Physician Compare Quality Measurement Technical Expert Panel (TEP) Summary Report

December 2013

Acumen, LLC
500 Airport Blvd., Suite 365
Burlingame, CA 94010

[This page is intentionally left blank.]

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 ABOUT THE TEP

The Centers for Medicare & Medicaid Services (CMS) has contracted with Acumen, LLC (henceforth "Acumen") to assist in the selection of physician quality measures for public reporting via the Physician Compare website. As part of the measures selection process, Acumen has convened the Physician Compare Quality Measurement Technical Expert Panel (TEP) with the objectives of obtaining expert input on physician quality measures that CMS has proposed for public reporting and seeking recommendations regarding future quality measures for public reporting. Specifically, the TEP will help meet the Affordable Care Act (ACA) requirement to ensure that data reported on Physician Compare provide an accurate and robust portrayal of physician performance. The TEP composition meets the CMS Measures Management Blueprint criteria, including the involvement of individuals who represent the perspectives of patient/caregiver and purchasers, and technical experts who can provide a broad range of technical experience and expertise in public reporting of performance measures, health care quality improvement, and quality measure development and testing. Table 1.1 lists the 16 individuals who comprise the TEP.

Table 1.1: TEP Member Composition

TEP Member	Position(s),Organization	Location
A.J. Yates, MD	Associate Professor, Department of Orthopedic Surgery/University of Pittsburgh School of Medicine	Pittsburgh, PA
Bettina Berman, RN, MPH	Project Director for Quality Improvement, Jefferson School of Population Health/Thomas Jefferson University	Philadelphia, PA
Dale Shaller, MPA **(TEP Chair)**	Principal, Shaller Consulting Group	Stillwater, MN
David Baker, MD, MPH	Michael A. Gertz Professor in Medicine, Chief of the Division of General Internal Medicine and Geriatrics, and Deputy Director of Institute for Public Health and Medicine at Feinberg School of Medicine, Northwestern University	Chicago, IL
David Casarett, MD, MA	Associate Professor, Division Geriatrics/University of Pennsylvania; Director of Hospice and Palliative Care at Penn Medicine; Medical Director for Research and Quality for the National Hospice and Palliative Care Organization	Philadelphia, PA
Emma Kopleff, MPH	Senior Policy Advisor, National Partnership for Women & Families	Washington, DC
Eric Holmboe, MD	Chief Medical Officer and Senior Vice President, American Board of Internal Medicine	Philadelphia, PA
Gregory Dehmer, MD	Cardiologist, American College of Cardiology	Temple, TX
Jeffrey P. Jacobs, MD	Professor of Cardiac Surgery, Johns Hopkins University	St. Petersburg, FL
Michael Mihlbauer, MS	Practice Administrator, Anesthesiology Associates of Wisconsin	Milwaukee, WI
Richard deBronkart	Co-Founder, Society for Participatory Medicine (ePatient Dave)	Nashua, NH
Robert Krughoff, JD	Founder and President, Center for the Study of Services/Consumers' Checkbook	Washington, DC
Sara Schoelle, DrPH	Vice President, Research & Analysis/National Committee for Quality Assurance	Washington, DC
Sherrie Kaplan, PhD, MSPH, MPH	Professor of Medicine and Assistant Vice Chancellor, Healthcare Evaluation and Measurement Executive Co-Director, Health Policy Research Institute School of Medicine/ University of California, Irvine	Irvine, CA
Ted von Glahn, MS	Senior Director, Performance Information and Consumer Engagement/Pacific Business Group on Health	San Francisco, CA
Thomas Smith*, MD, MS	Vice President of Research & Analysis, New York State Office of Mental Health / Columbia University Medical Center	New York, NY

**Dr. Smith was unable to participate in the teleconferences. He reviewed all materials and provided written feedback.*

To support CMS in the final selection of quality measures for public reporting in 2014, Acumen analyzed the empirical and statistical properties of the measures data and presented our findings to the TEP for consideration. Acumen collected the TEP's feedback via a teleconference meeting, which was held over two three-hour sessions (i.e., on September 30, 2013, from 3-6pm EST, and on October 1, 2013, from 3-6pm EST). Following the meeting, Acumen instructed members of the TEP to provide written feedback using a measure scorecard (see Appendix A). Section 2 summarizes our analysis findings for the measures proposed for public reporting in 2014. Sections 3 and 4 summarize the TEP's input.

2 SUMMARY OF MEASURES SELECTION ANALYSES

Per the 2012 and 2013 Medicare Physician Fee Schedule (PFS) Final Rules, the CMS intends to report, amongst other quality measures, a select set of the 2012 Physician Quality Reporting System (PQRS) - Group Practice Reporting Option (GPRO) Web Interface quality measures (i.e., Diabetes Mellitus [DM] and Coronary Artery Disease [CAD]) in 2014. CMS still has to select the specific DM and CAD quality measures for public reporting; for the 2012 PQRS-GPRO Web Interface data, CMS may report up to *eight* DM measures and *three* CAD measures. Table 2.1 summarizes these 11 candidate measures.[1] The first column lists the PQRS measure number. The second column provides the title of each measure. The third column describes the patient outcome or clinical care process that each measure assesses the group practice on. The fourth column identifies the measure developer, which includes the National Committee for Quality Assurance (NCQA), the Minnesota Community Measurement (MNCM), and the American Medical Association-Physician Consortium for Performance Improvement (AMA-PCPI). The final column indicates the measure type (i.e., outcome or process).[2]

Table 2.1: 11 Candidate Measures for Public Reporting

Measure Number	Measure Title	Measure Description	Measure Developer	Measure Type
GPRO DM-3	High Blood Pressure Control in Diabetes Mellitus	Percentage of patients aged 18 through 75 years with diabetes mellitus who had most recent blood pressure in control (less than 140/90 mmHg)	NCQA	Outcome
GPRO DM-5	Low Density Lipoprotein (LDL-C) Control in Diabetes Mellitus	Percentage of patients aged 18 through 75 years with diabetes mellitus who had most recent LDL-C level in control (less than 100 mg/dL)	NCQA	Outcome
GPRO DM-2	Hemoglobin A1c Poor Control in Diabetes Mellitus	Percentage of patients aged 18 through 75 years with diabetes mellitus who had most recent hemoglobin A1c greater than 9.0%	NCQA	Outcome
GPRO DM-10	Hemoglobin A1c Control (< 8%)	The percentage of patients 18 through 75 years of age with a diagnosis of diabetes (type 1 or type 2) who had HbA1c < 8%	NCQA	Outcome
GPRO DM-7	Diabetic Eye Exam in Diabetic Patient	Percentage of patients aged 18 through 75 years with a diagnosis of diabetes mellitus who had a dilated eye exam	NCQA	Process
GPRO DM-8	Foot Exam	The percentage of patients aged 18 through 75 years with diabetes who had a foot examination	NCQA	Process

[1] The measure specifications can be downloaded from: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/Downloads/2012_PhysQualRptg_GPRO_Measures_List_Specs_Release_Notes12152011.zip

[2] Appendices A, B, and C provide the measure specifications for the measures listed in Table 2.1.

