

Review of the Massachusetts Group Insurance Commission Physician Profiling and Network Tiering Plan

A Report to the Massachusetts Medical Society

Robert A. Greene, MD, FACP

Howard B. Beckman, MD, FACP

Gregory H. Partridge

Focused Medical Analytics, Rochester, NY

J. William Thomas, PhD

*Muskie School of Public Service, University of
Southern Maine*

November 17, 2006

Table of Contents

I. Executive Summary	Page 3
II. Introduction and Objectives	Page 8
III. Observations That Raised Concerns	Page 13
IV. Data Accuracy	Page 15
V. Cost-Efficiency Measure	Page 18
VI. Quality Measurement	Page 25
VII. Issues Related to Tier Assignment	Page 32
VIII. Process Improvement Opportunities	Page 38
IX. Conclusions	Page 40
Appendix I: Summary of Recommendations	Page 42
Appendix II: Review of the Resolution Health Quality Measures	Page 45
Disclosures and Endnotes	Page 51

Executive Summary

i. Introduction and Scope

To address rising health care costs and variations in the quality and safety of services delivered to employees and their dependents, the Massachusetts Group Insurance Commission (GIC) joined with Mercer Human Resources Consulting (Mercer) in 2003 in establishing the Clinical Performance Improvement (CPI) Initiative. Because of concerns voiced in the Massachusetts physician community about the performance measurement and tiering processes being utilized, the Massachusetts Medical Society (MMS) engaged Focused Medical Analytics (FMA) of Rochester, NY and J. William Thomas, PhD from the University of Southern Maine to examine these methodologies and, if appropriate, make recommendations for improving them.

The purpose of this report is to identify aspects of the CPI Initiative that might be improved in order to more effectively promote quality improvement and appropriate cost control. The CPI Initiative involves (a) constructing a consolidated, multi-plan claims database, (b) using the database to construct cost-efficiency and quality of care profiles for physicians, and (c) use of the profile information by health plans to partition their physician networks into preferred and non-preferred tiers. Underlying the CPI Initiative is the significant presumption that creating a tiered network is an appropriate and effective method for bringing about the GIC's goals. Analysis of that basic assumption was outside the scope of this report and is not addressed.

The report reviews the observations that raised concern about the CPI Initiative, and then discusses the areas of data accuracy, cost-efficiency measurement, quality measurement, tier assignment, and administration-physician process improvement. Information gathered for this report was based on interviews with Mercer and Resolution Health Inc. (RHI), as well documents provided to the MMS throughout the history of the CPI Initiative. Early in the process FMA requested claims level detail for approximately twenty physicians who wanted to better understand their cost measures. As of the date of this report, FMA continues to work through the administrative requirements for obtaining access to the data that produced those results. The Massachusetts Medical Society made a draft of this report available to Mercer, the GIC, and Resolution Health for fact checking and comment prior to its release.

ii. Data Accuracy

In the CPI Initiative, physicians' quality and cost efficiency profiles are developed from the health care claims databases of participating health plans. Data accuracy is critical to performance evaluation. Lack of accuracy decreases the usefulness of the evaluations for all stakeholders, increases the risk of unintended consequences, and increases physician distrust of the system.

How could profiling data based on paid claims be inaccurate? The problem is that the claims payment process will tend to promote accuracy only in those elements necessary to pay the claim, such as procedure codes and knowing who billed for the service. But performance measurement depends on other elements, such as diagnosis codes and knowing who actually ordered or performed a service. Although the GIC has worked with plans specifically to obtain other data, such as diagnosis codes, those data are not necessarily involved in claims payment and their accuracy must be evaluated separately. Information may be inaccurate or even missing from a database consisting of paid claims, thereby introducing inaccuracy into profiling calculations.

Most of the work on the accuracy of primary data has been left to the health plans. We suggest several areas in which plans and Mercer should pro-actively investigate data accuracy. These include the ordering physician information on radiology and pharmacy claims, coding of diagnoses, and claims related to physicians with unusually high or low scores. Providers also need to submit accurate claims, while at the same time navigating an extremely complex medical policy and claims system. In spite of all efforts to deal with these likely problem areas, increasing data accuracy must be approached as a continuous quality improvement (CQI) project aimed at improving the evaluation system itself.

The key to addressing data inaccuracy is implementing a feedback cycle which accepts possible errors from multiple sources, evaluates them, assesses their causes, and then improves the system.

Recommendation: To help identify systematic accuracy problems, physicians should be given patient-level drilldowns for the efficiency measure, and patient lists for the quality measures. There should be a formal feedback and correction mechanism so that errors uncovered by physicians, plans, and other analysts can contribute to improving the evaluation system.

Using the data for individual tiering exacerbates the impact of accuracy issues. Pharmacy prescribing and radiology ordering information, for example, will be more accurate within a group than for an individual.

Recommendation: Tier at a group level until data accuracy is improved and the methodology is further validated.

iii. Cost Efficiency Measure

Construction of physician level cost efficiency measures from health care claims data involves grouping claims into episodes of care, calculating the total cost of all claims in an episode, assigning episodes to individual physicians, calculating specialty specific expected costs for episodes of given conditions, summing total actual costs and total expected costs for each physician, and using these to generate a measure of cost

efficiency performance for each physician so that physicians can be compared on relative cost efficiency within their specialty.

While Mercer's profiling approach incorporates steps designed to try to assure accuracy, modifications can be made in order to further enhance accuracy of results.

Major Recommendations:

- **Improve cost outlier logic**
- **Attribute episodes of care on professional and prescribing costs together**
- **Enhance risk adjustment of episode expected costs**
- **Instead of using three years of episodes with weighting, use two, unweighted; increase sample size by truncating outlier episodes instead of trimming them, and by scoring physicians in groups instead of individually**
- **Measure costs only in conditions pertinent to a given specialty**
- **Remove the costs of preventive care and other under-utilized interventions**
- **Use a more reliable metric than an actual-to-expected ratio**
- **Assess and control for the effects of benefit differences across plans**

iv. Quality Measurement

Quality measures created from administrative data are based on a set of recommended interventions applied to a population of interest. The interventions and the patients are then attributed to the physician or group whose performance is to be measured. The quality indicator used by Mercer employs a combination of measures developed by RHI and logic developed by RHI and Mercer. The measures and logic chosen by Mercer and RHI are generally reasonable.

Major Recommendations:

- **Provide physicians with lists of patients and interventions counted towards each quality measure**
- **Do not try to increase sample size by adding the number of interventions and the number of opportunities across multiple conditions to create an adherence ratio, because the interventions and opportunities tend to cluster and therefore are not independent**
- **Instead, overcome the problem of small sample size by assessing adherence at a group level rather than at an individual physician level**
- **Set a reasonable target for each measure, rather than leaving the target open-ended and thereby creating an implied target of 100%**
- **If the measure is to be constructed across multiple conditions, use case-mix adjustment**
- **Use a consensus process to weight more important measures higher, rather than use a default of equal weight for all measures.**

v. Issues Related to Tier Assignment

This section discusses the issues raised by using the efficiency and quality scores to assign tiers to physicians or physician groups, and the process of implementing the tiering system.

Major Recommendations:

- **Develop a suggested uniform tier assignment protocol**
- **Require more than a small difference such as 0.01 units to cause shifting of tiers**
- **Use a data-driven process to consider which specialties to tier.**
- **Do not tier physicians whose practices are too new, too small to measure, or different from their peers**

Tiering on an individual basis will result in some offices with practitioners of different tiers. The increased administrative burden of having physicians in the same office assigned to different tiers offsets the potential system savings, especially until accurate physician tiering is validated. This is another factor in favor of assigning tiers at a group level rather than by individual physician at this time.

vi. Process Improvement Opportunities

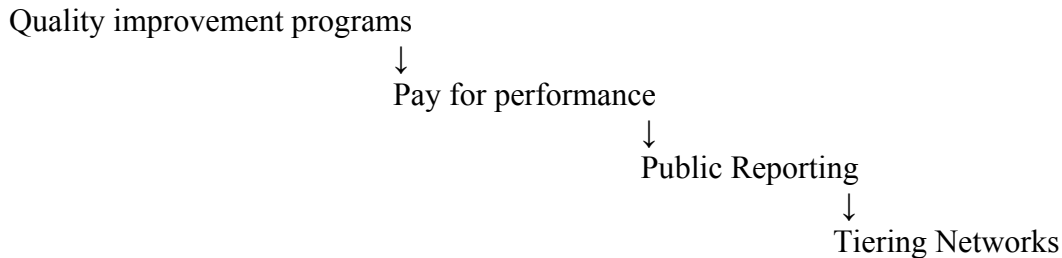
It has been the authors' experience that successful physician evaluation programs, including pay-for-performance, public reporting, and tiering systems, share the characteristic of good process in terms of relationship-centered administration.

Recommendations:

- **Clarify and then widely articulate the program's core values.**
- **Offer physicians meaningful input into the program.**
- **Share results with physicians well in advance of publicly reporting them.**
- **Give practitioners patient-level data and correct identified errors prior to public reports.**
- **Provide physicians with specific behaviors (action items) by which they can improve their results.**
- **Demonstrate and publicize a commitment to continuous improvement.**
- **Anticipate the need for physicians to buy in to consensus-based measures.**

vii. Conclusions

Physician performance evaluation has a spectrum of uses that range from quality improvement through to network tiering:



For each succeeding purpose, inaccuracies have more severe consequences for all stakeholders, and especially for physicians and their patients. Therefore, each successive use requires a correspondingly higher level of rigor to generate sufficiently accurate and usable results.

In order for the GIC to meet its goals, the CPI Initiative should be improved in phases over several years, in a prudent, thoughtful, and deliberate manner. A number of our recommendations involve performance measurement and tiering at a group level rather than at the level of the individual physician. As an interim measure, group level reporting could be a reasonable starting place. As the data and the instrument improve, they may reach a point where individual-level physician measurement, tiering, and accountability become appropriate. We have concerns that the CPI Initiative, as it now stands, does not meet the test of sufficient rigor for the purpose of individual physician measurement.

All parties should realize that physician profiling and accountability only affect one part of health care. This is particularly true for chronic disease care, where a system approach plays an especially important role. Health plans should support physicians through disease and case management. Such support also will help prevent unintended consequences such as avoiding sicker or more difficult patients.

The GIC and Mercer should engage the physician community in improvement of the profiling and tiering system and, ultimately, of patient care itself. By improving the data base, the tiering instrument, and the relationship with the physician community, the GIC is more likely to achieve its goals, which are shared by all stakeholders.

II. Introduction

As benefits administrator for 250,000 State employees, the Massachusetts Group Insurance Commission (GIC) has been concerned for many years about rising health care costs and variations in the quality and safety of services delivered to employees and their dependents. To address these concerns, GIC has implemented a number of strategies – managed care, increased co-payments, preventive health programs – but none of these approaches has been considered fully successful in reducing variation. In 2003, the GIC joined with Mercer Human Resources Consulting in establishing the Clinical Performance Improvement (CPI) Initiative to make quality and cost information available to the public, thereby increasing transparency of provider performance, and to encourage employees and their dependents to become more involved in improving their own health.¹ With this approach, the GIC is seeking to control costs and improve healthcare quality and safety while maintaining comprehensive benefits and a broad choice of providers.^{1, 2} By means of the CPI Initiative the GIC is evaluating Massachusetts physicians on the basis of cost-efficiency and quality of care performance. Furthermore, it is requiring that GIC participating health plans partition physician networks into preferred and non-preferred tiers on the basis of measured performance.¹ Co-payment differentials are used to encourage GIC members to utilize physicians in the preferred tier.

A key component of the CPI Initiative is the construction of a database comprised of three years of claims records from each of the six health plans that serve GIC members. Mercer processes these claims records through the Episode Treatment Grouper® (ETGs) software system from Symmetry Health Data Systems, and uses the results to construct a measure of physician cost-efficiency performance, the efficiency index (EI). Mercer is responsible for database development and for the analyses that underlie the efficiency measure.

The database is also analyzed by Resolution Health, Inc. (RHI) to produce measures of physician quality of care performance. Quality and cost-efficiency data are provided by Mercer to GIC's participating health plans, and the plans use the data as the basis for physician tier assignments.* In 2006, the first year in which physician network tiers were implemented, three of the six plans tiered physicians at the group level, one plan tiered by hospital affiliation (Tufts), one plan (Unicare) tiered physicians at an individual practitioner level, and one plan (Neighborhood Health Plan "Community Care") did not tier at all. Several changes to scoring and tiering methodologies are being implemented for 2007 (e.g., weighting recent data more heavily in the efficiency measure, and removing certain specialties from tiering).

