rwjf.org/programareas/resources/product.jsp?id=22571&pid=1142

Transparency and continuous improvement of measurement are necessary prerequisites to transparency and continuous improvement of performance. To this end, many measuring entities are now more attentive to measure and measurement improvement.

None of the efforts to optimize care will succeed without enhanced clinician engagement. Physician specialty societies are increasingly playing a leadership role in measurement and improvement activities. Included have been calls for increased professionalism – calling not only for the maintenance of clinical competence but also collaboration with other professionals to reduce medical error, increase patient safety, minimize overuse of healthcare resources, and optimize care outcomes.[2]

With the increasing deployment of clinician level measurement in the domains of process quality, resource use, and to a lesser extent, outcomes, wisdom is accumulating as to how measurement and improvement in each of these domains might be enhanced to improve the reliability and actionability of measurement results, facilitating greater physician acceptance and broader engagement in performance improvement. What follows is a domain-specific (process quality, resource use, and outcomes) discussion of the issues, some observations about experience in the field to date and recommendations for ways to improve the measurement processes.

III. Process measures

Donabedian spoke of process measures in 1966. He wrote, "*This approach [process measurement] requires that a great deal of attention be given to specifying the relevant dimensions, values and standards to be used in assessment. The estimates of quality that one obtains are less stable and less final than those that derive from the measurement of outcomes. They may, however, be more relevant to the question at hand: whether medicine is properly practiced.* [3] "

The contemporary use of process measures usually involves HEDIS-like measures that use software algorithms to parse electronic administrative data – usually claims data – into

---

[2] ABIM Foundation. Medical Professionalism in the New Millennium: A Physician Charter, Annals of Internal Medicine. 136:3 243-246. February 5, 2002 http://www.annals.org/cgi/content/full/136/3/243

[3] Donabedian A. Evaluating the quality of medical care. 1966, reprinted 2005. The Milbank Quarterly. http://www.milbank.org/quarterly/830416donabedian.pdf

numerators and denominators. The typical denominator represents the instances in which patients fulfilled criteria for the performance of a particular service thought important to the care of specific patient conditions. Perhaps the simplest example of denominator criteria would be age-range specific recommendations for a particular immunization. The numerator criteria would be whether or not the claims data included evidence that the immunization was billed and paid (and therefore assumed to have been administered). Many process measures involve complicated denominator logic requiring a minimum length of health plan eligibility, extensive diagnoses codes specifications for inclusion in the denominator and often extensive diagnoses and / or procedure codes defining instances where patients' clinical situations exclude them from the denominator.

Given that the provision of recommended care remains significantly suboptimal, process quality measures continue as highly relevant tools for identifying deficiencies and driving care improvement.

Implementation issues regarding process quality measurement

- **Binary numerator logic.** For each patient that is included in the denominator of a particular measure, the software looks to see if the numerator criteria are met, with the result being 'yes' or 'no'. A 'yes' answer is highly reliable – the desired service was performed by evidence of a claim in the dataset. A 'no' answer has two possible meanings. One is that the service was not performed and therefore no claim was generated. The other possibility is the service was rendered, but no claim for the service was present in the claims dataset. *Real world examples: 1. Patient has dual insurance and the required claim was billed to her other insurance company; 2. Software algorithm missing a new CPT code in the numerator logic that would have indicated that the required service was provided.*
    - **Potential mitigation or improvement.**
        - **Improved electronic numerator data capture.** If a provider or provider group has electronic systems capturing numerator events, such data could be added to the measurement dataset