Measure Number	Measure Title	Measure Description	Measure Developer	Measure Type
GPRO DM-11	Daily Aspirin Use for Patients with Diabetes and Ischemic Vascular Disease	Percentage of patients ages 18 to 75 years of age with diabetes mellitus and ischemic vascular disease with documented daily aspirin use during the measurement year unless contraindicated	MNCM	Process
GPRO DM-12	Tobacco Non-Use	Percentage of patients with a diagnosis of diabetes who indicated they were tobacco non-users	MNCM	Process
GPRO CAD-1	Antiplatelet Therapy	Percentage of patients aged 18 years and older with a diagnosis of coronary artery disease seen within a 12 month period who were prescribed aspirin or clopidogrel.	AMA-PCPI	Process
GPRO CAD-2	Lipid Control	Percentage of patients aged 18 years and older with a diagnosis of coronary artery disease seen within a 12 month period who have a LDL-C result < 100 mg/dL OR patients who have a LDL-C result ≥ 100 mg/dL and have a documented plan of care to achieve LDL-C < 100 mg/dL, including at a minimum the prescription of a statin	AMA-PCPI	Process
GPRO CAD-7	Angiotensin-Converting Enzyme (ACE) Inhibitor or Angiotensin Receptor Blocker (ARB) Therapy for Patients with CAD and Diabetes and/or Left Ventricular Systolic Dysfunction (LVSD)	Percentage of patients aged 18 years and older with a diagnosis of coronary artery disease seen within a 12 month period who also have diabetes OR a current or prior Left Ventricular Ejection Fraction (LVEF) < 40% who were prescribed ACE inhibitor or ARB therapy	AMA-PCPI	Process

Amongst other requirements, Section 10331 of the 2010 Patient Protection and Affordable Care Act (ACA) requires that the data made public via Physician Compare are statistically valid and reliable, including any risk adjustment mechanisms as appropriate. To support CMS in the final selection of quality measures for public reporting in early 2014, Acumen analyzed the empirical and statistical properties of group practices' data collected via the 2012 GPRO Web Interface and presented our findings to the TEP for consideration. The remainder of this section summarizes the analysis findings that were presented to the TEP.

2.1 Data Quality Assurance

To ensure the reliability and accuracy of the quality measures reported through Physician Compare, Acumen performed three sets of QA checks. First, Acumen verified that the group-level performance rates in the group-level files are largely consistent with the beneficiary-level performance rates across the DM and CAD quality measures (i.e. beneficiary-level data aggregate up to the group-level performance rates). There are minor discrepancies for three group practices in the diabetes measures, where Acumen found 1-2 extra eligible beneficiaries in the beneficiary-level data. After reaching out to the GPRO team, Acumen determined that that these group practices entered information for a few beneficiaries into the Web Interface without correctly submitting that data, leading to this minor discrepancy.

Second, Acumen confirmed that group practices have correctly adhered to CMS criteria in "skipping" or excluding patients in their DM and CAD samples when the patient was deceased. Across all reasons for skipping beneficiaries, the average skip rate among group practices is 8%. For both DM and CAD measures, about 85% of all skips resulted from not being able to confirm that the sampled patient had the condition.

Finally, Acumen leveraged Medicare Part D claims data from 2012 to confirm that group practices correctly identified the patients who met the measure numerator for **CAD-7**. [3] Acumen found that among sampled beneficiaries enrolled in Medicare Part D, 89% of the patients that group practices identified as meeting the measure numerator actually had ACE/ARB therapy prescriptions in their claims histories.

2.2 Measure Performance Rates

Because the Physician Compare website aims to provide actionable information for users to choose physicians and other health care professionals based on the quality of care, CMS should only publicly report performance rates that enable users to make "meaningful" comparisons across physicians and other health care professionals. The performance rate refers to the percentage of each group practice's sampled beneficiaries who met the criteria for a particular measure. Table 2.2 shows the distribution of group practice performance rates across the measures. For all measures except **DM-2,** meeting the measure is better than not, so a higher rate indicates better performance (e.g., patients meeting measure **DM-10** are successfully managing blood glucose levels better than those who do not). For **DM-2**, however, patients who

[3] Acumen could only confirm the measure numerator for **CAD-7** but not for other measures for several reasons. First, while Medicare claims (which Acumen can readily access) provide the information needed to verify the measure numerator for **CAD-7,** they do not provide the information needed to conduct a similar analysis for other measures. Second, although group practices agree to provide CMS access to review the Medicare beneficiary data on which PQRS GPRO submissions are based as part of the nomination process, Acumen does not have access to this data.

meet this measure have recorded blood glucose levels exceeding a healthy threshold – thus, not meeting the measure represents a better patient outcome (i.e., a lower rate indicates better performance).

Table 2.2: Distribution of Group Practice Performance Rates Across Measures

Measure		Minimum	25th Percentile	Median	75th Percentile	Maximum	Mean	Std. Deviation
Outcome	**DM-3**	56.2	65.3	70.8	75.5	88.3	70.6	7.01
	DM-5	31.4	50.6	56.6	63.5	74.0	56.4	9.47
	DM-2	6.3	13.0	16.3	23.6	39.6	18.8	7.88
	DM-10	48.3	67.2	73.0	77.2	85.7	71.5	7.74
Process	**DM-7**	5.0	44.7	60.4	69.3	94.7	57.1	18.7
	DM-8	17.7	61.6	71.1	85.6	100	69.7	19.6
	DM-11	31.0	80.9	88.2	91.5	100	83.4	14.9
	DM-12	28.7	77.9	82.7	86.7	92.4	80.1	10.9
	CAD-1	73.8	87.3	91.3	95.9	100	90.8	5.9
	CAD-2	46.7	68.1	79.8	89.4	100	77.6	13.8
	CAD-7	52.8	80.1	85.7	89.9	100	83.7	9.4

CMS should only publicly report performance rates that enable users to make "meaningful" comparisons across group practices. To assess whether the comparison of performance results across group practices is meaningful, Acumen examined the statistical significance of the measure performance rates by comparing the rate of each group practice with the mean rate across all group practices. Acumen conducted a two-sided test of the null hypothesis that the group practice's performance is not different from the mean performance of all group practices with at least 25 measure-eligible cases at the 5% significance level. Across all measures, *more than half* of the group practices' rates differ statistically from the mean, implying that there is enough variation between groups to make meaningful comparisons.