Because of concerns voiced in the Massachusetts physician community about the performance measurement and tiering processes being utilized in conjunction with the CPI Initiative, the Massachusetts Medical Society (MMS) engaged Focused Medical Analytics, LLC (FMA) of Rochester, NY and J. William Thomas, PhD from the

* Harvard Pilgrim's 2006 physician tier assignments are based on analyses of its own claims database.

University of Southern Maine to examine these methodologies and, if appropriate, to make recommendations for improving them. FMA personnel – Robert A. Greene MD FACP, Gregory H. Partridge, and Howard B. Beckman MD FACP – have been involved in physician performance measurement for more than seven years at the Rochester Individual Practice Association, Rochester, NY. Dr. Greene is on the Symmetry Medical Advisory Board, the AQA Physician Performance Measures Workgroup, and the Technical Expert Panel for the AHRQ project, “Identifying, Evaluating, and Categorizing Healthcare Efficiency Measures.” Mr. Partridge has over 25 years of experience analyzing health plan data. Dr. Beckman is a national thought leader on physician-patient and administration-physician relationships, and has published over 30 peer-reviewed papers. He is a member of the board of the American Academy on Communication in Healthcare. Dr. Thomas is Professor of Health Policy and Management in the Muskie School at University of Southern Maine, and has been engaged in research on provider performance measurement for more than 20 years.

Objectives of This Report

The purpose of this report is to identify aspects of the Clinical Performance Improvement Initiative that might be improved in order to more effectively promote quality improvement and appropriate cost control. The CPI Initiative involves (a) constructing a consolidated, multi-plan claims database, (b) using the database to construct cost-efficiency and quality of care profiles for physicians, and (c) use of the profile information by health plans to partition their physician networks into preferred and non-preferred tiers. We note that underlying the CPI Initiative is the significant presumption that creating a tiered network is an appropriate and effective method for bringing about the GIC’s goals. However, we do not comment on the utility or appropriateness of tiering per se as a method of promoting cost efficiency or quality improvement as that question was outside the scope of this report.

In theory, tiering physician networks may lead to cost control and quality improvement in two ways. First, by encouraging health plan members to shift their service use from poorer performing physicians to more cost-efficient, higher quality physicians, an immediate, one-time improvement in cost control and quality performance may be obtained. Second, by motivating all physicians – those selected for preferred tiers as well as those in non-preferred tiers – to work toward identifying and implementing more cost-efficient and/or better quality practices patterns, tiering may promote performance improvement on a continuing basis. Both outcomes depend, of course, on the degree to which variation in physician practice patterns actually influences the cost of care. To be viable strategies, however, both depend on the accuracy of physician profiling information.

Consider the actual case of a pediatrician placed in tier 2 because his case-mix adjusted relative cost of care was nearly twice that of the expected costs of other pediatricians. This ratio is referred to as an efficiency index. Analysis showed that the major driver of his efficiency index was the cost of appropriate care for two patients with hemophilia.

After removing their costs his efficiency index was under 1.0, that is, he was actually more efficient on average than his peers and qualified for tier 1. If health plans implement physician network tiers on the basis of inaccurate profiling information, their differential co-payments may encourage members to shift care from better performing to poorer performing physicians, and no quality or cost-control improvements would be observed. Indeed, in this case the opposite result would have occurred.

In *Crossing the Quality Chasm* the Institute of Medicine (IOM) defined six characteristics of ideal medical care.³ Specifically, the IOM proposed six important goals. These are making care safe, effective, patient-centered, timely, efficient, and equitable. Inaccurate physician performance evaluation has the possibility of interfering with each of these goals:

- Safety may be compromised because of interruption of continuity of care
- Effective care decreases if tiering inadvertently discourages appropriate care
- Patient-centeredness diminishes if physicians are disincented from attending to patient preferences
- Care is less timely when delays occur due to changing physicians, or access is decreased for patients who require unusual or expensive care
- Care may become *more* costly, as illustrated
- Equitable care suffers if patients with high severity diseases are inaccurately assessed higher co-payments, and if physicians hesitate before accepting them as patients

Inaccuracy interferes with the process of promoting change, as well, in that it undermines the face validity of the program. If physicians doubt the importance of selected measures and/or the accuracy of the reported results, they direct their attention and energy to proving the system wrong. If concerns are not resolved, practitioners focus on the monetary penalties or determine how best to undermine the program. The opportunity to engage physicians in quality improvement is lost.

Inaccuracy may also misdirect attention to costs or interventions that were mistakenly reported at a higher or lower than actual rate. The result would be physicians focusing on areas that are not in need of improvement while ignoring the real causes of system inefficiencies.

In summary, information perceived as inaccurate by physicians will not motivate them to seek cost and quality improvements in their practice patterns. In addition to profiling information accuracy, the potential for the tiered networks to promote continuous improvement depends on physician support of program objectives, i.e., on physician “buy-in.” If physicians feel that program goals are not compatible with their own perceptions of quality of care, or if they view the program as being imposed on them by outside forces, they will not actively participate, and the program’s objectives will not be achieved.

The main body of this report is divided into sections reflecting the process of generating tier assignments. The first section discusses data accuracy issues. Primary data accuracy is the foundation of performance measurement,^{4,5,6,7,8} and verification of accuracy is part of emerging standards for physician performance measurement.^{9,10} We next review the cost measure, the efficiency index. The purpose of these discussions is to help identify areas of possible improvement in both the underlying data base and the cost-efficiency metric itself.

The next part discusses the quality measures. For this area Mercer contracted with Resolution Health, Inc. (RHI) to use a subset of RHI's quality measures. RHI's role in the CPI Initiative has been somewhat limited. Mercer provides RHI the three-year claims data set, which RHI then processes through its quality rules and physician attribution logic. RHI returns the results to Mercer, which in turn passes them back to the health plans to help determine physician tiering. This section of the report is intended to highlight aspects of the quality measures that could be modified in order to enhance the measures themselves and also advance quality improvement.

We find it useful to consider both cost-efficiency and quality measures in terms of a second Institute of Medicine paradigm. To address the goals of providing effective, efficient care, the IOM endorses a focus on reducing overuse, underuse, and misuse of medical services.^{11,12} The IOM defines these terms as follows:

Overuse is the provision of a health care service under circumstances in which its potential for harm exceeds the possible benefit. *Underuse* is the failure to provide a health care service when it would have produced a favorable outcome for a patient. With *misuse* an appropriate service is provided, but a preventable complication occurs, and the patient does not receive the full potential benefit of the service.¹²

Decreasing overuse reduces costs by eliminating services that either are not needed at all or for which there are less expensive, equally effective alternatives. Decreasing underuse may increase costs in the short run, but either saves money in the longer term by reducing morbidity, or results in improved health status. Increasing mammography rates is an example. Cost efficiency generally reflects decreased overuse and misuse, while traditional quality measures are most often measures of underuse. Assessments of cost efficiency and quality of care improve when they each appropriately address overuse and underuse. We discuss potential ways to improve handling of overuse and underuse in each of the two corresponding sections. We do not discuss misuse further, since assessing preventable complications is difficult from administrative data and is not specifically addressed in the CPI Initiative measures.

After discussing CPI Initiative methodologies for cost-efficiency and quality performance measurement, we then review the conversion of efficiency and quality scores into tier assignments. We recognize that Mercer is not implementing tiering programs; the health plans design and implement tiering themselves. Thus, Mercer cannot, on its own, address all aspects of tiered network programs that might impact physician buy-in and the ability

of the tiered network programs to achieve continuing improvement goals. Lastly, we turn to the process through which the CPI Initiative has been rolled out in the patient and physician communities, and discuss ways in which improved physician-administration interpersonal process would help all parties.

Information gathered for this report was based on interviews with Mercer and Resolution Health Inc. (RHI), as well documents provided to the MMS throughout the history of the CPI Initiative. Early in the process FMA requested claims level detail for approximately twenty physicians who wanted to better understand their cost measures. As of the date of this report, FMA continues to work through the administrative requirements for obtaining access to the data that produced those results.

Our recommendations are based on our current level of knowledge of cost-efficiency and quality profiling methodologies. We, like Mercer, RHI, and all other organizations engaged in this activity, continue to learn about what works and what does not, about actions that promote improvement and actions that do not, about information that is useful and information that has little value. We expect that a year from now, we will be aware of other aspects of the CPI Initiative that, if changed, would lead to further success in controlling cost and promoting quality of care. The recommendations included here represent our current best attempt at identifying positive CPI Initiative program changes, ones that we believe will enable it to better serve the objectives of GIC, participating health plans, and physicians who serve GIC employees and their dependents.

The Massachusetts Medical Society made a draft of this report available to Mercer, the GIC, and Resolution Health for fact checking and comment prior to its release.

III. Observations That Raised Concerns

After the CPI Initiative was rolled out, a number of physicians reported observations and anecdotes that raised concerns about data accuracy and potential methodological issues affecting their tier assignment. It is important to the context of this report to list some of the most significant and frequently recurring comments. Along with publicly available information, they suggested a number of issues that warranted further examination.

- A pediatrician's efficiency index decreased from almost 2.00 to less than 1.00 after two hemophiliacs were removed from the calculation
- Many physicians noted being evaluated on the cost of conditions that do not occur in their specialty; sometimes those are very expensive conditions which affected their efficiency score greatly
- Extremely high surgical costs or facility costs were observed in conditions where those costs were not expected
- A specialist was thought to be three times the specialty average in cost, yet the most common high-cost procedures in the specialty did not appear in his data
- A family practice physician had an efficiency index of 0.64 in one plan and 1.15 in another; there were more than 30 episodes in each plan (i.e. either would have exceeded the minimum episode number to qualify the physician for measurement)¹³
- A podiatrist had a case mix index (a measure of relative patient severity) of 0.72 for one plan but 7.08 across all plans¹⁴
- Physician efficiency and case-mix scatter plots showed many physicians' scores over 2.0 and as high as 3.0 (two to three times the expected cost or severity)¹³
- Well-child visits and preventive measures were included in the cost-efficiency metric
- Hospital-based physicians assigned to the non-preferred tier were concerned about systematic bias due to patient severity being higher for them than for outpatient physicians with ambulatory practices
- A neurosurgeon was profiled as a general surgeon, raising issues about specialty assignment accuracy
- A pediatrician with over a thousand episodes of pharyngitis on the efficiency score had fewer than 20 in the quality score; and among those, there was a very high inappropriate antibiotic prescribing rate despite an office policy not to give antibiotics for non-bacterial sore throat.
- A group that had received multiple quality awards scored poorly on CPI Initiative quality measures for individual physicians
- A number of physicians felt that lab or other test data was available but not reflected in the administrative data base
- Prescribing physician name inaccuracies were found
- There were coding issues for diabetic eye exams
- Physician practice addresses were found to be inaccurate, leading to concern about accurate group assignment for tiering
- New physicians were assigned to the non-preferred tier for unclear reasons

- Physicians felt angry and disempowered because of
 - lack of opportunity to review scores before patients see them
 - lack of opportunity to review data
 - lack of opportunity to correct data
- A difference of 0.01 units, from 1.00 to 0.99, resulted in non-preferred tiering
- Physicians were assigned to the non-preferred tier based on small sample size
- Among four specialists in the same group, two were assigned to one tier and two to another
- A nurse practitioner's work was attributed to different supervising physicians, who may be in different tiers
- Many noted an increased administrative burden from having physicians assigned to two different tiers within the same office

IV. Data Accuracy

Data accuracy is critical to accurate evaluation of physician performance. Lack of accuracy decreases the usefulness of the evaluation for all stakeholders, increases the risk of unintended consequences, and increases physician distrust of the system. Mercer and GIC have left most of the primary data accuracy work to the plans. There are several areas that are especially important for plans to examine, and there are ways Mercer can help address accuracy issues.

How could profiling data based on paid claims be inaccurate? The problem is that the claims payment process will tend to promote accuracy only in those elements necessary to pay the claim, such as procedure codes and knowing who billed for the service. But performance measurement depends on other elements, such as diagnosis codes and knowing who actually ordered or performed a service. Although the GIC has worked with plans specifically to obtain other data, such as diagnosis codes, those data are not necessarily involved in claims payment.