        - **Willingness to troubleshoot software programming if patterns of missing data are identified. Third party validation of algorithmic logic.**
        - **Aggregation of data across payers.**
        - **Detailed patient and quality measure reporting; reconsideration and attestation processes.** If physicians are provided with detailed reports of patients meeting denominator criteria and whether or not each patient met the numerator criteria, physicians or their staff may review the clinical record for evidence of performance of the required service. If evidence is found, physician may add the missing data by contacting the plan or other measuring entity, attesting to meeting the criteria. Plan then recalculates performance results and updates dataset. The plan reserves the right to audit physician data supplementation.
- **Denominator logic errors.** Inclusion / exclusion criteria may have programming errors, leading to over or under-inclusion of patients for whom a quality metric might apply.
    - **Potential mitigation or improvement.**
        - Troubleshooting algorithmic logic when frequencies of patients meeting denominator logic are highly variant from typical benchmarks. Third party validation of denominator logic.
- **Assuring sufficiency of observations ("n value") for performance reporting.** Physician measurement programs have varying minimum "n" requirements. The minimum requirements to meet the same statistical test vary based on a range of factors, including: whether an individual practice or practice group are being measured; measure characteristics; and the level of differentiation that will be applied based on the measure. Physician concern has been expressed about the minimum requirements of most health plan programs. Some plans have used as few as 10 cases (instances when denominator criteria are met) across *all* process measures (i.e., as few as 1 case for each of 10 process measures generate a physician's overall process quality score). *Reliable clinician level measurement of process indicators is highly constrained by the available number of observations.*

- **When measurement results are used for performance ranking or tiering.** When results are used to create a ranking or tiering of physicians, an evaluation of the statistical chance of misclassification into the wrong placement in the ranking or tier may be more meaningful than a set minimum number of observations. In such a case, the minimum number required to meet a misclassification risk statistic may vary between measures.
    - **Potential mitigation or improvement of number of observations and misclassification risk.**
        - Consider consensus standards for the minimum number of observations on a per indicator basis, a total indicator basis and/or the acceptable level of misclassification risk. Such standards appear necessary to gain strong clinician acceptance and engagement, especially when used for P4P, tiering, and transparency.
        - Data aggregation across payers may provide an increased number of observations
        - Self-reporting by physicians or physician groups that capture process indicator performance across payers
- **Applicability of process measures to physician specialty.** Measurement programs vary as to which process indicators apply to physician practice specialties. Most create specialty specific measure subsets. Some deploy differential weighting of indicators depending on specialty. For example, an endocrinologist's performance results on HbA1c may have more impact than screening mammography. Such a schema may have disproportionate impact at the individual physician level. For example, a female endocrinologist within a group of mostly male endocrinologists may see a disproportionate number of women in her practice. Certain aggregate scoring methods may place her at a total score disadvantage if she has more lower-weighted quality indicators such as screening mammography and pap smears in her quality indicator mix.
    - **Potential mitigation or improvement.**
        - Consider consensus standards for the applicability of process measures to specialties
        - Evaluate unintended consequences of differential weighting of process measure performance

- **Attribution of process measures to clinicians.** Beyond determining which process measures are applicable to which specialties, there exists multiple methods by which patients' quality indicator results are assigned to individual physicians. In general, attribution of these indicators may be episode-linked or panel-linked. In episode-linked attribution (only possible if episode of care software is also applied to the dataset for cost of care measurement) quality indicators for certain conditions are attributed to the physician to whom episode(s) of care for that condition are attributed. So a physician who is assigned a patient's diabetes episodes may also be assigned that patient's diabetes process measures. Ostensibly this is done to attempt to link condition specific cost and quality performance. Panel-linked attribution is more common, and HEDIS measures fall into this category. In panel-linked attribution, the presence of the patient in a physician's "panel" (i.e., the patient was seen at least once -- sometime twice is required -- by the physician and a claim to that effect is present in the dataset) makes the physician eligible for assignment of that patient's quality indicator(s) (again assuming that the indicator is applicable to the physician's specialty). If more than one physician in an applicable specialty saw the patient, additional logic is required unless the measuring entity applies what is referred to as "team-based attribution." Under team-based attribution, all physicians of applicable specialties for the patient's quality indicators receive credit if the measured services are provided, and conversely all would be "debited" if the services were not provided. In such a case a PCP would receive credit if the patient's Ob/Gyn provider ordered and the patient obtained a screening mammography. In non-team-based approaches logic drives assignment to only one provider based on rules such as the presence of relevant diagnosis codes on physician claims, or to the physician who saw the patient more often. *Real world example: A PCP sees the patient and codes the visit as preventive care. The patient's gynecologist sees the patient for urinary tract infection and codes the visit as such, but also orders screening mammography. There is no claims record of a mammogram being ordered by the PCP, and the PCP is "debited" for the screening mammography process measure.*
    - **Potential mitigation or improvement.**
        - While episode-linked attribution might provide a degree of cross-correlation of condition-specific cost and process quality performance, the