2.3 Reliability Scores

Measure reliability refers to the extent to which differences in each quality measure were due to actual differences in group practice performance versus variation that arises from measurement error. Statistically, reliability depends on performance variation for a measure across group practices, the random variation in performance for a measure within a provider's panel of attributed beneficiaries, and the number of beneficiaries attributed to the provider. High reliability for a measure suggests that comparisons of relative performance across group practices are likely to be stable over different performance periods, and that the performance of one group practice on the quality measure can confidently be distinguished from another. Potential reliability values range from zero to one, where one (highest possible reliability) means that all variation in the measure's rates is the result of variation in differences in performance

across group practices, while zero (lowest possible reliability) means that all variation is a result of measurement error.[4]

Acumen fit a beta-binomial model to calculate reliability scores for each measure. Acumen concluded that reliability was high across all measures; the 25th percentile ranged from 0.89 for **DM-3** to 0.99 for **DM-8**, which is well above the range considered acceptable for drawing inferences about group practices (i.e., 0.70 – 0.80). Additionally, to measure between group-practice variation and within group-practice variation, Acumen calculated the test-retest reliability using the intra-class correlation coefficient (ICC). ICC values that approach 1 indicate that the fraction of the total variance due to between-group variation is high. The ICC values across all measures range from 0.79 for **DM-3** to 0.97 for **DM-8**, indicating that most of the total variation is due to between-group practice variation. Table 2.3 shows the reliability coefficients and ICC values for each measure.

Table 2.3: Reliability of Group Practice Performance Rates

Measure		Reliability Coefficient					ICC
		Minimum	25th Percentile	Median	75th Percentile	Maximum	
Outcome	DM-3	0.76	0.89	0.90	0.92	0.95	0.79
	DM-5	0.84	0.93	0.93	0.94	0.95	0.89
	DM-2	0.79	0.93	0.94	0.95	0.97	0.90
	DM-10	0.74	0.90	0.92	0.93	0.95	0.87
Process	DM-7	0.96	0.98	0.99	0.99	1.00	0.96
	DM-8	0.97	0.99	0.99	0.99	1.00	0.97
	DM-11	0.73	0.91	0.95	0.97	1.00	0.84
	DM-12	0.84	0.95	0.96	0.97	0.99	0.92
	CAD-1	0.59	0.93	0.96	0.97	1.00	0.88
	CAD-2	0.84	0.98	0.98	0.99	1.00	0.96
	CAD-7	0.62	0.90	0.92	0.95	1.00	0.80

2.4 Impact of Case-Mix on DM Outcome Measures

Case-mix adjustment refers to the statistical process of identifying and adjusting for differences in population characteristics (i.e., risk factors) before comparing outcomes of care. While case-mix should *not* be applied to certain structure and process measures, case-mix adjustment may be necessary for outcome measures that are not fully within the measured providers' control (e.g., blood sugar or cholesterol control).

[4] For more information about reliability testing for physician performance measurement, as well as the methodology for constructing the reliability score reported on Table 6, see "Reliability of Provider Profiling: A Tutorial" by John Adams, RAND. http://www.rand.org/pubs/technical_reports/TR653.html.

The DM outcome measures (i.e., **DM-3**, **DM-5**, **DM-2**, and **DM-10)** do not include case-mix adjustment as part of their specifications. To determine the impact of patient characteristics on group practice performance rates across measures, Acumen adjusted the group performance rates for certain patient attributes that are outside the control of a group practice (e.g., demographic characteristics and pre-existing health conditions) and evaluated how group practice performance rates changed with case-mix adjustment.

To compare the impact of different sets of case-mix factors, Acumen constructed two predictive models; Model 1 includes selected basic demographic characteristics and health status variables (e.g., age, gender, reason for Medicare eligibility) that Medicare commonly uses as part of case-mix adjustment for other publicly reported measures. However, group practice performance rates may vary systematically based on racial and regional attributes that Medicare does *not* typically use for case-mix adjustment; Model 2 is an expanded model that includes these additional characteristics (e.g., race, income, and region type). Based on these models, Acumen reached the following conclusions about the impact of case-mix adjustment on the DM outcome measures:

(1) In Model 1, the case-mix consisting of the basic demographic indicators (i.e., age, and gender) and health status variables (i.e., Medicare eligibility, Medicaid eligibility, and the presence of assorted health conditions) does predict, on the patient-level, the probability of meeting a measure. However, this case-mix does not differ significantly across the 67 group practices included in the sample. By comparing actual performance rates with performance rates predicted by this case-mix, Acumen found that case-mix adjustment has only a small impact on the predictive ability of Model 1.

(2) In Model 2, when the case-mix was expanded to include additional demographic and regional characteristics (i.e., race, region, region type, household income, and home value), predicted performance rates differed from actual performance rates on the group practice level.

Although DM outcome measures, as specified, are *not* risk-adjusted, the TEP generally concluded it would be acceptable to move forward with publicly reporting a subset of the DM outcome measures in the short-term *without* risk adjustment, because public reporting targets only fairly large group practices at this time (see Section 3.1). Additionally, as part of the NQF endorsement process, risk adjustment was deemed not appropriate for these measures; consumer representatives and multi-stakeholder groups have reviewed and supported these measures. That said, the TEP felt there would need to be strong messaging around risk adjustment to accompany the public reporting efforts, including an acknowledgment of the limitations of claims data and how CMS will consider risk adjustment for reporting outcome measures at the individual clinician-level going forward.

3 TEP INPUT ON DM AND CAD MEASURES

As part of the measure evaluation process, the TEP reviewed the measure specifications and analysis findings above to assess the appropriateness of the measures for public reporting. Acumen instructed members of the TEP to rate each performance measure on a scale from one to five, where one corresponds to "not appropriate for public reporting" and five corresponds to "very appropriate for public reporting." For measures that received a rating of *less than* 3, Acumen instructed the TEP to recommend changes to the measure specifications and data collection method that would make the measures more suitable for public reporting. 13 out of the 16 TEP members rated the measures. Table 3.1 below displays the average rating for each measure.

Table 3.1: Average TEP Ratings for each Measure

Measure	Average Score (Out of 5)
CAD-2: Lipid Control	4.23
DM-5: Most Recent LDL-C Result	4.15
DM-3: Blood Pressure Management	4.00
DM-10: Most Recent HbA1c Results (High)	3.92
CAD-7: Diabetes / LVSD and ACE-I / ARB	3.85
DM-2: Most Recent HbA1c Result (Low)	3.62
DM-11: IVD / Aspirin Use	3.62
DM-12: Tobacco Non-Use	3.46
CAD-1: Antiplatelet Therapy	3.00
DM-7: Eye Exam	2.69
DM-8: Foot Exam	2.46

Overall, both lipid control measures (**CAD-2** and **DM-5**) received the highest average ratings. Based on the comments, TEP members generally felt these measures were clinically relevant, and the locus of control was with the provider (rather than the patient). Blood pressure management (**DM-3**) and treatment with ACE inhibitors or ARBs (**CAD-7**) also received high scores for similar reasons.