To understand ordering, servicing, and billing provider information, consider the following three cases. If a primary care physician in solo practice performs and bills for an electrocardiogram then all three – the ordering, servicing, and billing providers – are the same. Physicians writing a medication prescription are the ordering physician but the pharmacy is the servicing and billing provider. A physician who sends a patient for an MRI of the spine is the ordering physician, but the radiologist who performs the study is the servicing physician, and the radiology group is the billing entity.

The data elements that are not used to pay a claim may be incorrect. That includes the ordering and servicing physician, and also diagnosis (ICD-9-CM) codes. Those data must be evaluated separately from billing data. Information may be inaccurate or even missing from a data base consisting of paid claims, thereby introducing inaccuracy into the profiling data base.

Radiology and medication utilization are among the most important factors under physicians' control. Therefore their accurate measurement is particularly important. However, radiology practices and pharmacies do not necessarily have any intrinsic incentive to identify the correct ordering physician. In fact, prescriptions may list all physicians in a group, and given that pharmacists often have trouble identifying physicians from their signatures, a classic accuracy problem is their assigning the prescription to the first physician listed in the group. Similarly, radiologists may not know which individual physician in a group ordered a radiology study, since the test may have been scheduled by a mid-level practitioner or by a secretary.

Under the right circumstances, assigning the total cost of an episode to one physician may lessen the accuracy issue for the efficiency index. The total cost of an episode does not depend on who actually ordered the service, presuming the episode is attributed accurately, and the costs are truly part of the episode. Those are significant qualifications, however. Any costs used for attribution must be accurate. We recommend including

pharmacy costs in the attribution rule (Cost Efficiency Measure, recommendation 2), which would require accurate prescribing physician information. Such information must be accurate for reporting purposes, without which there will be difficulty in effecting change in physician behavior.

Physician specialty assignment is especially important to assess. An evaluation system presumes that physicians are compared on equivalent patient populations. A neurosurgeon assigned to general surgery will have a different patient population, and this will have an effect among all conditions seen by both neurosurgeons and general surgeons. (See also the discussion of conditions belonging to a given specialty, below). Physician identity codes themselves may not be accurate, and furthermore must be linked from one plan to the next, introducing another possible source of error.

Physician disease coding also can make a large difference. Encounter coding determines the condition in which cost comparisons are made. For example, malignant hypertension is considered a separate disease from the much more common (and much less costly) condition, benign hypertension. The assignment of care to one or the other depends on using ICD-9 code 401.0 for malignant hypertension and 401.1 for benign hypertension. But software systems often interpret a claim submitted with the digits “401” as “401.0” and generate an episode of care in the wrong condition. A similar situation occurs when low back pain is coded as “muscle strain” instead of lumbar disc disease. These coding inaccuracies may cause some physicians to appear artificially more or less expensive than average. While acknowledging these provider coding accuracy issues, they should be viewed in the context of providers’ concurrent need to navigate an extremely complex medical policy and claims system.

Data accuracy affects the measurement of quality performance as well. Accurate prescribing information is important in two ways: first because it is used along with encounter information to assign the primary specialist to whom patient care will be attributed, and second because it is of course scored in many measures. Radiology ordering physician information also appears in several measures, for example imaging of low back pain. Thus ordering physician accuracy affects many components of scoring in vital ways.

We have mentioned several common problems. In fact, any given physician performance evaluation system will have multiple other data accuracy issues. These cannot be totally defined and addressed a priori. Increasing data accuracy must be approached as a continuous quality improvement (CQI) project aimed at improving the evaluation system itself. Such projects require multiple cycles of defining and identifying problems, measuring them, analyzing them, and ultimately improving and controlling their accuracy. This need not require a full Six-Sigma approach because important aspects can be implemented in a straight forward manner. The key to addressing data inaccuracy is implementing a feedback cycle which accepts possible errors from multiple sources, assesses their causes, and then improves the system.

The CPI Initiative includes a Technical Advisory Committee, and recently has added a physician advisory group. These could form one part of the feedback and improvement mechanism needed. The other parts include input from physicians who have examined patient-level detail, input from other sources (such as analysts at the health plans), a system for evaluating possible errors, a way to develop corrective action, and a communication plan that informs those concerned – physicians, plans, and when necessary the public – of the actions taken.

There may be a concern that accepting feedback and changing the system will lead to a large administrative burden. For the many reasons detailed later in this report, we believe this work is not only necessary but desirable. Fortunately, the administrative burden is not large, in our experience. It involves an initial investment in setting up the feedback and change mechanisms. At the outset there will be many issues regarding data accuracy, but as the data and measurement tools improve the amount of work needed for further refinement decreases. The process must be ongoing because problems are continually uncovered, and new issues arise. An ongoing system provides accuracy monitoring and control with relatively small effort, and helps build confidence in the measurement program.

In the spirit of this CQI approach to data accuracy, FMA requested claims level detail for approximately twenty physicians. These physicians were MMS members who wanted to better understand their results on the cost measure. Our intention in examining these physicians' data was not to generalize from that data to the whole program, nor to examine a statistically valid sample. Rather, it was to perform a targeted investigation to identify problems that would not otherwise be apparent. This type of analysis is useful for finding additional methodological issues that would not be apparent a priori, and this type of investigation often identifies additional ways to improve the system, which is our goal. As of the date of this report, FMA continues to work through the administrative requirements for obtaining access to the data that produced those results.

Recommendations: Physicians should be given patient-level drilldowns for the efficiency measure, and patient lists for the quality measures, to help identify systematic accuracy problems. Patient-level detail will generate accuracy issues as physicians examine their own data. The health plans and Mercer should develop a formal feedback and correction mechanism so that errors uncovered by physicians, plans, and other analysts can contribute to improving the evaluation system. Plans should initiate programs to audit their primary data, especially in the specific important areas discussed above. Mercer and the plans should examine low and high outliers in detail as this targeted evaluation is likely to identify systematic accuracy issues. Mercer should review its data base for unusual disease coding patterns.

Using the data for individual tiering exacerbates the impact of accuracy issues for the reasons noted above. Pharmacy prescribing and radiology ordering information, for example, will be more accurate within a group than for an individual. Therefore we recommend tiering at a group level until data accuracy is improved and validated.

V. Cost-Efficiency Measure

Construction of physician level cost efficiency measures from health care claims data involves the following five steps:

1. Claims are processed through “episode grouper” software, which aggregates each member’s claim records into “episodes of care,” where an episode of care refers to a period during which a disease process is present and is being managed – diagnosed and treated – by health care providers.
2. An actual cost figure is calculated for each defined episode by summing costs of all claims included in the episode, including those for physician services, inpatient and outpatient facility services, prescription medications, and other services.
3. Responsibility for each episode’s actual and expected costs is attributed to a physician or physician group based on an attribution rule.
4. An episode expected cost is calculated for each defined episode.
5. Sums of actual costs and of expected costs are calculated for each physician or physician group based upon his or her attributed episodes, and physician- or physician group-level cost efficiency measure are calculated. Physicians are compared, within specialty, on relative cost efficiency performance.

In constructing its physician cost-efficiency profiles, Mercer processes claims records through Symmetry’s Episode Treatment Group® (ETG®) software, which is the most widely used episode grouper. For episode costs, Mercer standardizes claim costs for professional services using CPT codes, for facility outpatient services using HCPCS codes, for outpatient surgery using ambulatory surgery center rates, for facility inpatient services using per-diem rates for each of seven different bed types, and for prescription medications using NDC code and amount dispensed. After individual episode costs are summed, cost outlier episodes – those that exceed Symmetry’s ETG-specific trim points – are removed from the analysis. Then, responsibility for each defined episode is attributed to the physician who accounts for the greatest percentage of professional costs (excluding costs of anesthesiologists, radiologists, emergency and other non-managing physicians), provided that the percentage is at least 25%. Episode expected costs are calculated as specialty specific, ETG averages – i.e., as the average costs of episodes within each ETG managed by physicians in a specified specialty. Finally, Mercer sums actual and expected costs across all attributed episodes, and calculates the physician ratio-of-actual (observed)-to-expected costs (O/E Ratio). An O/E ratio is often also referred to as an efficiency index (EI). Ratios above 1.0 are considered to indicate relatively cost-inefficient performance, those below 1.0, relatively cost-efficient performance.

The greatest threat to accuracy of cost-efficiency scores is unreliability associated with basing performance calculations on small samples of episodes. Mercer attempts to deal with this issue in three ways. First, Mercer has constructed a consolidated database that includes claims from all health plans serving GIC members. Cost efficiency scores can be

based on episodes from multiple plans, and thus profiles are more reliable – and more accurate – when based on larger samples of episodes. Second, Mercer utilizes three years of claims from each of the participating health plans. While use of multiples years of claims increases number of episodes available for profiling, and thus increases reliability, it may simultaneously reduce the relevance of profiles since they reflect non-current performance. Furthermore, physicians may be penalized for behaviors that have already changed. To try to adjust for this problem, Mercer for its 2006 profile release – which is based on claims data from 2003, 2004, and 2005 – weighted recent episodes more heavily than earlier episodes. Episodes from 2005 were given a weight of 3.0; episodes from 2004, a weight of 2.0; and episodes from 2003 were weighted 1.0. The third step that Mercer takes to try to assure profile reliability is to require a minimum of 30 complete non-outlier episodes for any profile. As a consequence, practices are not profiled when they account for 29 or fewer attributed episodes in the three-year, multiple plan database.

After calculations are completed, Mercer provides each of the participating health plans with three levels of result reporting. First, for each physician profiled, health plans receive a composite O/E Ratio based upon data from all participating plans. Second, health plans receive physician O/E Ratios based only on their own data. Third and most detailed, the plans receive physician specific reports that show, by ETG, a physician's average cost compared to his/her specialty peers for inpatient care, surgery, medical management other than surgery, outpatient services, and prescribed medications. For 2006, Mercer is also providing health plans with copies of the ETG-processed combined claims database so that the plans can perform their own analyses with the multi-plan data

While Mercer's profiling approach incorporates steps designed to increase accuracy, we believe that modifications can be made in order to further enhance accuracy of results. Specific concerns and the changes that we recommend are as follows:

- 1. Definition of cost outlier episodes.** Mercer currently utilizes Symmetry's ETG specific trim points for identifying cost outlier episodes. These trim points are recalculated periodically by Symmetry based on its own national claims database. But the cost distributions of ETGs in Symmetry's database may not be (actually, are highly unlikely to be) the same as those in Mercer's Massachusetts database. As a result, percentages of episodes defined as outliers and removed from analysis will vary from ETG to ETG, and these variations can lead to biased cost-efficiency scores. Also, Mercer uses a procedure of removing cost outlier episodes from the analysis, i.e., "trimming" cost outliers. Thomas (2006) has shown that truncating (or "Winzorizing") outlier costs to a specific percentile of the ETG cost distribution works as well as trimming outliers from the analysis in terms of both measurement reliability and measurement bias.¹⁵ Further, because truncating retains episodes in the cost efficiency analyses, more physicians satisfy minimum episode volume requirements and larger percentages of physicians can be profiled.

Recommendation: Establish ETG specific trim points (e.g., at the 2% and 98% levels of the ETG cost distribution) using current Massachusetts claims data, and truncate outlier cost instead of trimming outlier episodes from analysis.

2. **Episode attribution.** Mercer attributes episode responsibility to the physician accounting for the highest portion of professional cost in the episode, so long as the physician's portion equals at least 25%. Ideally, episode responsibility should reflect all costs – professional, inpatient, outpatient, and pharmacy. However, health plan inpatient facility claims provide only service level aggregate costs (e.g., total room and board costs, total laboratory costs), and as a consequence it is not possible to link physicians with the costs of services for which they may be responsible. For facility outpatient services such as laboratory and imaging tests, the HCFA 1500 billing form does provide space for identifying ordering physician. However, this information is usually missing on completed forms and not included in health plan databases, and so, again, it is normally not possible to link physicians with the costs of services ordered. However, for pharmacy claims, prescribing physician often is identified in health plan databases.

Recommendation: Base episode attribution on combined professional and prescribing costs rather than on professional costs alone. This will be possible only if prescribing physician is reliably identified in pharmacy claims across all participating health plans. If prescribing physician is currently not reliably identified, GIC should encourage health plans to undertake steps to improve source data quality (see Section IV above).