controversy surrounding episode attribution itself would seem to make panel-based attribution less controversial. Panel-based attribution must be used for HEDIS measures in any case, due to the HEDIS methodology requirements

    - Team-based attribution provides benefit of the doubt to individual physicians whenever a process quality measure requirement is met. However, when the criterion is not met, "blame" is shared. Not surprisingly, physicians prefer team-based attribution in the former case, but not in the latter.
    - Consensus standards development may be appropriate for this issue

- **Minimum number of applicable measures for specialties.** Studies by the author and others have shown that if fewer than three process measures are applicable to a specialty, a "normal" or "bell-shaped" distribution of the results are highly unlikely within that specialty. In addition, a small number of indicators inherently places disproportionate importance on a few measures that may not fully represent the breadth of practice process quality within that specialty.
- **Measurement period.** The time frame from the beginning to the end of the data being analyzed is referred to as the measurement period. Two related implementation issues arise with respect to the measurement period and these are applicable to both process quality measures and resource use measures. The first is the length of the measurement period. The second is how recent or "fresh" the data is. Longer measurement periods provide more cases (or episodes, in the case of resource use) for analyses and reporting by increasing the "n" values. However, as the measurement period increases, the "staleness" of the data increases, and recent improvement of performance may be missed. It is important to capture recent data in the analysis; the degree of physician concern regarding the analysis is correlated with the age of the data set.

IV. Resource use measures

As with process measures, healthcare resource use measurement has been enhanced in recent years by improved informatics capabilities. Improvements in data processing speed, lower-cost

data storage leading to the creation of massive data warehouses, and episode of care analytic methods have provided new ways to characterize resource use than in the past.

Past characterization of resource use typically was focused on units of service, such as the number of specific types of services rendered based on CPT codes, as well as categorical resource use such as office visits, hospital bed days by service line (medical, surgical, ICU, etc.), outpatient surgeries, lab, X-ray and pharmacy. In addition to raw counts of services, rates of services (e.g., hospital bed days per 1000 health plan members per month), and per capita resource use (e.g., cost per health plan member per month) have long been tracked and frequently were reported to physicians and physician groups. The focus of such reporting was often on outliers – for process of care measures the focus was usually on low outliers representing under provision of recommended care, while for resource use measures the focus was most often on high outliers potentially indicating overuse of resources.

Historical resource use reporting in this fashion was problematic in that it was provider rather than patient focused. Frequently such reports lacked practice case-mix or patient level risk adjustment. Thus, outlier status alone was not enough to characterize a practice as an inappropriately high user of resources. Such a determination required more intense clinical review of the nature of the practice, the case-mix of treated conditions, and the application of evidence-based practice guidelines where available. Unless such a time-consuming and costly focused review occurred with the involvement of the practicing physician, such results were often dismissed by the measured physician with the assertion that "my patients are sicker."

The healthcare informatics industry responded to these challenges with the development of episode of care analytics. These new approaches were patient and condition centric, and methodologically improved resource use measurement to get closer to an "apples to apples" comparison.