Two measures of blood sugar control are under consideration: **DM-2** measures the proportion of diabetic patients with high blood sugar (HbA1c > 9%), while **DM-10** measures the proportion of diabetic patients who have controlled blood sugar (HbA1c < 8%). The TEP felt a blood sugar measure was clinically important, but that reporting both measures would be redundant. Unlike all other measures where better performance (i.e. better care) is indicated by a higher rate, better **DM-2** performance is indicated by a lower rate, which could be confusing to consumers.

The TEP felt that the clinical processes captured by **DM-11** and **DM-12** may not be fully under the control of providers. **DM-11** evaluates group practices on whether their diabetic patients with ischemic vascular disease take aspirin daily; however, aspirin is an over-the-counter drug that does not require a prescription and therefore, may not be reliably captured in patient records. **DM-12** assesses group practices on whether their patients with diabetes are also tobacco smokers; however, performance on this measure is heavily influenced by the prevalence of smoking among the patient population.

Finally, the TEP noted that **CAD-1**, **DM-7**, and **DM-8** were collected by the PQRS for program year 2012 but will no longer be collected starting in program year 2013 and beyond. Publicly reporting these measures for program year 2012 but not for subsequent years would be confusing to health consumers; several TEP members agreed there is value in maintaining a more consistent measure set for public reporting over time. Sections 3.1, 3.2, and 3.3 present the TEP's feedback regarding the DM outcome measures, DM process measures, and CAD process measures respectively.

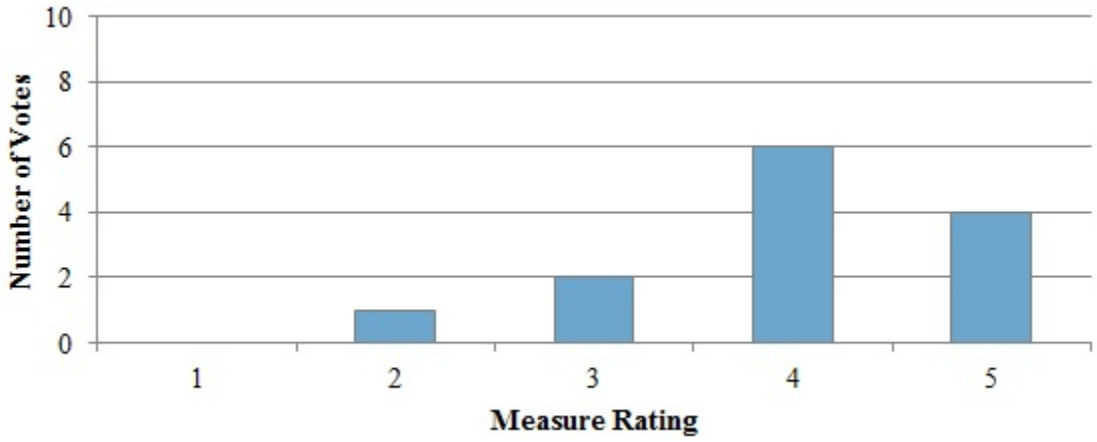
3.1 TEP Input on DM Outcome Measures

CMS is considering four DM outcome measures for public reporting via the Physician Compare website. **DM-3, DM-5,** and **DM-10** are among the top five quality measures by average TEP rating (see Table 3.1). **DM-5** received the second highest average score (4.15). **DM-3** received the third highest average score (4.00). **DM-10** received the fourth highest average score (3.92). The remainder of this section describes the TEP feedback received on each measure in greater detail.

3.1.1 DM-3: Blood Pressure Control

DM-3 received an average score of 4.00. Figure 3.1 below shows the distribution of the TEP's measure ratings for **DM-3**; 10 TEP members assigned a score of 4 *or greater*, 2 TEP members assigned a score of 3, and 1 TEP member assigned a score *less than* 3.

Figure 3.1: Distribution of TEP Ratings for DM-3



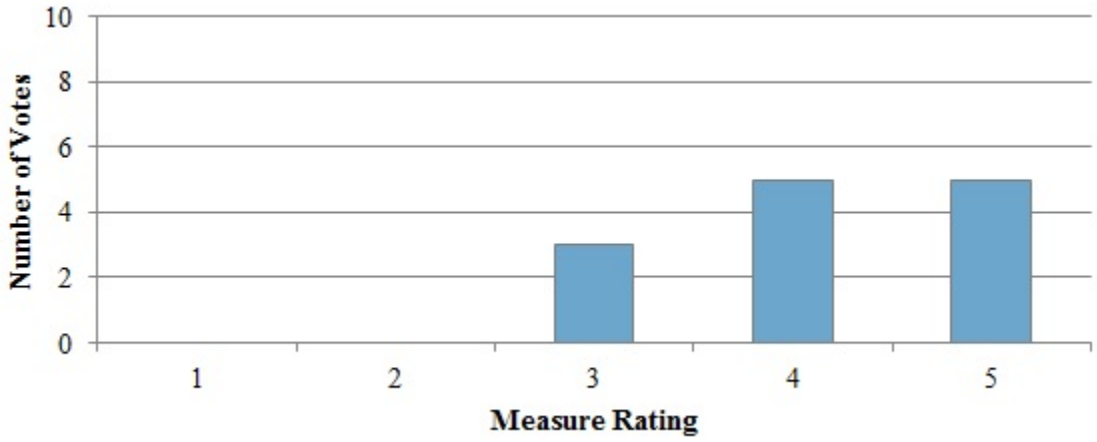
Of those who assigned a score of 4 *or greater,* several TEP members stated that it was acceptable to report the group-practice level measures without case-mix adjustment. One TEP member's reasoning was that **DM-3** was NQF endorsed without case-mix adjustment. Another noted that case-mix adjustment based on patient socioeconomic status, race, ethnicity and other non-clinical patient characteristics is not recommended, since there is a need for quality improvement resources and efforts to be appropriately directed to areas in which disparities exist (which can only be identified without adjusting away disparities). Moreover, case-mix adjustment analysis conducted by Acumen demonstrates limited effect of patient characteristics on patient outcomes at the group practice level. However, several TEP members emphasized that case-mix adjustment will be critical in the future when reporting at the individual eligible professional-level.

One TEP member who assigned a score of 3 noted that 140/90 mmHg is a reasonable blood pressure level target for most but not all patients (due to underlying patient characteristics). One TEP member who assigned a score of 2 stated that there was significant movement in the rankings with the expanded model, and commented that he/she would be fully in favor of this measure if the expanded case-mix model was used for reporting.