3. **Risk-Adjustment of Episode Expected Costs.** Like nearly all users of ETG software, Mercer defines episode expected costs based on ETG and specialty averages. In the new release of its Episode Resource Group (ERG) software, Symmetry now provides the capability for recognizing the potential influence of ETG specific co-morbid conditions on episode costs. With this new system, each episode defined during ETG processing is assigned a risk level to reflect the treatment cost impact of co-morbid diseases present.

Recommendation: Although this software release is too new to allow us to estimate the magnitude of risk adjustment effects, Mercer should implement the new system and incorporate risk adjustment of episode expected costs into its profiling process.

4. **Number of Years of Data and Number of Episodes Analyzed.** Mercer uses three years of claims from each of the participating health plans when constructing physician cost efficiency scores. As noted above, use of three years of data, instead of two years or one year, increases the number of physicians who satisfy the minimum-number-of-episodes criterion for profiling. However, as also noted, use of data that are three years old reduces the relevance of profile scores

as indicators of current performance. Although Mercer has attempted in 2006 to address the age-of-data problem by weighting more recent data more heavily than older data, we believe that the problem would be better addressed by limiting profiling analyses to two years of claims data. With two years of claims, fewer physicians would satisfy the 30 episode minimum requirement, but profiles would be more descriptive of current performance and use of the subjective weighting scheme would not be necessary. It should be noted that, to our knowledge, there are no methodologically strong studies that support 30, or any other specific number, as the appropriate minimum sample size. Implementation of recommendation 1 above – truncation of outlier cost instead of trimming outlier episodes from analysis – would ameliorate this problem to a degree, and scoring physicians in groups instead of individually would lessen the problem significantly.

Recommendation: Use the most recent two years of claims data to assess performance, and do not use a weighting scheme; truncate instead of trim outlier episodes; score physicians in groups instead of individually in order to increase sample size and thereby increase reliability.

- 5. Definition of Physician Specialty.** Mercer defines a physician's specialty based on credentialing records of the majority of participating health plans. Thus, if a physician is listed as an internist by three plans and a pulmonologist by two others (the physician may have a primary specialty of internal medicine and a secondary specialty of pulmonology), s/he would be evaluated only as an internist. All episodes attributed to the physician would be included in the physician's profile, including, in this example, internist related episodes, pulmonologist related episodes, and the small number of additional episodes that are attributed to the physician but relate to neither specialty. This approach is problematic for at least two reasons. First, in terms of pulmonology performance, the physician is compared only to other internists who manage pulmonology episodes. For rarely occurring combinations of primary and secondary specialty, virtually all episodes in a sub-specialty category might be associated with a single physician, causing the specialty specific ETG expected costs to be based on that physician's own performance. Second, inclusion of rarely occurring (for a specific specialty) episode types (ETGs) in physicians' profiles increases profile unreliability. Both of these problems would be eliminated if Mercer were to define specialties in terms of types of ETGs managed; i.e., general internists manage ETG 0281, benign hypertension; 0678, minor inflammation of skin & subcutaneous tissue; 0333, acute sinusitis; etc. Cardiologists manage ETG 0265, ischemic heart disease, except CHF, w/o AMI; 0281, benign hypertension, w/o comorbidity; 0311, cardiovascular disease signs & symptoms; etc. The particular set of ETGs that define a specialty could be identified as those accounting for 90%, 95%, or some other selected percentage of episodes managed by physicians whose credentialing records identify them with that specialty.

Recommendation: Define specialties in terms of the types of ETGs managed, and use only the particular ETGs that define a specialty to calculate the efficiency index.

- 6. The efficiency index calculation includes the cost of preventive measures and some chronic disease care.** Primary care physicians are scored on the ETG that includes well child visits and the ETG that includes vaccinations. Diabetes and asthma care ETGs may be included. Our concern is that in the IOM paradigm, preventive health measures are underuse measures. Similarly, many of the important elements of chronic disease care, such as HbA1c testing and asthma controller medication use, are also under utilized. In order to improve quality, the number or rate of these interventions should be increased. Therefore, it is desirable to spend more on these measures. However, costs associated with these interventions increase the observed to expected cost ratio. The current scoring system has the potential to reward those with lower screening or preventive measure performance and to discourage increasing the use of those interventions.

In a similar fashion, some preventive measures such as routine physicals or screening colonoscopy may appear in other ETGs that are not oriented specifically to preventive care. For example, if a physician codes sinusitis at a well child visit, as well as the diagnosis code for a general exam, the cost of the well child visit could appear in a sinusitis episode. Because well-child visits are both preventive and much more expensive than a routine office visit, the practitioner is doubly penalized (first by the cost of a preventive, underuse intervention, and second by having a relatively expensive episode of sinusitis). In other words, a physician who is saving the system money by handling the acute problem at a well-child visit actually can appear less efficient than average when that visit appears in the acute problem's ETG.

It has been observed many times that there is no correlation between physician performance on cost-efficiency and quality (underuse) measures. It would seem therefore that physicians have not been either hurt or rewarded by including preventive measure costs in the EI. However, this is only true when considered *retrospectively*. If one of the goals of the system is quality improvement *in the future* then the incentive to be in a preferred tier must be aligned with the incentive to provide underused interventions.

From the physicians' and patients' points of view, removing underuse measures avoids misaligning cost-efficiency measure with underuse quality measures. Avoiding distortion of costs when physicals are included in acute care episodes improves accuracy and face validity. For GIC and its participating plans, removing these costs would help promote underuse quality work and improve plan measures of quality such as HEDIS scores.

There are a number of strategies for addressing this issue. For example, ETGs primarily composed of preventive measures such as vaccinations could simply be

removed from EI calculation. Claims representing costs of preventive measures or underuse interventions in chronic disease could be removed from the data base. Costs of such claims could be zeroed out. Mercer should analyze such strategies to find the ones most appropriate to given conditions or ETGs.

Recommendations: Remove the costs of preventive measures and underuse interventions from the EI calculation. When a routine physical appears in a non-preventive care ETG convert its cost to that of an average office visit (e.g. a 99213 cost).

7. **Cost-Efficiency Metric.** Mercer uses a ratio of observed to expected costs as the metric for assessing physicians' cost-efficiency performance. This metric has the advantage of being widely used and intuitive. However, it also has a substantial disadvantage. With the O/E Ratio metric, physicians whose available samples of episodes are small compared to those of their specialty colleagues are at significantly greater risk of being misclassified as either cost inefficient or cost efficient, because the sampling variance of cost estimates is larger when based on fewer episodes.¹⁶ As an alternative to the O/E Ratio, Thomas et al. proposed the Standardized Cost Difference (SCD) metric.¹⁶ The SCD is in essence a Z score statistic with difference between actual and expected costs as the numerator. The MedStat Group also uses a Z statistic as its cost efficiency metric, but MedStat's measure is based on the O/E Ratio; i.e., it is the O/E Ratio adjusted for episode sample size differences across physicians. Both SCD and MedStat's O/E Z statistic control for episode sample size differences across physicians.

Recommendation: Use the Standardized Cost Difference (SCD) metric instead of a ratio of observed to expected costs.

8. **Control for Effects of Benefit Differences across Health Plans.** The issue of small sample sizes has long plagued health plan efforts at episode-based economic profiling of physicians. In a single health plan's database, large percentages of network physicians often fail to satisfy minimum episode volume requirements for profiling. Further, especially for physicians represented by relatively small samples of episodes in each of several health plan data sets, different health plans can reach very different conclusions about cost efficiency performance from their analyses of the small samples of episodes. To address this problem, GIC and Mercer three years ago devised the approach of combining claims from multiple health plans and from both HMO and PPO populations into a single consolidated profiling database. As a result, larger percentages of network physicians satisfy minimum episode sample size requirements and thus are eligible for profiling, and physician performance measures, which are based on larger samples of episodes, represent more precise estimates of physicians' true cost efficiency performance. Although the consolidated database approach is now being adopted in other geographic areas, and indeed is currently being promoted in an Agency for Health Care Research and Quality (AHRQ) demonstration program, an important methodological question remains to be

answered. The question is: do benefit differences among participating health plans lead to measurement bias when physician performance scores are calculated using a consolidated database? If one health plan provides more generous coverage for specific services than another health plan, members of the first health plan might utilize those services more frequently than members of the second. As a result, a physician whose practice includes a greater percentage of patients from the first health plan might appear to be cost inefficient compared to another physician whose practice includes fewer of those patients, not because of the physician's own decisions, but instead because of the proclivity of the first health plan's members to seek services. Data shown in slides from Mercer's February 18, 2005 (slide 21) and June 29, 2005 (slide 15) presentations on the Clinical Performance Improvement Initiative suggest that this may be a problem with the CPI Initiative consolidated data set. In slide 21 of the February 18 presentation, data are shown for a family practice physician who has a provider efficiency index of 0.64 for Harvard Pilgrim patients and 1.15 for Tufts Health Plan patients. There are several possible explanations for such differences, but until benefit differences across plans is ruled out as an explanation, GIC and Mercer must be concerned about significant bias in individual physician scores.

Recommendation: The effect of benefit design should be assessed and factored out of the physician cost efficiency metric.

VI. Quality Measurement

Quality measures based on administrative data have several factors in common:

1. A rule that describes the desired intervention. An example of such a rule is “Every patient with diabetes should have a hemoglobin A1c measurement once a year.” Ideally these are evidence based, but often they are based on consensus (expert opinion).
2. A population of interest. The patients all share a common characteristic: they have diabetes, or they are female and between ages of 50 and 69. These patients form the denominator of quality measure rates.
3. Interventions of interest. Counting the interventions forms a numerator for a quality measure.
4. Connecting the population specific to the rule (the denominator) to a given physician or group (attribution).
5. Crediting the recommended intervention to a given physician or group, another attribution issue.

The quality indicator developed by Mercer implements these factors through a combination of measures developed by RHI and logic developed by RHI and Mercer. The CPI Initiative used an internal consensus process to select a 59 measure subset of all of RHI’s measures for 2005. An additional 17 measures were selected for use in September 2006.¹⁷ Measures were chosen based on the ability to apply them to administrative data. For example, the consensus process resulted in the exclusion of measures that could not be applied reliably using claims data alone and would have required chart review (e.g. a measure that looked at aspirin use, since aspirin would likely be purchased without a prescription, or a measure that considered patient symptoms). In addition, potential quality measures that included undefined terms (e.g. “pre-operatively”) were not employed.

Mercer and RHI chose a number of safeguards at each step of the quality measurement process they performed. The measures were selected from a number of nationally recognized sources, such as AQA, HEDIS, and AHRQ. Measures that might be influenced by benefit design were limited to patients with that benefit. For example, appropriate asthma medication use was only measured among patients with a pharmacy benefit.* To identify the population of interest, the measures required more than a single marker of the disease or condition of interest. Numerators (interventions) had to be available in claims data. In order to increase sample size, Mercer rolled up all denominators and numerators to form a single ratio of interventions accomplished divided by opportunities among the population, across all diseases applicable to that physician’s specialty.

The process begins with the same three-year claims data set used to generate the efficiency index. Mercer passes the data to RHI, along with a Master ID and Specialty

* For similar reasons the efficiency index was calculated only on patients with a pharmacy benefit.

for each physician as determined by Mercer. Denominator populations are identified from the first two years of data. Interventions are assessed for those patients from data in year three. RHI makes sure that patients in the denominator have an opportunity to be in the numerator of interest by checking for continuous enrollment during the measurement period. RHI then generates the numerators and denominators for each measure that are attributed to a given provider, and passes those back to Mercer, which in turn gives both plan specific and aggregated numbers to each plan.

The attribution rules that RHI chose are reasonable.* To link a denominator patient to a physician, RHI found the physician who provided the most outpatient services (encounters and prescriptions) to that patient in the last 18 months of the data set. If the measure applied to more than one specialty, the process was repeated for each specialty. In this way a “primary specialist” was determined for each specialty and its applicable measures. The interventions, on the other hand, were attributed to each member. In that way, a primary specialist would receive credit for an intervention no matter who was responsible for getting it accomplished. Thus if a patient with coronary artery disease received a lipid profile, both the primary internist for that patient as well as the primary cardiologist would receive credit, whether the internist, the cardiologist, or even a consulting endocrinologist actually ordered the lipid profile. This is an appropriate, patient-centered methodology that is used nationally.

In terms of validating the measures, Mercer and RHI have been generally transparent about the construction of the measures. Descriptions of the initial 59 measure set were reviewed for this report.¹⁸ RHI has made detailed specifications, including ICD-9 and CPT codes, available to physicians through participating plans.