As compared to the "raw" reporting of aggregate resource use, these episode of care approaches advanced resource use reporting in several ways:

- The unit of analysis was at the patient condition level, and the most widely used tools offered over 500 separate conditions, inherently allowing for case-mix variations within a practice
- The tools captured resource usage across all providers for specific patient conditions

- By further adjusting for patient and disease factors, the final output of these proprietary measurement systems based on episodes of care were thought to significantly diminish the credibility of the "my patients are sicker" argument.

Many health plans subsequently deployed these proprietary tools in their physician practice measurement and reporting systems. Physicians were quick to dismiss the tools as "black boxes" as the analytic processes were complex, difficult to understand, and peer-reviewed literature as to their proven utility to drive more efficient and appropriate use of resources was sparse or non-existent. They were further perplexed when they received "report cards" from multiple health plans, with at times widely differing depictions of their episode-based cost of care performance.

Earlier in this decade the stakes of physician measurement were dramatically increased when large employers and their health benefit consultants began to embrace the theoretical construct of "high performance networks" as a means to drive value-based purchasing of health plans and physician services. This construct hypothesized that using administrative (claims) data, process quality of care algorithms, and episode of care approaches, that "higher value" physicians could be identified, and if employees shifted their care to higher performing physicians, improvements in quality and reductions in cost trends would follow. Health plans scrambled to meet the resulting new employer-driven requirements by creating tiered networks, benefit differentials, physician-level quality and cost rankings in provider directories, and inclusion of relative resource use in pay for performance programs. These increased stakes have brought about much more active physician engagement and broader stakeholder scrutiny of current episode of care approaches. From this certain observations can be made as to the ongoing challenges for broad physician acceptance of episode of care measurement.

It is helpful to view episode of care measurement as consisting of three sequential phases: pre-grouper[4] processing, grouper processing, and post-grouper processing. Implementation

[4] The term "grouper" is used generically to denote the episode of care software products. These tools are referred to as groupers because they sift through claims and group claims together at the patient and condition level. Grouper methodology is not discussed at length here, but can be reviewed at the vendors' web sites. Simply put, the groupers identify the first claim that initiates an episode of care for a patient condition and the last claim that ends the episode of care, and aggregates all claims from all providers in the data set belonging to that episode of care for a specific patient's specific condition. The definition of each condition is unique to each vendor's product. Chronic diseases typically are evaluated as multiple year-long episodes.

issues related to clinician level episode based resource use measurement are discussed in relationship to each of these phases.

Implementation issues associated with pre-grouper processing

- **Provider identification.** Payer provider databases are complex and almost always of unique architecture. Identifying providers across health plans is especially problematic, as until recently, there have not been universally unique provider identification numbers. Even within a single health plan a provider might have multiple provider id numbers if he/she practiced in more than one office location or was affiliated with more than one practice group. In addition, physicians display significant mobility over the period of measurement, creating challenges to the full capture of their practice data. Significant progress has been made by vendors via the creation of id number crosswalks and the use of probabilistic algorithms for provider identification.
- **Patient identification.** Within a given health plan, patient identifiers tend to be more reliable than provider identifiers. Data aggregation across payers may be a challenge if cross plan identifiers such as Social Security numbers are unavailable
- **Data completeness.** Many measurement efforts routinely exclude mental health and substance abuse data. The two most common reasons for this are the greater degree of privacy afforded such information and the fact that many health plans "carve out" their mental health benefits management and claims payment to third parties. These third parties may be unwilling or unable to submit data files for analysis. This is unfortunate given that some mental health disorders are serious comorbidities that affect resource use. Less commonly, but still frequently, especially for historical Medicare data, is the lack of pharmacy data. Pharmacy "carve-outs" with administration by third parties are the most common causes of missing pharmacy data. Every effort should be made to obtain and include pharmacy data as medications are an integral treatment option for many episode types, and those episode results may be inappropriately skewed in the absence of such data. Some payers create "plug" numbers (frequently an overall expected average from benchmark data or from those episodes including pharmacy data) if the data are not available. While this may reduce the skewing effect, obtaining actual costs provides for more reliable assessment and reporting.