3.1.2 DM-5: Cholesterol Control

DM-5 received an average score of 4.15. Figure 3.2 below shows the distribution of the TEP's measure ratings for **DM-5**; 10 TEP members assigned a score of 4 *or greater,* 3 TEP members assigned a score of 3, and no TEP members assigned a score *less than* 3.

Figure 3.2: Distribution of TEP Ratings for DM-5



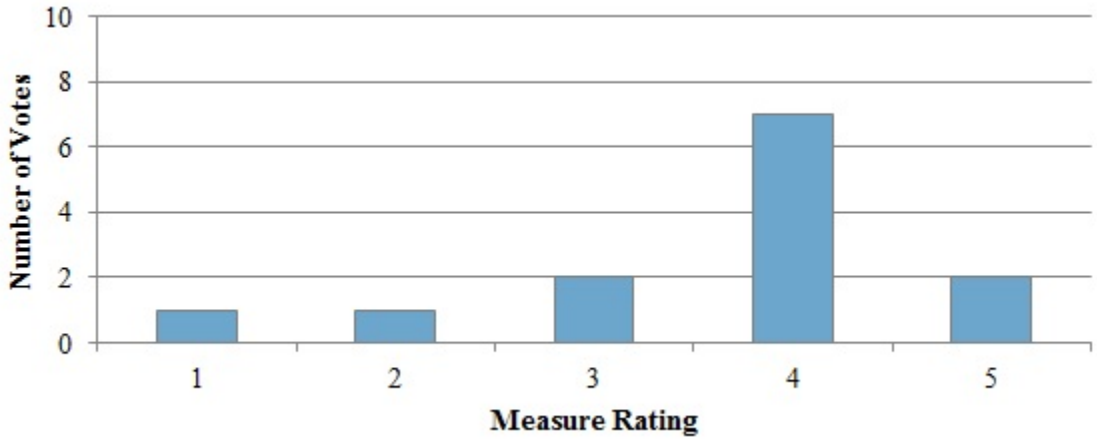
Of those who assigned a score of 4 *or greater,* several TEP members stated that it was acceptable to report the group-practice level measures without case-mix adjustment. One rationale echoed by several TEP members was that **DM-5** was NQF endorsed without case-mix adjustment. Case-mix adjustment based on patient socioeconomic status, race, ethnicity and other non-clinical patient characteristics is not recommended, since there is a need for quality improvement resources and efforts to be appropriately directed to areas in which disparities exist (which can only be identified without adjusting away disparities). Moreover, case-mix adjustment analysis conducted by Acumen demonstrates limited effect of patient characteristics on patient outcomes at the group practice level. There was also minimal change in the rankings in the expanded case-mix model. However, case-mix adjustment will be critical in the future when reporting at the individual eligible professional-level. Of those who assigned a score of 3, one TEP member stated that the performance rates across group practices were relatively low, and that publicly reporting low-performing measures would make it more difficult to engage physicians in quality improvement.

During the TEP discussion, a TEP member raised the concern that if DM-5 is publicly reported, it may lead to unintended consequences, as it may encourage increased use of medicines that are not evidence-based in an attempt to get patient's cholesterol level below the threshold needed to meet the measure.

3.1.3 DM-2: Poor Blood Glucose Control

DM-2 received an average score of 3.62. Figure 3.3 below shows the distribution of the TEP's measure ratings for **DM-2**; 9 TEP members assigned a score of 4 *or greater*, 2 TEP members assigned a score of 3, and 2 TEP members assigned a score *less than* 3.

Figure 3.3: Distribution of TEP Ratings for DM-2



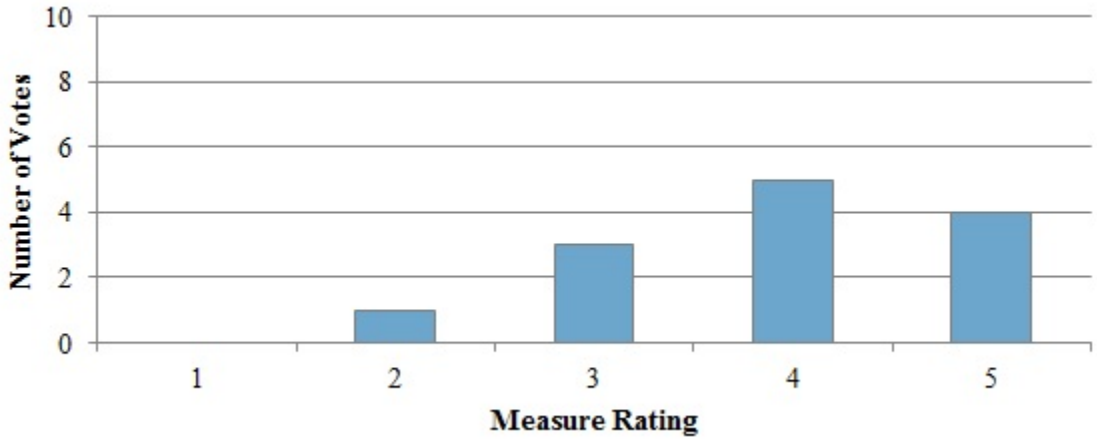
Of those who assigned a score of 4 *or greater,* several TEP members stated that it was acceptable to report the group-practice level measures without case-mix adjustment. One rationale echoed by several TEP members was that **DM-2** was NQF endorsed without case-mix adjustment. Case-mix adjustment based on patient socioeconomic status, race, ethnicity and other non-clinical patient characteristics is not recommended, since there is a need for quality improvement resources and efforts to be appropriately directed to areas in which disparities exist (which can only be identified without adjusting away disparities). Moreover, case-mix adjustment analysis conducted by Acumen demonstrates limited effect of patient characteristics on patient outcomes at the group practice level. However, case-mix adjustment will be critical in the future when reporting at the individual eligible professional-level.

Of those who assigned a score of 3 *or less,* several TEP members recommended that CMS report either **DM-2** or **DM-10** but not both. Since these two measures assess the same outcome (i.e., HbA1c control), reporting both simultaneously may cause consumers to be confused. Additionally, **DM-2** could be confusing because a lower value indicates better performance. Another TEP member stated that there was significant movement in the rankings with the expanded case-mix model and commented that patient outcomes for **DM-2** are often influenced by psychosocial issues or health care access barriers rather than quality of care.

3.1.4 DM-10: Good Blood Glucose Control

DM-10 received an average score of 3.92. Figure 3.4 below shows the distribution of the TEP's measure ratings for **DM-10**; 9 TEP members assigned a score of 4 *or greater,* 3 TEP members assigned a score of 3, and 1 TEP member assigned a score *less than* 3.