Mercer and RHI have made many reasonable choices in developing the quality indicator. Nonetheless there are a number of concerns with the current quality measurement system. A number of revisions would make the quality indicator more robust, more useful for quality improvement, and more likely to accomplish the goals of the GIC:

- 1. Lack of patient lists and interventions credited for quality measures.** Giving physicians the list(s) of eligible patients with information on completed interventions would help address several problems at the same time. First, such lists would yield more accurate data. Lists can be inaccurate in any of the three characteristics listed above. Patients may be inappropriately assigned a condition (denominator issue). Interventions may not be captured (numerator issue). Because encounter coding and prescribing physician data may be inaccurate, physicians may not in fact be responsible for their care (attribution issue). As in the discussion of efficiency, accurate data benefits everyone. Similarly if there is a mechanism for receiving corrections, the underlying data base becomes more accurate. Although information regarding which patients are assigned to the denominator and numerator of each measure for each physician is available and

* Note that the attribution rule used by RHI for quality performance measurement is different from the rule used by Mercer for episode attribution in cost efficiency analysis.

has been supplied by RHI, health plans have not, to date, provided this information to physicians.

Second, the list becomes a registry that can be used for true quality improvement. Physician participants would not have to purchase or use electronic health records because the data has already been aggregated under the CPI Initiative. Good quality measurement should promote improvement. It would make more sense to focus the intervention on improving each physician's practice process so that all patients receive desired care.

Lastly, providing action items as well as opportunities to correct the data empowers the physicians to act and would go a long way towards defusing physician frustration and anger while promoting improved care.

Physicians need to be familiar with the detailed specifications of each measure in order to understand why patients were assigned to it, and which interventions count. RHI has made detailed specifications available to physicians on request from participating plans. Plans should systematically review the specifications with their physicians so that physicians understand what is expected of them and what interventions their patients need.

Recommendation: The health plans should provide physicians a list of patients counted towards each quality measure, as well as the interventions for each patient that counted towards that measure, especially since RHI has provided this information. Plans should systematically review the details of the specifications with their physicians so that physicians understand which patients are affected and what interventions are expected.

2. **Trying to overcome the problem of small sample size by rolling up multiple measures' numerators and denominators.** This methodology assumes that the measurements are independent. However, there are at least three known ways in which the measures are not independent of one another. Types of patients tend to cluster around a given physician. Some physicians like to take care of patients with certain conditions, such as diabetes. Some physicians collect emotionally needy patients. Greenfield and colleagues have described concerns about clustering of patients and the influence that has on quality assessment accuracy.¹⁹ Conversely, a single patient can have a cluster of measures. One diabetic contributes multiple different interventions (HbA1c, lipid measurement, retinal exam, etc.). Therefore one non-compliant patient may affect multiple measures. Furthermore, a number of the measures are low frequency occurrences that increase the potential for patient behavior rather than physician quality performance to influence outcome. This "lumpiness" makes it risky to add up a large number of different conditions or diseases where each is applying to only a small number of patients. Lastly, one intervention fulfills multiple measures. For example, under care of hypertensives there are five measures that a given patient can accomplish with a single blood draw. While we applaud the attempt to

address the problem of small sample size, the uncertain independence of the numerators and denominators is a significant statistical issue that has not been addressed.

There are several ways to ameliorate this problem. The first is to select a smaller number of clinically important conditions with significant population impact, and then score those in separate domains. Unfortunately, an individual physician may still have too small a sample size for statistical significance. For example, to reliably evaluate a physician's care of diabetes might require as many as 100 patients.²⁰ Measuring physician groups instead of individual physicians would resolve the problem to a large extent by directly increasing sample size.

The difficulties related to clustering can be addressed through statistical methods that analyze the data for degree of independence. Adjustments can be developed that overcome these problems, and allow an adjusted roll up methodology to be used to increase sample size for individual physician measurement. Until then, however, it would be more appropriate to use these measures at a group level, where the sample size can be addressed by aggregating data across a number of individual physicians.

Recommendation: Select a smaller number of measures that address conditions that are both prevalent and clinically significant. Require an appropriate minimum sample size before counting a measure towards the overall quality performance evaluation. Profile as groups instead of individual physicians in order to overcome the issue of sample size, at least until the numerator-denominator roll-up methodology has been modified and validated to take into account the fact that its components are not independent measures.

3. **The quality measure has an implied target of 100% adherence.** The overall quality indicator is a “rolled up” value. There are no specific targets for given conditions. Furthermore, when plans convert adherence into performance relative to specialty average, quality is only measured relative to peers. Therefore the target is open-ended: even if average adherence to a given measure were to be acceptably high, individual physicians would need to perform better still. Both these factors create an implied target of 100%.

As physicians push past 80-90%, they find themselves attempting to influence patients who may prefer not to receive the appropriate test or treatment. One of the core principles of quality care is respect for a patient's autonomy. By making the implicit target for a measure 100% we are saying that for each decision, we know what is best for each patient. That then becomes a factor related to accepting screening measures, preventive services, or monitoring treatment. Dr. Albert Mulley of Boston has been active nationally in promoting the incorporation of patient informed choice into these measure calculations.²¹

Our intent here is not to downplay the importance of striving for excellence. The problem is that in reaching for perfection the system runs the risk of devaluing patient preference, interfering with the doctor-patient relationship, and disrespecting the “art of medicine” in the sense that care must include some clinically appropriate variation. We also note that use of performance relative to specialty average additionally does not address the situation in which performance is generally poor.

If we are to define quality as incorporating patient choice into measurement we have two choices. The first is to create exclusion criteria for the measure denominators such as is being done in the UK Pay for Performance program.^{22,23} However, the method by which patients can or are excluded must be articulated clearly. This is different from allowing exclusions for counseling patients regarding desired behaviors. We would not recommend such exclusions because they ignore the variation in physician counseling and interviewing skills.

Another way to respect patient autonomy is to set a reasonable target for the measure so that physicians are not held accountable for patients that have a legitimate reason for non-adherence. As an example, one might use the 90th percentile nationally for a target.

Without appropriate targets for specific conditions, the present quality assessment process runs the risk of ignoring the patient’s role in adherence and preferences in defining quality. In evaluating the “quality” of a physician’s care, the determinants of quality may be different for different patients. The measures provided are often ones that an informed patient could request without seeing a “quality” doctor. For example, many women know they should have a mammogram or a colonoscopy. Patients with coronary disease have heard of aspirin and statins. In these instances accomplishing recommended health interventions is a combination of practitioner organization and knowledge and patient involvement in their care.

Setting a target of less than 100% begs the question, what should be done for that last 10 or 15% of patients? As described in Ed Wagner’s chronic disease model, quality care is collaboration between an engaged patient, an activated informed physician, and a community prepared to support the doctor-patient dyad by providing access to needed, quality services.^{24,25} Therefore we would suggest that some of the responsibility lies with the health plans themselves. Are they giving patients an incentive to do the right thing? Are plans intervening with individuals who are not responding to a physician’s recommendations? For example, if one socio-demographic group has lower adherence to screening procedures, rather than changing the tier for the physicians in that community (which might impair access and increase costs to some who could afford them least), the plan and the community could explore the reasons for lower adherence and address them.

Recommendation: In addition to choosing a smaller number of clinically important measures, plans in conjunction with physician leadership should determine appropriate targets for each measure. Plans should provide support for intervening with patients who do not adhere to interventions despite their physicians' best efforts. The goal should be system change to attain a target level, not merely "better than average" quality.

- 4. Case-mix adjustment when combining quality measures.** The specialty average for quality measures will vary from measure to measure. Some of that variation will be due to intrinsic differences in the difficulty of accomplishing the recommended intervention. A diabetic eye exam requires a two-hour trip to a specialist and a copayment. That makes it more difficult than, say, obtaining a blood test for warfarin monitoring on a walk-in basis. The appropriate targets set for each individual measure would in part reflect these differences. If multiple quality measures are combined there would need to be case-mix adjustment for this effect. Otherwise a physician could be penalized simply for treating patients with a certain mix of diseases.

Recommendation: Apply case-mix adjustment if multiple quality measures are combined into a single score.

- 5. Attribution is not adjusted for quality measures that reflect overuse rather than underuse.** As discussed above, when a diabetic patient has an intervention accomplished (i.e. underuse is addressed), it is appropriately credited to the physician most involved in care regardless of who actually provided the intervention. In overuse use situations, the overused intervention could be inappropriately attributed to the physician who provided most of the care but not the particular overuse. If a patient with a viral respiratory infection receives antibiotics from an on-call or emergency room physician (i.e. overuse occurs), the blame for the overuse accrues to the patient's usual physician. This may explain why one pediatrician noted many antibiotic prescriptions for viral URI in his data even though he strictly does not prescribe antibiotics in that situation.

Recommendation: For quality measures that involve overuse, the instance of overuse should attach to the attributed physician only when that physician actually generated the overuse. (Individual measures are identified as underuse or overuse in the Appendix.)

- 6. There is no weighting of the measures, so all are considered equal.** The scoring system treats all measures as equally important. That is not the case: some measures are clearly connected to improved outcomes (for example, use of controller medications in asthmatics) while others are only expert opinion (for example, specific follow up frequencies). Because a subset are evidence based

measures with demonstrated benefit in terms of survival or reducing morbidity, it would seem appropriate to give those measures more weight.

Choosing a smaller number of clinically significant measures would enable weighting, first by the choice itself, and then by assigning variable weights to different measures as appropriate. The weighting could be determined through a consensus process involving the physician community, thereby helping achieve physician buy-in to the final product. Unfortunately there is little objective evidence by which to set precise weights. Left alone, the current system's default is an implicit decision in favor of equal weighting. We feel it is preferable to have an explicit discussion and reach a consensus on this issue.

Recommendation: Physicians working with CPI Initiative sponsors should investigate, propose, and discuss weighting of quality measures through a process that is transparent. Strong consideration should be given to assigning higher weight to quality measures that are evidence-based and most clearly connected to improved outcomes.

VII. Issues Related to Tier Assignment

The preceding two sections have addressed issues related to generating cost efficiency and quality measures. This section discusses the issues raised by the use of those scores to assign tiers to physicians or physician groups, and the process of implementing the tiering system. There are several opportunities for improving the tier assignment process. We do not comment on the utility or appropriateness of tiering per se as a method of promoting cost efficiency or quality improvement; that question is outside the scope of this report.

To this end we note the following concerns and make the following recommendations:

- 1. Lack of a uniform tier assignment protocol.** Details of generating the tiers for the most part appear to have been left to the plans themselves. As of 2006, only one plan tiered at an individual level, with the rest using group assignments. There are differing cut-off points for preferred and non-preferred tiers. There is variation in the precedence of efficiency versus quality score in generating tiers. Without further discussion there will be variation in the methodology for rolling up individual scores into a group score. The result is that some physicians are in one tier for one plan, and another tier for another. For a plan tiering at the individual level, that individual's tier may differ from his group-scored tier, resulting in one or another physician in a given practice being in more than one tier. When a single physician or group is assigned to different tiers by different plans, confusion and doubt have resulted. The confusion arises for physicians, patients, office staff, and administrators. Doubts are cast on the overall system when different tiers are assigned to the same physician or group.

The tiering system should be transparent and predictable. A doctor assigned two different tiers for the same pool of (GIC) patients becomes suspicious of accuracy of both designations. Uniform tiering helps reduce the administrative burden for offices. It would generate improved face validity and physician acceptance. The impact of the CPI Initiative will be strengthened if GIC members seen by a physician group are at a preferred or non-preferred tier.

The GIC, Mercer, and plans should involve a broad range of stake holders in developing an outline for a suggested tiering protocol for use across all plans. The discussion should include the basic issue of individual vs. group profiling. By bringing the physician community in on the process from as early a point as possible, listening to their comments, and using their ideas to improve the program, the CPI Initiative would gain trust and respect.

Recommendation: The GIC and Mercer, in conjunction with a wide array of physician leaders, should develop a suggested tier assignment protocol that all plans could use.

- 2. Very small absolute performance differences result in a yes-no decision on assignment to the preferred tier.** Multiple MMS members have noted that they or their colleagues were assigned to the non-preferred tier by only 0.01 or 0.02. There have been insufficient data accuracy audit and statistical tests to say that these small differences are not the results of data errors or chance. If they are, the system becomes unreliable and unstable over time. Such differences could easily be due to data inaccuracy, ETG selection, or other methodological differences. For the pediatrician described in the introduction, two patients' costs changed his efficiency index by over 0.90. Even with perfectly accurate data, random effects over time decrease measurement precision. Instability of tiering assignment over time, if unrelated to physician practice pattern changes, decreases the utility of the tiering system for the GIC and increases the possibility of harm to GIC patients due to loss of continuity of care.