- **Mitigation of pre-grouper issues.**
    - The ability for a physician to reconcile his/her reports at the patient level is important to make sure that patients and their episodes are assigned to the correct clinicians. Physicians seeking to validate their performance measurement results are hobbled when the patients included in their performance assessment are de-identified. It is strongly recommended that measurement programs provide this functionality. There are countervailing patient privacy concerns that have led some data aggregators to de-identify patients in reports provided to physicians. Some have created independent, secure websites accessible only to the patient's physician of record to provide lookup capability as to the actual identity of patients comprising the physicians' episode set. This applies to process of care measures as well.
    - All efforts should be made to acquire a complete data set for analysis. There should be transparency as to which data sources are unavailable, and ideally the impact of that lack of availability should be assessed and described.

Implementation issues associated with grouper processing –

- **Complexity.** Each grouper has unique and proprietary grouping mechanisms representing substantial investments by grouper vendor companies. In essence, the groupers map the codes present on provider billing forms (HCFA-1500 and UB-92) and pharmacy claims to unique episodes. They are able to track multiple episodes for patients, including concurrent episodes. Given the complexity of the underlying logic that assigns claims to episodes, providers may apply the "black box" moniker. This complexity frustrates physicians and their practice administrators when they seek to deconstruct episodes in a fashion similar to process quality measures. It is a difficult task even for the experienced. It is especially difficult for primary care specialties, where there may be 75 or more episode types present in the analysis. Certain specialties may have fewer than 10 episode types appearing in their reports.
- **Patient comorbidities and disease severity risk adjustment.** Groupers deploy varying techniques to attempt to normalize for variations in patient factors such as comorbidities and disease severity. One can legitimately assert that each physician's practice is unique. It is unique because of case-mix, patient and disease factors but also because each

physician practices medicine in his/her own unique way, calling upon training, experience, and individual clinical judgment. For many conditions, physicians may vary widely as to the resources expended in the treatment of those conditions. The resource use measurement process attempts to identify variation from peers that may present an opportunity to more appropriately use fewer or more resources for specific conditions. In order to best characterize those opportunities, variations in resource use due to patient factors, including their disease severity and comorbidities, must be addressed and normalized to the extent possible. The groupers attempt to do this using a variety of sophisticated approaches. One approach builds in disease severity staging and further adjusts raw results by calculations involving the patient's overall health and illness factors. Another creates separate numerical adjustments for patient factors as well as for each comorbidity. It is fair to say that the current approaches are not perfect, that they are continually improving, and do allow an assessment of potential unwarranted practice resource use variation. The current systems may in certain circumstances require individual clinical review to augment automated assessment processes.

- **Mitigation of grouper processing issues.** Unfortunately, the nature of what groupers do is indeed complex. Mitigation of this frustration is best accomplished by education, ideally with physician specialty society engagement. Specialty societies may be able to create a centralized expertise for their members, helping to educate and providing advocacy when methodological errors or undue methodological compromises are suspected.

Implementation issues with post-grouper processing

- **Assuring sufficiency of observations ("n value").** As with process quality measurement and reporting "N" size is a concern with resource use measurement analyses and reporting,. The ability to differentiate physician performance is highly correlated with the number of episodes studied. Figures 2a and b represent work done by Bill Thomas[5]. Using varying sample sizes of 10, 20, and 50 episodes and for three different physicians, he was able to demonstrate that reliable differentiation between physicians' performance

[5] Thomas JW. Enhancing Validity of Physicians' Economic Profiles. Presentation at Academy Health's Annual Research Meeting, Seattle, WA. June 2006.

results was related to the number of observations. The requisite number of observations for reliable measurement may well vary by type of episode, and further research is required. The question of the requisite minimum number of episodes occurs at the individual episode level, at the aggregate episode level for each physician, and for the specialty/episode type combination (the comparative group). As episode size minima increase, the fewer the number of physicians whose performance is reported. Again, when results are used to create a ranking or tiering of physicians, an evaluation of the statistical chance of misclassification into an incorrect placement in the ranking or tier may be more meaningful than simply depicting the underlying number of episodes. In such a case, the minimum number of observations required to meet a misclassification risk statistic may vary between episode types.