Figure 3.4: Distribution of TEP Ratings for DM-10



Of those who assigned a score of 4 *or greater,* several TEP members stated that it was acceptable to report the group-practice level measures without case-mix adjustment. One rationale echoed by several TEP members was that **DM-10** was NQF endorsed without case-mix adjustment. Case-mix adjustment based on patient socioeconomic status, race, ethnicity and other non-clinical patient characteristics is not recommended, since there is a need for quality improvement resources and efforts to be appropriately directed to areas in which disparities exist (which can only be identified without adjusting away disparities). Moreover, case-mix adjustment analysis conducted by Acumen demonstrates limited effect of patient characteristics on patient outcomes at the group practice level. However, case-mix adjustment will be critical in the future when reporting at the individual eligible professional-level.

One TEP member who assigned a score of 3 commented that although there was some movement in the rankings with expanded case-mix model, the extent of the movement was less than **DM-2**.

3.2 TEP Input on DM Process Measures

CMS is considering four DM process measures for public reporting via the Physician Compare website. The DM process measures are among the bottom five quality measures by average TEP rating (see Table 3.1). **DM-8** received the lowest average score (2.46). **DM-7** received the second lowest average score (2.69). **DM-12** received the fourth lowest average score (3.46). **DM-11** received the fifth lowest average score (3.62). The remainder of this section describes the TEP feedback received on each measure in greater detail.

3.2.1 DM-7: Dilated Eye Exam

DM-7 received an average score of 2.69. Figure 3.5 below shows the distribution of the TEP's measure ratings for **DM-7**; 4 TEP members assigned a score of 4 *or greater*, 3 TEP members assigned a score of 3, and 6 TEP members assigned a score *less than* 3.

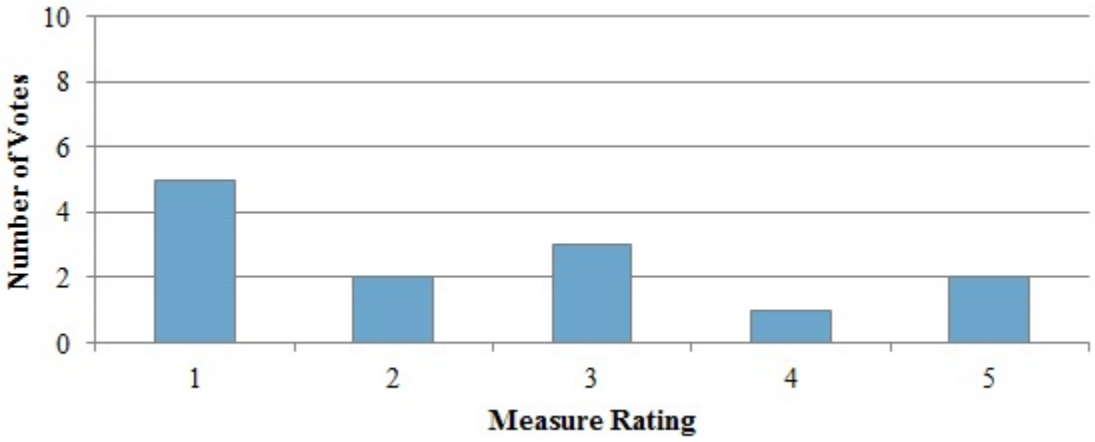
Figure 3.5: Distribution of TEP Ratings for DM-7

One TEP member who assigned a score of 5 stated that process measures in general offer a better introduction to public reporting, are likely to be more meaningful to consumers compared to outcome measures, and are more likely than outcome measures to get buy-in from physicians. Of those who assigned a score of 3 *or less*, several TEP members cited the fact that **DM-7** is no longer collected as part of PQRS after 2012. Publicly reporting this measure for program year 2012 but not for subsequent years may be confusing to health consumers; several TEP members agreed there is value in maintaining a more consistent measure set for public reporting over time. One TEP member, who assigned a score of 1, stated that the measure is not consistent with current clinical guidelines to do an eye exam every two years for patient who has good glycemic control.

3.2.2 DM-8: Foot Exam

DM-8 received an average score of 2.46. Figure 3.6 below shows the distribution of the TEP's measure ratings for **DM-8**; 3 TEP members assigned a score of 4 *or greater*, 3 TEP members assigned a score of 3, and 7 TEP members assigned a score *less than* 3.

Figure 3.6: Distribution of TEP Ratings for DM-8



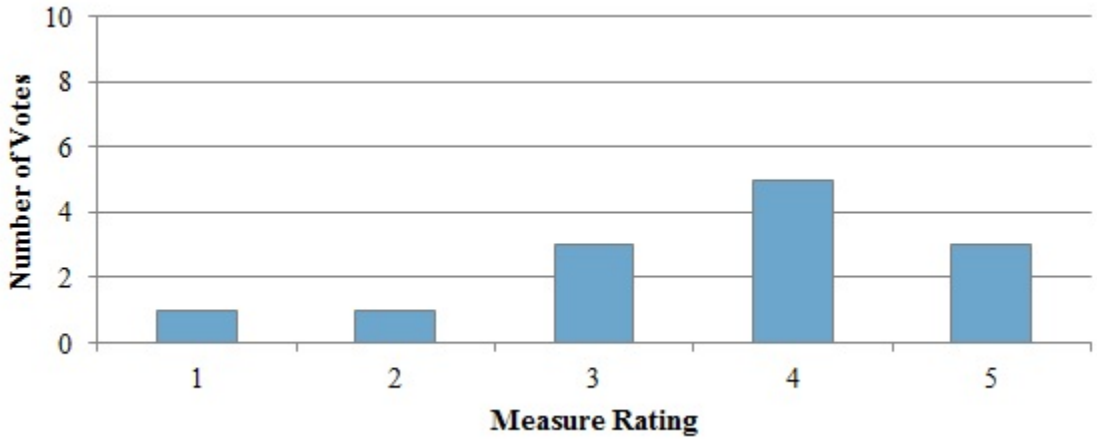
Of those who assigned a score of 3 *or less*, several TEP members cited the fact that **DM-8** is no longer collected as part of PQRS after 2012. Publicly reporting this measure for program year 2012 but not for subsequent years may be confusing to health consumers; several TEP members agreed there is value in maintaining a more consistent measure set for public reporting over time. One TEP member, who assigned a score of 1, stated that the measure is not consistent with recommended clinical guidelines.

During the TEP discussion, a TEP member commented that foot exam is essential to good care and wanted to know why **DM-8** was being retired from PQRS. Another TEP member stated that it was difficult to get documentation on foot exam.