To overcome these effects, the reproducibility of tiering methodology should be validated so that there is a reasonable probability that tier assignments are not due to chance. This will mean that there are some groups with too small a sample size to determine their tier reliably. That applies even more strongly to tiering at an individual level. Physicians or groups with a statistically too small sample size are by definition not necessarily any worse performers than others, and it would be misleading to assign them to the non-preferred tier, as apparently has occurred in some plans.

Furthermore, a data driven process should be applied to those who are close to average to determine and then indicate that they are not appreciably different from average. One technique that could be applied, for example, would be finding natural breaks that help differentiate those clustered around the mean from others. More formal techniques could be investigated that would generate confidence intervals. Then physicians whose confidence interval overlaps the mean would be deemed average. Note that any of these techniques would create different percentages of above, below, and average performers for one specialty versus another.

Therefore in addition to tier assignment there should be categories such as “not assignable” and “not different from average” as appropriate to the corresponding situation. The result will be improved reliability, stability, and respect for the system.

Recommendation: Validate tiering statistically so that there is a reasonable probability that tier assignments are not due to chance. Calculate and make available figures on the reliability of tier assignments. Create categories of “tier not assignable” for physicians for those who cannot be evaluated in a statistically reliable manner and “not different from average” where that is the case.

3. **Tiering of physicians within a practice.** Tiering at an individual physician level will result in physicians within a single practice location being assigned to different tiers. This has already been seen for the plan now using individual physician tiers. While there is some evidence that individual accountability results in greater change than group measurement and reporting, all parties involved must recognize the costs and problems that could result from physicians in the same practice being assigned to different tiers. Cross coverage becomes a problem: differing tiers among physicians creates confusion for patients and bureaucratic problems for staff because of the difference in co-pays. There is another layer of complication for the nurse practitioner who works with a different physician from time to time, and therefore who apparently changes co-pay from day to day. The result is increased administrative burdens to practices.

These costs and problems should be balanced with the benefits expected to accrue from meeting the GIC's goals. Methodological and data accuracy issues decrease the benefit at this time. Until accurate tier assignment is more probable, these concerns form another reason to begin with group-level tiering instead of individual physician profiling.

Tiering at the group level leads to the question of group assignment. There is both the accuracy of the initial assignment, and the issue of physicians who change practice locations or groups. Do physicians bring their old tier with them or adopt that of the new group? A policy should be established to handle this issue. It should be based on core principles and data-driven where possible. Plans need to have mechanisms to review and re-assign groups on a yearly basis, in time for the next round of tiering assignment.

Recommendation: The increased administrative burden of having physicians in the same office assigned to different tiers offsets the potential system savings, especially until accurate physician tiering is validated, and therefore tiering should be assigned at a group level rather than by individual physician at this time. In any case, the CPI Initiative should take into account the added costs and issues that would occur if physicians in the same office were assigned to different tiers.

4. **Certain specialties by their nature are less amenable to profiling on a finer level than the group.** Cardiology is perhaps the clearest example. Cardiology groups often have a mixture of generalists and procedural-oriented practitioners. Interventional cardiologists will tend to look more expensive than general cardiologists across a variety of conditions, while cardiologists whose main work is echocardiography may appear to be more cost-efficient. The appropriate unit of analysis may thus be the cardiology group rather than individual cardiologists. Orthopedic surgery is another area in which groups with mixes of sub-specialists may be the appropriate analytic unit. True group practice may occur in some specialties, such as pediatrics, where physicians may not have their own panel of

patients, but rather see patients for the group. These questions can be investigated by analysis of the actual profiling data.

Recommendation: Because of the nature of their specialty care or of a practice's organization, certain specialties or practices may not be amenable to profiling on a finer level than a group. Irrespective of other issues around the question of individual versus group tier assignment, in these cases assign tiers at the group level.

- 5. Potential for unintended consequences.** In response to any evaluation system, physicians have an incentive (whether conscious or not) to avoid or discharge patients with unusual costs, diseases, severity, or compliance issues. Tiering at an individual physician level increases these pressures. Leaving aside ethical issues, such actions would create problems with access and with continuity of care. The evaluation and tiering system can be constructed so that such incentives are minimized. Examples are removing rare, high cost patients from scoring (the hemophiliacs in our first example; see also cost-efficiency measure recommendation 5). Another element is the use of reasonable targets, so that physicians are not under pressure to dismiss the non-compliant patient who keeps them from reaching 100% adherence (as discussed in quality measurement recommendation 3). Risk adjustment also decreases the pressure to deselect complex patients (cost-efficiency measure recommendation 3 and quality measurement recommendation 4). Plans in addition should monitor network performance such as the rate of patients changing physicians, which could indicate inappropriate patient discharges. Plans should also work with non-compliant patients using mechanisms such as case management, and use aligned physician incentives, such as rewarding participation in case management.

Recommendations: Include in the tiering system elements that decrease the incentive to avoid caring for unusual, complex, or difficult patients. Tiering by group instead of by individual is one such element. Develop indicators to monitor for unintended consequences. Mechanisms outside the tiering system should support desired consequences.

- 6. Handling of physicians who are different from peers or new to practice.** Efficiency index calculations presume that each practitioner has a similar patient severity distribution. A given practitioner must be comparable to others in their specialty. Even with adoption of our recommendation on severity adjustment (cost-efficiency measure recommendation 3), this assumption can be violated in several ways:

- A physician may be dual certified. For example, an internist who is also certified as an endocrinologist might have systematically more severe

diabetics or hypertensives. The specialty assignment logic may assign this physician to internal medicine.

- A physician may have developed a cluster of unusual patients. For example, in an actual case from the authors' (FMA) experience, a neurologist who did not mind caring for end-stage ventilator-dependent ALS or multiple-sclerosis became known within the provider community as the "last resort" for those patients. He had much higher costs in these conditions when compared to other neurologists.
- Physicians, especially specialists, with hospital-based practices could be expected to have higher costs due to a systematically different patient population (hospitalized patients and their follow up instead of a more purely outpatient population).
- Sub-specialists, such as a dermatologist who specializes in skin cancer surgery (Mohs technique) will have a systematically higher severity population than other dermatologists. Academic medical centers are especially sensitive to this issue for obvious reasons. Risk adjustment by co-morbidity is not sufficient to correct this problem, because there may be unadjusted risk within the condition itself.
- New physicians tend to have higher costs because new patient office visits are more expensive than established patient visits. These physicians also may see a disproportionate number of patients who are "doctor shopping" and therefore form a different and potentially higher cost population.

These physicians are high cost because of appropriate care for different groups of patients, not because of inappropriate use (overuse) of medical resources. Often these physicians are performing important services for the patients, the plan, and the physician community. They help preserve access for patients with severe or unusual disease. They should not be penalized for having a different patient practice.

Placing new and part time physicians in the higher co-pay tier may indirectly make it more difficult to recruit new physicians to the area by adding a barrier to soliciting new patients. A separate category would be more appropriate for these practitioners.

This change would avoid penalizing physicians who are performing a service to the system, and help preserve access for patients with severe or unusual diseases.

Recommendation: The health plans, GIC, Mercer, and physician leadership should develop a written specification and policy, along with a data driven process, to determine if physicians are different from peers. As part of the formal feedback process, allow physicians to appeal as different from peers. Plans in conjunction with Mercer should review physicians (especially those in the non-preferred tier) to understand if their practices are systematically different from expected. These physicians should be designated as "different from peers" rather than tiered.

- 7. Selection of specialties to be tiered.** Tiering may be more appropriate for some specialties than others. For primary care, there are numerous quality measures, a large number of episodes, and a tendency to relatively low costs per episode. The measures are thus more finely graduated. A specialty such as hematology-oncology has the opposite problem: few quality measures, many high cost episodes, and difficulty in defining overused medical services. Selection of specialties to be tiered should be reviewed along the lines of the principles discussed in this report.

Recommendation: Develop a process for selecting which specialties to tier. The process should be guided by data and general principles, with input from a wide array of stakeholders.

VIII. Process Improvement Opportunities

Good inter-personal process is based upon certain core values. These include honesty, integrity, and respect for others. These values and principles are put into action by listening to others' feedback, by admitting mistakes openly, and by acting on that feedback when appropriate. A related practice is to avoid judging *persons* as good or bad, but rather to evaluate *behaviors* as more or less productive or appropriate. While we are certain that all involved with the CPI Initiative personally share these values and principles, the ways in which the program was developed and rolled out have not consistently reflected them.

Good inter-personal process respects the IOM goal of patient-centered care, and has been shown to be important to the patient-physician relationship.²⁶ Researchers have now extended that concept to the administration-physician relationship through a model called relationship-centered administration.²⁷

It has been the authors' experience that successful physician evaluation programs, including pay-for-performance, public reporting, and tiering systems, share the characteristic of good process in terms of relationship-centered administration. We have touched upon several issues in the preceding sections that have significant process implications. Below we highlight several areas in which improved process would make the CPI Initiative more effective by decreasing contentiousness and improving physician engagement.

1. **Recommendation: Clarify and then widely articulate the program's core values.** These then should be used as a touchstone for the goals, process, and outcomes of the CPI Initiative.
2. **Recommendation: Offer physicians meaningful input into the program.** A successful roll-out process proceeds in a certain order: convening of stakeholders, discussion of values, principles, and goals, shared decision making, private reporting, opportunity to review and correct, public reporting, use of public reporting (tiering). The more closely that future interactions with the physician community follow this pattern, the more likely they will meet with success.
3. **Recommendation: Share results with physicians well in advance of publicly reporting them.** When results are shown to patients before their physicians, the physicians will naturally feel unfairly treated. Such lapses in protocol predictably create anger and distrust, and are virtually guaranteed to create physician backlash. Furthermore, they do not meet emerging national standards for physician evaluation programs.^{5,28,29}
4. **Recommendation: Give practitioners patient-level data and correct identified errors prior to public reports.** It is not enough merely to share results. The opportunity for physicians to review the data about the quality of their work and point out data errors is especially important. We have already discussed this point

in terms of continuously improving the accuracy of the tiering system. However, it is extremely important in terms of process. Listening and then changing the system will generate allies rather than antagonists.³⁰ In addition, there should be an appeal process to rectify errors. Correction of errors increases mutual trust. Plans should develop the policies, processes, and infrastructure that would allow systematic acceptance of feedback and corrections of data, process, and methodology.

5. **Recommendation: Provide physicians with specific behaviors (action items) by which they can improve their results.** Without action items there is no opportunity for improvement, and that means the system functions purely as judgment. Action items must withstand the test of face-validity, while a lack of actionability bewilders physicians about what to do with their scores. Judgment without ability to change is a formula for increasing frustration. Mercer and plans should develop a mechanism as soon as possible for distributing action items related to both the cost-efficiency and quality measures. These would consist of so-called drilldowns for the efficiency index and patient lists/registries for quality measures. Action items that reduce overuse and underuse help bring about real change and true cost savings. Focusing on action items encourages all practitioners to improve and provides partial assistance to lower performing practices and practitioners.
6. **Recommendation: Demonstrate and publicize a commitment to continuous improvement.** Start making changes that are straightforward improvements to the system. An example of an improvement that could be accomplished immediately is changing the efficiency index calculation to exclude hemophilia for general pediatricians and preventive care for all specialties. The sooner that physicians see the system being changed in response to their concerns, the sooner the CPI Initiative will start building trust and concomitant physician acceptance.
7. **Recommendation: Anticipate the need for physicians to buy in to consensus-based measures.** A number of the measures selected are not evidence based, but rather the result of expert opinion. Examples are numbers of visits for newborns, the frequency of follow-up lab testing in diabetes, and the measurement of lab tests for new hypertensives. There are different regional norms for many measures based on expert opinion. Physicians may not necessarily know or agree with what was expected of them for a given such measure. When quality measures are consensus measures, the recommended behaviors have to be shared in advance of measurement.