- **Peer comparison group assignment.** This has been a problematic area for health plans deploying episode of care measurement. In many instances their provider specialty databases were out of date. In other instances, certain specialty types were highly variable as to underlying case-mix. For example, the specialty of cardiology may be quite heterogeneous. Some cardiologists are generalists, some are interventionalists, some electrophysiologists. There exists a broad continuum from practice to practice as to the volume of procedures performed. Another example is orthopedics, where in urban areas orthopedists may focus their practice on arthroscopy, joint replacement, and ankle or hand surgery. This returns us to the "every practice is unique" discussion. Some measuring entities use case-mix to define the comparison specialty group, but there are limits to how many practice specialty peer groups can be created and still have enough physicians within the peer group for comparison of performance.
- **Attribution of episodes to physicians.** Episodes of care, once created, are typically then attributed to "responsible" clinicians. Multiple approaches to attribution exist, and each is a compromise to some extent. In an HMO gatekeeper model life was a bit simpler as an assigned PCP could be assigned responsibility for overall care of a patient. Given the current predominance of the PPO model and the mobility of patients between practitioners, identifying a single physician to whom responsibility for a patient's condition resides may be problematic. In non-Medicare populations approximately 85% of episodes involve only one managing clinician. It is therefore the remaining 15%, and

for Medicare likely around 35%, that presents potential attribution challenges. The most common method for attribution in such cases is to assign the patient to the managing clinician (exempts radiologists, pathologists, etc.) who provided the majority of physician care. In many instances a minimum percentage is required for attribution, typically 25-30%, and various tie-breaker rules exist. *Real world example: a cardiac electrophysiologist places a pacemaker in a patient, seeing the patient the night before and the day after pacemaker placement. Other physicians were responsible for the patient's ongoing course, which was complicated and the hospital stay prolonged. The episode was attributed to the electrophysiologist because his billing generated the greatest per physician share of the aggregate professional costs.*

- **Outlier handling.** While the groupers have some outlier handling capability, many entities handle outliers in the post-processing phase, allowing for more sophisticated approaches. Extraordinarily high or low episode resource use within an episode often warrants closer evaluation of the episode via review of the underlying claims and clinical situation. Some instances may warrant exclusion or truncation of the value at certain maxima. When a physician's average resource use across all episodes is unusually low or high, a review of the underlying episode mix and peer comparison group assignment may be warranted. *Real life example: A family practitioner's average episode costs were significantly higher than others within his specialty. His case-mix review showed a high frequency of HIV episodes. The physician's special interest in the treatment of HIV patients skewed his results when compared to other family practitioners.* When results are used for tiering, transparency, or P4P, every effort should be made to understand outlier results that would place clinicians in the lower performing tier(s). This requires clinician engagement in review of his/her results, and may require a conference between the measuring entity's medical director and the measured clinician.
- **Comparison against benchmarks.** The measuring entity will typically compare grouper generated results against vendor benchmark values at various stages of grouper and post grouper processing to be alert for methodological errors. When working with new data sources, one can expect multiple iterations prior to final reporting.
- **Levels of reporting.**