3.2.3 DM-11: Aspirin Use

DM-11 received an average score of 3.62. Figure 3.7 below shows the distribution of the TEP's measure ratings for **DM-11**; 8 TEP members assigned a score of 4 *or greater*, 3 TEP members assigned a score of 3, and 2 TEP members assigned a score *less than* 3.

Figure 3.7: Distribution of TEP Ratings for DM-11

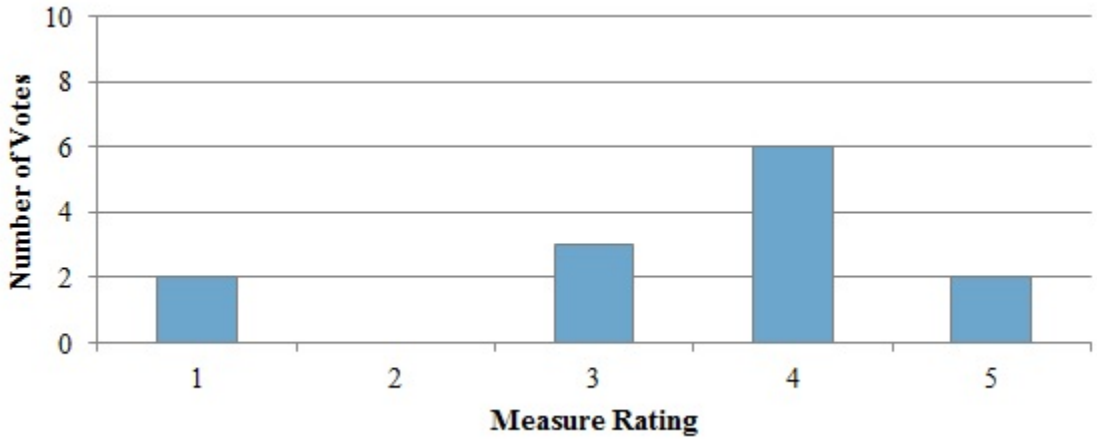


One TEP member who assigned a score of 4 stated that there exists ample variation in performance across group practices and clear clinical evidence supporting the importance of this measure. One TEP member who assigned a score of 3 commented that including **DM-11** may be confusing to consumers, since it is unclear to consumers whether the process assessed by the measure is a provider or patient-driven process. One TEP member who assigned a score of 2 stated that academic medical centers participating in the PQRS GPRO have shown difficulties in capturing **DM-11**, since aspirin is an over-the-counter drug, and so aspirin use might not be reliably captured in the patient records. Finally, one TEP member who assigned a score of 1 commented that the evidence underlying **DM-11** is uncertain.

3.2.4 DM-12: Tobacco Non-Use

DM-12 received an average score of 3.46. Figure 3.8 below shows the distribution of the TEP's measure ratings for **DM-12**; 8 TEP members assigned a score of 4 *or greater*, 3 TEP members assigned a score of 3, and 2 TEP members assigned a score *less than* 3.

Figure 3.8: Distribution of TEP Ratings for DM-12



One TEP member, who assigned a score of 4, commented that **DM-12** is an important process measure to capture and that there is ample variation in the performance for **DM-12**. One TEP member, who assigned a score of 3, commented that including **DM-12** may be confusing to consumers, since it is unclear to consumers whether the process assessed by the measure is a provider or patient-driven process. Another TEP member, who assigned a score of 1, stated that performance on **DM-12** is not fully under the control of providers, as it is influenced by the prevalence of smoking among the patient population.

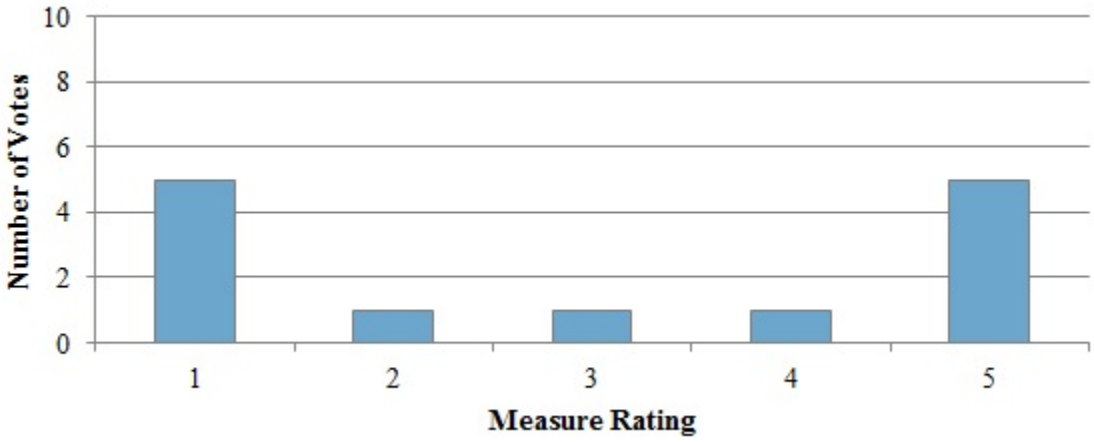
3.3 TEP Input on CAD Process Measures

CMS is considering three CAD process measures for public reporting via the Physician Compare website. **CAD-2** and **CAD-7** are among the top five quality measures by average TEP rating (see Table 3.1). **CAD-2** received the highest average score (4.23). **CAD-7** received the fifth highest average score (3.85). The remainder of this section describes the TEP feedback received on each measure in greater detail.

3.3.1 CAD-1: Anti-Platelet Therapy

CAD-1 received an average score of 3.00. Figure 3.9 below shows the distribution of the TEP's measure ratings for **CAD-1**; 6 TEP members assigned a score of 4 *or greater*, 1 TEP members assigned a score of 3, and 6 TEP members assigned a score *less than* 3.