IX. Conclusions

Physician performance measures could be used for many purposes:

Quality improvement programs



Pay for performance



Public Reporting



Tiering Networks

For each succeeding purpose along this spectrum, from quality improvement through to network tiering, inaccuracies have more severe consequences for all stakeholders, and especially for physicians and their patients. Therefore, each use requires a corresponding level of rigor that generates sufficiently accurate and usable results. We are not the first to note this relationship.^{7,31,32} Quality improvement programs involving internal reporting work toward improvements at population levels and require less methodological rigor and data accuracy compared with pay for performance, for example, so long as the system measures do indeed improve. Public reporting and tiering of networks, with greater potential repercussions for physicians, patients, and health plans, requires greater rigor.

In order for the GIC to meet its goals, the CPI Initiative should be improved in phases over several years, in a prudent, thoughtful, and deliberate manner. As an interim measure, group level reporting could be a reasonable starting place. However, even at the group level, physicians must have access to cost-efficiency drill downs and quality measure patient lists, and there must be a mechanism to accept their feedback, correct data when appropriate, and refine the measurement system through a continuous quality improvement process.

As the data and the instrument improve, they may reach a point where individual-level physician measurement, tiering, and accountability become appropriate. We have concerns that the CPI Initiative, as it now stands, does not meet the test of sufficient rigor for that purpose.

In this report we have discussed a number of ways in which the CPI Initiative could be modified to better increase the value of the medical care delivered to patients. Underlying the initiative is the significant presumption that creating a tiered network is an appropriate and effective method for bringing about the GIC's goals. With the recognition that we do not address that assumption, since it was outside the scope of the report, our recommendations can be summarized in the following major points:

- Provide physicians patient level data for both the efficiency and quality measures. In that way practitioners can review, correct, and act upon the information.

- Make technical improvements to the efficiency and quality metrics as suggested in this report to increase their reliability and utility.
- Incorporate the changes suggested to improve tier assignment methodology
- Develop a suggested standard tier assignment protocol that can be used by all plans
- Use group rather than individual tiering at least until changes to the system are investigated and improvements are implemented

All parties should realize that physician profiling and accountability only affect one part of health care. This is particularly true for chronic disease care, where a system approach plays an especially important role.^{24,25} Physicians also need help in reaching non-adherent patients. Health plans should support physicians through disease and case management. Such support also will help prevent unintended consequences such as avoiding sicker or more difficult patients.

GIC and Mercer should engage the physician community in improvement of the profiling and tiering system and, ultimately, of patient care itself. To be successful, physician profiling and evaluation programs should adhere to all or most of the following principles: clearly articulate the reason for and values underlying the evaluation program, engage practitioners early on in measure selection and the reporting process, make methodology transparent, provide action items for improvement, assure accurate measurement and present an appeal process for mistakes made, and inform practitioners of the process well in advance of involving their patients. It is our experience that the respect inherent in following these principles will decrease physician frustration and anger. By improving the data base, the tiering instrument, and the relationship with the physician community, the GIC is more likely to achieve its goals, which are shared by all stakeholders.

Appendix I: Summary of Recommendations

Data Accuracy

1. Physicians should be given patient-level drilldowns for the efficiency measure, and patient lists for the quality measures, to help identify systematic accuracy problems.
2. The health plans and Mercer should develop a formal feedback and correction mechanism so that errors uncovered by physicians, plans, and other analysts can contribute to improving the evaluation system.
3. Plans should initiate programs to audit their primary data, especially in specific important areas of concern.
4. Mercer and the plans should examine low and high outliers in detail as this targeted evaluation is likely to identify systematic accuracy issues.
5. Mercer should review its data base for unusual disease coding patterns.
6. Tiering physicians as individuals exacerbates accuracy issues. Therefore tiering should occur at a group level until data accuracy is validated and improved.

Cost-Efficiency Measurement:

1. Establish ETG specific trim points (e.g., at the 2% and 98% levels of the ETG cost distribution) using current Massachusetts claims data, and truncate outlier cost instead of trimming outlier episodes from analysis.
2. Base episode attribution on combined professional and prescribing costs rather than on professional costs alone, presuming reliable prescribing physician information.
3. Implement the new risk adjusted ETG system and incorporate risk adjustment of episode expected costs into its profiling process.
4. Use the most recent two years of claims data to assess performance, and do not use a weighting scheme; truncate instead of trim outlier episodes; score physicians in groups instead of individually in order to increase sample size and thereby increase reliability.
5. Define specialties in terms of the types of ETGs managed, and use only the particular ETGs that define a specialty to calculate the efficiency index.
6. Remove the costs of preventive measures and underuse interventions from the EI calculation. When a routine physical appears in a non-preventive ETGs, convert its cost to that of an average office visit (e.g. a 99213 cost).
7. Use the Standardized Cost Difference (SCD) metric instead of a ratio of observed to expected costs.
8. The effect of benefit design should be assessed and factored out of the physician cost efficiency metric.

Quality Measures:

1. The health plans should provide physicians a list of patients counted towards each quality measure, as well as the interventions for each patient that counted towards that measure, especially since RHI has provided this information. Plans should systematically review the details of the specifications with their physicians so that

- physicians understand which patients are affected and what interventions are expected.
2. Select a smaller number of measures that address conditions that are both prevalent and clinically significant. Require an appropriate minimum sample size before counting a measure towards the overall quality performance evaluation. Profile as groups instead of individual physicians in order to overcome the issue of sample size, at least until the numerator-denominator roll-up methodology has been modified and validated to take into account the fact that its components are not independent measures.
 3. In addition to choosing a smaller number of clinically important measures, plans in conjunction with physician leadership should determine appropriate targets for each measure. Plans should provide support for intervening with patients who do not adhere to interventions despite their physicians' best efforts. The goal should be system change to attain a target level, not merely "better than average" quality.
 4. Apply case-mix adjustment if multiple quality measures are combined into a single score.
 5. For quality measures that involve overuse, the instance of overuse should attach to the attributed physician only when that physician actually generated the overuse. (Individual measures are identified as underuse or overuse in the Appendix.)
 6. CPI Initiative sponsors should investigate, propose, and discuss weighting of quality measures through a process that is transparent. Strong consideration should be given to assigning higher weight to quality measures that are evidence-based and most clearly connected to improved outcomes.

Issues Related to Tier Assignment:

1. The GIC and Mercer, in conjunction with a wide array of physician leaders, should develop a suggested tier assignment protocol that all plans could use.
2. Validate tiering statistically so that there is a reasonable probability that tier assignments are not due to chance. Calculate and make available figures on the reliability of tier assignments. Create categories of "tier not assignable" for physicians for those who cannot be evaluated in a statistically reliable manner and "not different from average" where that is the case.
3. The increased administrative burden of having physicians in the same office assigned to different tiers offsets some of the potential system savings, especially until accurate physician tiering is validated, and therefore tiering should be assigned at a group level rather than by individual physician tiering at this time. In any case, the CPI Initiative should take into account the added costs and issues that would occur if physicians in the same office were assigned to different tiers.
4. Because of the nature of their specialty care or of a practice's organization, certain specialties or practices may not be amenable to profiling on a finer level than a group. Irrespective of other issues around the question of individual versus group tier assignment, in these cases assign tiers at the group level.

5. Include in the tiering system structural elements that decrease the incentive to avoid caring for unusual, complex, or difficult patients. Develop indicators to monitor for unintended consequences. Mechanisms outside the tiering system should support desired consequences.
6. The health plans, GIC, Mercer, and physician leadership should develop a written specification and policy, along with a data driven process, to determine if physicians are different from peers. As part of the formal feedback process, allow physicians to appeal as different from peers. Plans in conjunction with Mercer should review physicians (especially those in the non-preferred tier) to understand if their practices are systematically different from expected. These physicians should be designated as “different from peers” rather than tiered.
7. Develop a process for selecting which specialties to tier. The process should be guided by data and general principles, with input from a wide array of stakeholders.

Process Improvement Opportunities

1. Clarify and then widely articulate the program’s core values.
2. Offer physicians meaningful input into the program.
3. Share results with physicians well in advance of publicly reporting them.
4. Give practitioners patient-level data and correct identified errors prior to public reports.
5. Provide physicians with specific behaviors (action items) by which they can improve their results.
6. Demonstrate and publicize a commitment to continuous improvement.
7. Anticipate the need for physicians to buy in to consensus-based measures.

Appendix II: Review of Resolution Health Quality Measures selected for the GIC Select and Save Program

Measure 1 – Beta 2 agonist use in moderate to severe asthma. Percentage of patients with persistent asthma diagnosed over the past 6 months who have filled a prescription for a short term beta-2 agonist inhaler after their diagnosis. Concern – after diagnosis, patients may require rescue treatment as controller doses are being initiated. The measure requires only 1 prescription, a low threshold for fulfilling the recommendation. Better measures might be the use of a beta 2 agonist without concurrent steroid inhaler use, or overuse of beta-2 agonists. An underuse measure as it stands.

Measure 2 – Moderate-severe asthma patients not receiving non-selective beta blocker. References about value of using selective beta blockers not provided. An overuse/misuse measure (should be attributed to the physician actually prescribing the non-selective beta blocker).

Measure 3 – Chronic inhaled steroid use for asthmatics requiring oral steroids. Filling one prescription is minimally acceptable. The literature suggests 50% adherence as a minimum effective adherence rate over time; the goal would be steroid inhaler use for most of the year. An underuse measure.

Measure 4 – Checking theophylline level for those on theophylline. The numerator is one level during the measurement year. The measure applies only to patients on chronic theophylline as determined by refill patterns. An underuse measure.

Measures 5 – Controller medication use in persistent asthma. Same considerations as measure 3. The denominator is patients diagnosed in the first six months of the measurement period. Current literature suggests that this may result in unreliable identification of patients with true asthma. The equivalent HEDIS measure requires the patient the meet the definition in both the identification and measurement years. This measure should be changed accordingly. An underuse measure.

Measure 6 – Theophylline level testing for new theophylline use. The measure requires a theophylline level be measured within a week of starting theophylline for new patients with COPD. Others have argued that using theophylline in COPD should be avoided so a better measure may be the percentage of patients placed on theophylline. The evidence for benefit for this measure is scant. An underuse measure.

Measure 7 – Asthma patients not over-using beta 2 agonists. The criteria for rescue use abuse (the number of scripts used) and the target are critical here. An overuse measure.

Measure 8 – Appropriate strep testing before antibiotic use in pharyngitis. An underuse measure.

Measure 9 – Antibiotic use in viral URI. It is important to avoid antibiotics for non-bacterial infections. An overuse measure.

Measure 10 – Coumadin prescription for patients with atrial fibrillation – Risk factors and exceptions to Coumadin use are defined in the specifications. These are factors that likely would benefit from physicians’ being able to review and correct the data based on their patient charts. An underuse measure.

Measure 11 – Same as 10 but age > 65. Same concerns. An underuse measure.

Measure 12 – Patients with AFib started on warfarin who have INR checked within one week. Is hospital data available? Is lab data reliably available for all plans? Again, likely would benefit from review and correction of primary data. An underuse measure.

Measure 13 – Patients on warfarin > 6 months getting a PT within 3 months – Some concerns as 12. INR is the correct test. An underuse measure.

Measure 14 – CHF ejection fraction check for patients newly diagnosed as outpatient: This is a useful measure, but the situation to which it applies is of low frequency occurrence. An underuse measure.

Measure 15 – Heart failure patients on ACE-I with potassium checked in measurement year. As above, this measure assumes accurate and equal availability of lab data, which is not always the case. An underuse measure.

Measure 16 – CHF patients not taking a calcium channel blocker other than Norvasc or Plendil. An overuse measure.

Measure 17 – Serum creatinine measured for each patient with HF on ACE-I. Can be accomplished by one blood draw for creatinine, potassium. Evidence for outcome improvement is small. As above, this measure assumes accurate and equal availability of lab data, which is not always the case. An underuse measure.

Measure 18 – Heart failure patients on beta blocker. Good measure. Exclusions are listed in the specifications. Another measure where physician review of patient list would be helpful. An underuse measure.

Measure 19 - Percentage of HF patients taking ACEI or ARB. An important measure. An underuse measure.

Measure 20 – Post MI patients without second or third degree block on beta blocker – Good measures. Depends on accurate coding. MI patients are low frequency events for an individual PCP. An underuse measure.

Measure 21 – Post MI patients on a statin. A very important measure. An underuse measure.

Measure 22 – CAD patients with one lipid test per year. Accuracy of lab data an issue, as above. An underuse measure.

Measure 23 - Patients with CAD NOT on triptans or ergotamine – Good measure, however patient preference may be a factor especially for patients with mild CAD. An overuse measure.

Measure 24 – Post MI patients on persistent beta blockers. A good measure, but MI patients are low frequency events for an individual PCP. An underuse measure.