- **Episode level reporting.** Those working with detailed episodes of care data involving a large mix of episode types (typically primary care) can attest that an individual physician's comparative performance reflects a mix of above and below average performance. Even the highest performing clinicians typically have opportunities to reduce unwarranted variation. While lower performing clinicians are likely to have more improvement opportunities, they may find the measurement process more acceptable if their higher performing episode types are identified along with the lower performing areas.
- **Condition based clinician level reporting.** Clinically it is logical to aggregate certain diagnoses or condition types together to create condition-based clinician level reports. These types of reports are of most use to patients with the measured conditions. Such an approach may be a more inclusive one when it comes to the clinician network. An overall average performing clinician who demonstrates high quality and efficient resource use in the treatment of his / her diabetic patients probably should receive a disproportionate share of diabetic patients. Especially if capacity is an issue, tiering may at times be more effective based on conditions than overall aggregate physician performance,
- **Aggregate (overall) based clinician reporting.** Given what we know about each physician's practice being a mix of above average and below average performance across episode types, one might question the value of aggregate based performance reporting and tiering. To date employers and purchasers have requested that aggregate clinician performance be measured and reported to encourage overall accountability. While overall accountability is critical, physician engagement and improvement, however, is likely to require engagement at the condition level, with upside to areas of high performance and targeted improvement required in areas of low performance. Healthy patients are perhaps more interested in physicians' performance in preventive care and health maintenance than how diabetics are treated within a practice.
- **Point estimates and confidence intervals in reporting.** Those measurement efforts that have involved physicians in their measurement activities have gravitated to reporting results as point estimates with associated confidence

intervals. Doing so recognizes the fact that probabilistic theory plays a large role in the genesis of results. Assigning scores or rankings without recognizing the inherent imprecision involved diminishes the reputation of a measurement effort. Humility about the measurement tools and methods teamed up with passion to improve care at the *highest common denominator, the patient,* are critical success factors for physician acceptance and engagement.

V. **Outcome measures**

Donabedian offered insight in 1966 as to the challenges of attribution of outcomes measures. He wrote, "Many advantages are gained by using outcome as the criterion of quality in medical care. The validity of outcome as a dimension of quality is seldom questioned…"

"…Many factors other than medical care may influence outcome, and precautions must be taken to hold all significant factors other than medical care constant if valid conclusions are to be drawn. In some cases long periods of time, perhaps decades, must elapse before relevant outcomes are manifest. In such cases the results are not available when they are needed for appraisal and the problems of maintaining comparability are greatly magnified."[6]

Donabedian was focusing on factors outside of medical care in this discussion. We can extend his arguments further into the variability of medical care itself. In many instances outcomes measures, to a much greater extent than process and resource use measures, are dependent upon elements of care beyond the clinician's control. Poor post-operative care processes and inadequate institutional staffing can imperil even the best surgeon's efforts. Laboratory errors can lead to patient mismanagement. Multiple physicians may be involved the course of a patient's treatment.

This is not to say that the complexity of outcomes measurement precludes accountability, even accountability at the clinician level. At one extreme is the case of wrong site surgery. While one could simply say that a hospital's system of care failed in such an instance, to do so abdicates the ultimate responsibility of all *individual* care professionals to patients under their care. Towards the other end of the spectrum may be the institution whose care processes show substantial variation from best practices and in aggregate those variations produce poorer

---

[6] Donabedian, *ibid.*

outcomes. The responsibility of all involved individual care professionals to seek and drive improvement remains.

Care outcomes vs. clinical outcomes – broadening our definition of outcomes. Outcomes from the consumer and patient perspective may vary significantly from the medical community's perspective, although the rise of consumerism and patient-centeredness may at some point create more of a confluence. For now, let us consider care outcomes as a larger universe than clinical outcomes, encompassing such elements as consumer and patient experience, quality of life, and functional status. Certain patient-assessed care outcomes may be attributable at the clinician level, such as clinician-specific satisfaction surveys. Other elements such as quality of life and functional status will require more complicated approaches to attribution.

Implementation issues of outcome measures

- **Complexity of clinical outcomes attribution.** As discussed briefly above, clinical outcomes, especially those remote from the related care delivery, may represent significant challenges to attribution of accountability. While attribution of acute care outcomes in patients with few comorbidities and other confounding factors may be reasonably straightforward, patients with chronic disease(s) often interface with multiple care providers, and the outcomes of interest may not be manifest for months or years. Clearly patient compliance as well as socioeconomic and cultural factors may play an operative role requiring sophisticated risk stratification and adjustment. Such sophistication is in turn linked to the need for robust electronic capture of relevant confounding factors and in many instances, as yet developed analytic methods. And as more robust methods become available and analytical subsets are defined, the number of observations in each subset declines, complicating statistical inference.
- **Need for enhanced electronic data capture relevant to outcomes.** Once one moves into the realm of outcomes measurement and reporting, electronic capture of clinical information beyond that contained in administrative data is often necessary. While many have expressed interest in more robust outcomes reporting, the relative lack of necessary data in electronic format has created cost barriers to more widespread outcomes analyses. Those who anticipated that electronic health records would obviate this challenge have at least initially

been disappointed by the capabilities of electronic systems in this regard. Substantial efforts are underway through the Quality Workgroup of America's Health Information Community to interface with other elements of the Quality Alliance Steering Committee to better define and more quickly deploy additional electronic clinical quality capture and reporting capabilities.

- **Need for better understanding of systems performance.** The behavior of health systems has only relatively recently been a subject of intense analyses. Performance measurement tools that provide actionable information and improve accountability at the system level are needed. As with physicians' behaviors, modification of systems' behaviors to support the goals of higher quality and more cost-effective care is no easy task, especially, as mentioned previously, where existing revenue generation models are imperiled.
- **Need for more extensive consumer / patient data and engagement; evolution to consumer / patient "performance" assessment.** Given the importance of patient factors such as socioeconomic status, personality type, health literacy, cultural preferences, and family and support system dynamics to health system interactions and behavior and many if not most outcomes, such additional information would advance the capabilities of outcomes measurement. We have seen the development of "patient health records" (PHRs) of varying design, capabilities, and integration into administrative and clinical data repositories. To date the dominant model has focused on information "push" capabilities in support of healthcare consumer empowerment. Enriching outcome analytics will depend on greater consumer / patient engagement and information "pull" capabilities. One can envision the potential for more sophisticated risk adjustment by more clearly identifying operative patient variables affecting outcomes. Comprehensively characterizing outcomes may well require complementary assessments of the contributions of individual clinicians, the care system, and consumer / patient factors and behaviors.

Perhaps some clinicians would agree that, to the extent that outcomes measures are in close proximity to the related care delivery, provision of care is dominated by identifiable clinicians, that care processes are largely controlled by those clinicians, and patient and disease factors may be reliably adjusted for, outcomes attribution at the clinician level may be appropriate.

Purchasers of care and many consumer advocacy groups would strongly agree, and would further argue that accountability for care outcomes need not be a perfect science to save lives and

improve care. Rather blunt tools deployed with public transparency have been shown to improve hospital performance. While most stakeholders would agree that we need better outcomes measurement systems, one can expect to see continued contentiousness as to how good is good enough.

VI. Conclusion

So we return to the two separate worlds – clinicians seeking precision and actionability while purchasers, payers and consumer advocates demand measurement, transparency and accountability.

We find ourselves in a situation where both sides are right – and wrong. Such situations are best dealt with by identifying overarching areas of agreement and alignment around common goals.

Optimizing patient care, as earlier defined, is a unifying theme for all stakeholders. Measurement, transparency and accountability efforts need to drive clinician engagement and improvement activities to be most effective. Clinician engagement and subsequent clinical improvement is dependent on the reliability and actionability of measurement and the alignment of practice revenue with performance.

Our understanding of the implementation issues related to clinician level measurement has increased substantially over the past few years. This understanding should enable ongoing refinement of clinician level measurement, improving actionability and the potential for actual clinical improvement. It is hoped that this improved understanding will not only lead to improved data capture, improved analytics, improved methodology and reporting, but also to informed compromises to drive improvement while awaiting better tools and approaches.