Measure 25 – Patients hospitalized with MI on beta blockers within 30 days of discharge. Same issues as measure 24. An underuse measure.

Measure 26 – Newly diagnosed hypertensives and UA within one year. A consensus measure. The evidence that demonstrates improved outcomes is low. Newly diagnosed hypertensives are low frequency events for an individual physician. An underuse measure.

Measure 27 – Newly diagnosed HTN getting blood glucose checked. Expert consensus, not evidence based. An underuse measure.

Measure 28 – Newly diagnosed hypertensives getting serum potassium. Reasonable measure. With several other blood tests counting towards hypertension care, one lab draw is heavily weighted compared to other interventions (see above measures as well). An underuse measure.

Measure 29 – Newly diagnosed hypertensives getting a lipid panel. Same as for 28.

Measure 30 – Newly diagnosed hypertensives getting a triglyceride level. Same as for 28.

Measure 31 – Newly diagnosed hypertensives getting a serum creatinine. Same as for 28.

Measure 32 – Hypertension patients receiving short acting CCBs. A reasonable measure. An overuse measure.

Measure 33 – HbA1c or fructosamine test for diabetics. HEDIS measure, well accepted. Generally HbA1c is now the preferred test. An underuse measure.

Measure 34 – Diabetic retinal exam. Acceptable measure. An underuse measure.

Measure 35 – Diabetics with lipid panel. Acceptable measure. An underuse measure.

Measure 36 – Patients with DM, HTN AND/OR nephropathy who are on ACEI or ARB. Concerns are accuracy of ICD9's for renal disease causing many missed patients and the

measure being satisfied by one prescription. The goal is prevention of renal disease which requires long term therapy, not filling one prescription. An underuse measure.

Measure 37 – Diabetics without renal disease measured for microalbuminuria. Denominator describes patient s with PH of diabetes, hypertension OR nephropathy while numerator is diabetics WITH nephropathy. In addition, the measure’s intent is to prevent nephropathy while the numerator is patients with nephropathy. An underuse measure.

Measure 38 – Percent of patients started on lipid lowering drug with lipid level checked within 3 months. No literature suggesting different outcomes if tested at 2, 3, 4, 5 or 6 months. The measure is based on expert opinion, not evidence of efficacy. Measures like this one should be shared before used. An underuse measure.

Measure 39 – Check LFT 3 months after starting statins. Similar concern as in measure 38. An underuse measure.

Measure 40 – Two measurements of Lipid levels before start lipid lowering treatment for patients without CAD. Diabetes is considered a CAD equivalent. Were diabetics included in the CAD diagnostic list? Second, is there evidence that two measurements are needed to more accurately determine appropriateness? A consensus opinion. An underuse measure.

Measure 41 – Rubella test for pre-natal women. Many women will have previous results from a prior pregnancy. If that is outside the three year window there is no way for the system to pick it up. Therefore this is another situation where it is vital for physicians to review the patient list and provide information from outside the current system. The system has to be able to track this information for more than the 3 year data window, otherwise physicians will be providing the same information every year. An underuse measure.

Measure 42 – Hep B test before pregnancy. Same issues as above. An underuse measure.

Measure 43 – Percent of pregnant diabetic women who are not taking an oral agents. An overuse measure.

Measure 44 – Patients with newly diagnosed depression being treated with only anti-anxiety drug. Accuracy of ICD9 depression coding will need to be addressed. An underuse measure (underuse of antidepressants).

Measure 45 – Three f/u visits in 12 weeks after diagnosis of depression. HEDIS measure. Evidence is weak. Phone calls can substitute and are all depression diagnoses included or major depression where more appropriate. An underuse measure.

Measure 46 – Newly diagnosed depression treated at least 12 weeks with antidepressants. – HEDIS measure. An underuse measure.

Measure 47 – Percent of newly depressed patients started on antidepressant who are maintained for 6 months. In primary care, very problematic. Mix of minor and major depression plus adjustment disorders significant. Also SAD. PCPs need education in use of codes prior to including measures. An underuse measure.

Measure 48 – High risk pap testing. An underuse measure.

Measure 49 – Pap within 3 years 18-64. HEDIS measure. The problem with in most plans is overuse (i.e. a yearly Pap smear in low risk patients). How large is the underuse problem in Massachusetts? An underuse measure.

Measure 50 – Women >67 with a fracture who get BMD or treatment within 6 months. Problem, with no upper age limit, the decision about treatment is complex. Many 85+ patients may decide not to be treated and if there is no decision to treat, testing is not appropriate. An underuse measure.

Measure 51 – Chlamydia screen for women 16-25 and sexually active. HEDIS measure. There is a great deal of push back nationally on the need to test everyone. An underuse measure.

Measure 52 – Post breast cancer patients with mastectomy receiving mammogram. An underuse measure.

Measure 53 – Mammogram for women 50-69. HEDIS measure. An underuse measure.

Measure 54 – Peds measure – 4 visits during first 6 months. AAP measure. For visit measures, there is no evaluation of what occurred at the visit. These are based on expert opinion rather than demonstrated efficacy. An underuse measure.

Measure 55 – Peds measure – 3 office visits for children ages 6-12 months. AAP measure. An underuse measure.

Measure 56 – Peds measure – 3 office visits ages 12-24 months. AAP measure. An underuse measure.

Measure 57 – Another visit measure – 1/yr for patients 2-6 years old. AAP measure. An underuse measure.

Measure 58 – another visit measure – 1 visit/2 years of ages 7-18. AAP measure. An underuse measure.

Measure 59 – Patients with low back pain being imaged within 28 days of diagnosis. There may be special populations of practitioners who see referred in patients with red flag symptoms, and there are patients with red flag symptoms that require imaging more urgently. This overuse measures especially requires a different approach from underuse

measures. The measure may have to exclude doctors with more complex practices where indications for imaging are met more often. An overuse measure.

Measure 60 – New ulcer patient receiving H Pylori test within one month – First, the one month is arbitrary and not evidence based. Need target to account for patients with a contraindications for scoping. Depends on accurate coding. If someone codes for ulcer based only on clinical grounds, or prior history, the patient may enter the denominator. As per rubella testing, there needs to be physician chart input and long-term tracking.

[Note: the CPI Initiative describes using 59 measures but there are actually 60 in the documentation that was available for this report.]

Financial Disclosures:

Dr. Greene and Mr. Partridge are the principals of Focused Medical Analytics (FMA), LLC, and Dr. Beckman has equity in the company. FMA provides medical data analyses as well as consulting services related to physician behavior change. Dr. Greene receives an honorarium for participation on the Symmetry Medical Advisory Board. Dr. Thomas has recently consulted about economic profiling of physicians with several health plans and with the American Medical Association, and he is currently examining network tiering options for Mercer and the Massachusetts Group Insurance Commission.

¹ The Commonwealth of Massachusetts Group Insurance Commission. 2006-2007 Guide to Select and Save Plans. (Accessed September 8, 2006, at <http://www.mass.gov/gic/annualenroll2006/selectandsave.pdf>.)

² Mercer Human Resource Consulting. GIC Clinical Performance Improvement Initiative Provider Profiling Analysis Update. Presented June 29, 2005. (Accessed September 8, 2006, at <http://www.unicare-cip.com/PDF/06-29-05%20Commission%20Update.pdf>.)

³ Institute of Medicine. Crossing the Quality Chasm: a new health system for the 21st century. Washington DC: National Academy Press. 2001:39-40.

⁴ American Medical Association. Physician pay for performance initiatives. Chicago. 2004.

⁵ American Medical Association. Principles for pay-for-performance programs. February 24, 2005. (Accessed November 12, 2006 at <http://www.ama-assn.org/ama1/pub/upload/mm/1/finalpfppinciples.pdf>.)

⁶ American Medical Association. Guidelines for pay-for-performance programs. February 24, 2005. (Accessed November 12, 2006 at <http://www.ama-assn.org/ama1/pub/upload/mm/1/finalpfpguidelines.pdf>.)

⁷ Massachusetts Medical Society. Principles for profiling physician performance. Waltham, Massachusetts. 1999.

⁸ Massachusetts Medical Society. Guidelines for measuring, reporting, and rewarding physician performance. Waltham, Massachusetts. May, 2005.

⁹ National Committee for Quality Assurance. Quality Plus Program for Managed Care Organizations and Preferred Provider Organizations. Physician and Hospital Quality. Washington, DC. March 31, 2006:38.

¹⁰ AQA Alliance. AQA Principles in the Use of Registries for Enhancing Quality of Care through Performance Measurement. October 24, 2006 (Accessed November 12, 2006, at <http://www.aqaalliance.org/files/RegistryPrinciplesDocumentV1Approved.doc>.)

¹¹ Chassin MR, Galvin RW. National Roundtable on Health Care Quality. The urgent need to improve health care quality. JAMA 1998;280:1000-1005.

¹² Crossing the Quality Chasm, p. 192.

¹³ Mercer Human Resource Consulting. GIC Clinical Performance Improvement Initiative Provider Profiling Study. Presented February 18, 2005. (Accessed September 8, 2006, at <http://www.unicare-cip.com/PDF/02-18-05%20Health%20Plan%20meeting%20deck.pdf>.)

¹⁴ Mercer presentation of June 29, 2005

¹⁵ Thomas JW, Ward K. Outlier treatment and episode attribution rules for economic profiling of physician specialists. *Inquiry* (forthcoming).

¹⁶ Thomas JW, Grazier KL, Ward K. Economic Profiling of Primary Care Physicians: Consistency among Risk-Adjusted Measures. Health Services Research 2004;39:4, (Part I): 985-1003.

¹⁷ Mercer presentation of September 18, 2006

¹⁸ Resolution Health, Inc. RHI physician quality measures. April 13, 2006. (Accessed September 8, 2006, at <http://www.unicare-cip.com/PDF/RHI%20PQP%20MEASURES%20BOOKLET%204-06.pdf>.)

¹⁹ Greenfield S, Kaplan SH, Kahn R, Ninomiya J, Griffith JL. Profiling care provided by different groups of physicians: Effects of patient case-mix (bias) and physician-level clustering on quality assessment results. Ann Intern Med 2002;136(2):111-121.

-
- ²⁰ Hofer TP, Hayward RA, Greenfield S, et al. The unreliability of individual physician “report cards” for assessing the costs and quality of care of a chronic disease. *JAMA* 1999;281(22):2098-2105.
- ²¹ Sepucha KR, Fowler FJ Jr, Mulley AG Jr. Policy support for patient-centered care: the need for measurable improvements in decision quality. *Health Aff* 2004;Suppl Web Exclusive:VAR54-62.
- ²² Doran T, Fullwood C, Gravelle H, et al. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med* 2006;355:375 - 84.
- ²³ Galvin R. Pay-for-performance: Too much of a good thing? A conversation with Martin Roland. Lessons learned from the U.K. experience with pay-for-performance. *Health Affairs* 2006;25:w412-419;10.1337/hlthaff.25.w412.
- ²⁴ Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. *JAMA* 2002;288:1775 - 1779.
- ²⁵ Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness: The chronic care model, part 2. *JAMA* 2002;288:1909 - 1914.
- ²⁶ Safran DG, Miller W, Beckman H. Organizational dimensions of relationship-centered care. *J Gen Intern Med* 2006;21:S9-15.
- ²⁷ Marvel K, Bailey A, Pfaffly C, Gunn W, Beckman H. Relationship-centered administration: Transferring effective communication skills from the exam room to the conference room. *J Healthcare Management* 2003;48(2):112-123.
- ²⁸ National Committee for Quality Assurance. Quality Plus Program for Managed Care Organizations and Preferred Provider Organizations. Physician and Hospital Quality. Washington, DC. March 31, 2006:47.
- ²⁹ American College of Physicians. The use of performance measurements to improve physician quality of care. Philadelphia, April 19, 2004.
- ³⁰ Beckman H, Suchman AL, Curtin K, Greene RA. Physician reactions to quantitative individual performance reports. *Am J Med Qual* 2006;21:192-197.
- ³¹ James BC. Future trends in pay-for-performance contracting. Presented at the Third Annual World Congress Leadership Summit on Healthcare Quality & Pay-for-Performance Contracting. Boston, Massachusetts. August 2, 2005.
- ³² Lee TH. Measurement of efficiency: Perspective of one integrated delivery system. Presentation at the ABIM Foundation Forum, Colorado Springs, Colorado, July 30, 2007.