Figure 3.9: Distribution of TEP Ratings for CAD-1

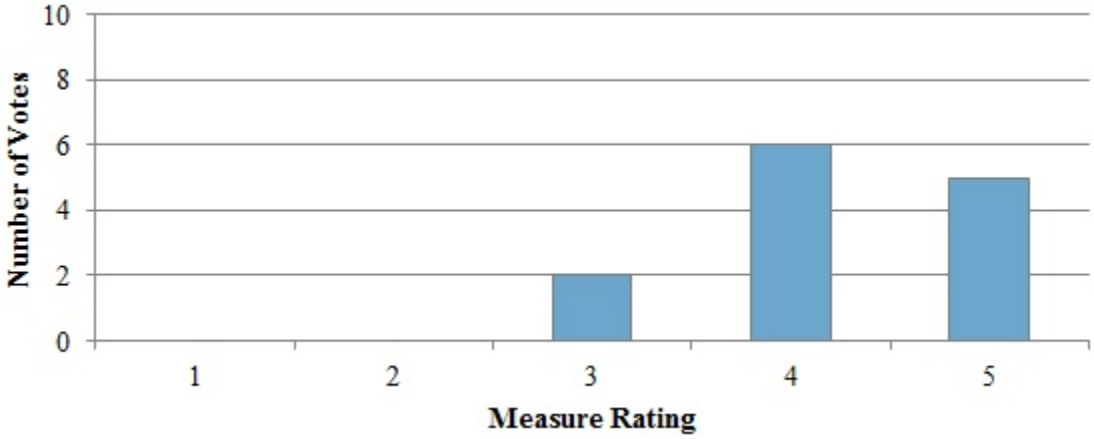


Of those who assigned a score of 3 *or less*, several TEP members cited the fact that **CAD-1** is no longer collected as part of PQRS after 2012. Publicly reporting this measure for program year 2012 but not for subsequent years may be confusing to health consumers; several TEP members agreed there is value in maintaining a more consistent measure set for public reporting over time. One TEP member, who assigned a score of 1, commented that the existing measure CAD Symptom and Activity Assessment Management (PQRS #196) would be an appropriate substitute for this measure.

3.3.2 CAD-2: Cholesterol Control or Plan of Care Including Statin

CAD-2 received an average score of 4.23. Figure 3.10 below shows the distribution of the TEP's measure ratings for **CAD-2**; 11 TEP members assigned a score of 4 *or greater*, 2 TEP members assigned a score of 3, and no TEP members assigned a score *less than* 3. **CAD-2** received strong endorsement by physicians on the TEP.

Figure 3.10: Distribution of TEP Ratings for CAD-2

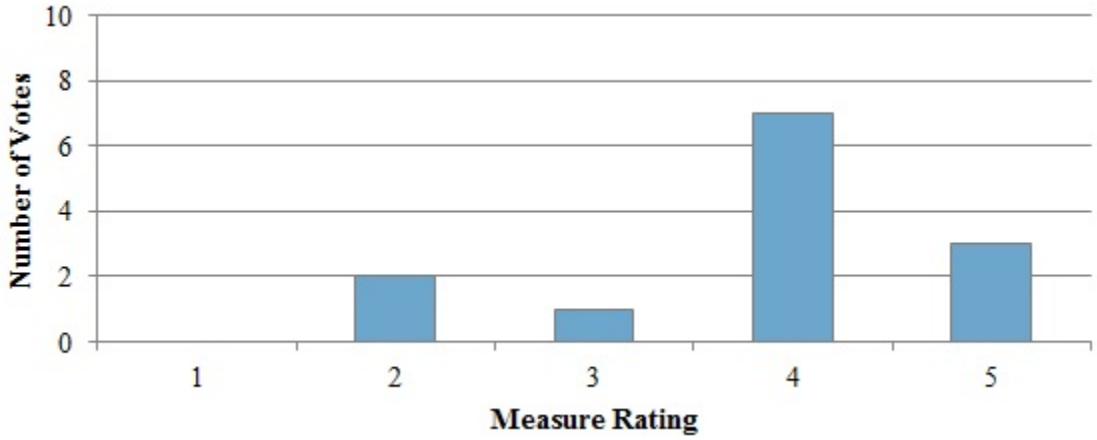


One TEP member, who assigned a score of 3, expressed concerns about the impact of the upper age range and patients with limited life expectancy on measure performance. Another TEP member, who also assigned a score of 3, suggested deleting the "documented plan of care" component of the measure.

3.3.3 CAD-7: ACE/ARB Therapy

CAD-7 received an average score of 3.85. Figure 3.11 below shows the distribution of the TEP's measure ratings for **CAD-7**; 10 TEP members assigned a score of 4 *or greater*, 1 TEP member assigned a score of 3, and 2 TEP members assigned a score *less than* 3. **CAD-7** received strong endorsement by physicians on the TEP.

Figure 3.11: Distribution of TEP Ratings for CAD-7



Of those who assigned a score of 3 *or less*, one TEP member commented that although **CAD-7** is clinically important, the measure relies on measurement of LVEF, which is not a readily reproducible measurement and often changes with therapy. Another TEP member stated that academic medical centers participating in the PQRS GPRO have shown difficulties in capturing **CAD-7**, since drug interactions/contraindications to prescribing ACE inhibitor or ARB therapy might not be reliably captured in the patient records.

4 GENERAL TEP COMMENTS

In addition to comments specific to the DM and CAD candidate measures described in the previous section, the TEP offered general comments that spanned several topic areas, including comments about the 2012 GPRO web interface measures, benchmarking, individual clinician-level reporting, etc. Table 4.1 summarizes these comments by topic area; the first column presents the topic area, and the second column provides the TEP comments that fall under each topic area. CMS will take the TEP's input into consideration to inform future Physician Compare measures selection activities.

Table 4.1: Summary of TEP Comments by Topic Area

Topic Area	TEP Comments
2012 GPRO Web Interface Measures	• A TEP member stated that the metrics presented to the TEP for indicating physician quality, measure outcomes on a population basis. While statistically sound and capable of showing differences between large group practices, these measures show value from a purchaser's perspective rather than from a consumer's perspective. • A TEP member commented that given that population outcomes are purchaser oriented value indicators, using them in a consumer context makes them, by definition, proxy measures from a consumer viewpoint. Proxy measures, arguably, are not the gold standard for consumer value. • A TEP member stated that the limited nature of the first two measure sets (i.e., DM and CAD) needs to be clearly communicated to health care consumers. Some public reporting web sites present this educational message by laying out the overall reporting framework that would eventually be filled in as the scope, number and quality of measures improve.[5]
Benchmarking	• A TEP member mentioned that raw performance results would be difficult to interpret and would disadvantage clinicians that serve sicker or more vulnerable populations. Rather than adjusting away differences in the performance results, it would be desirable to apply approaches that allow for fair comparisons across groups of physicians. The benchmarks and scoring could take into account performance and improvement over time. Benchmarks could be set separately for clinicians and groups practices serving high and low risk populations.
Securing Physician Buy-In	• A TEP member stated that publicly reporting quality measures on which the majority of group practices are likely to perform poorly may generate anxiety and skepticism and pushback from health care professionals. • A TEP member commented that if publicly reported measures are not case-mix adjusted, messaging efforts need to make it clear to group practices that analyses were performed which demonstrated case-mix adjustment had limited effects on group practices' performance rates.

[5] A TEP member mentioned that an example of this approach can be seen at GetBetterMaine, a web site maintained by the Maine Health Management Coalition: http://www.getbettermaine.org/.

Topic Area	TEP Comments
Individual Clinician-Level Reporting	• A TEP member commented that large sample sizes would be required to have sufficient reliability at the individual clinician-level. Large sample size requirements would in turn create a reporting burden for individual clinicians. • A TEP member commented that we would need to think about statistical issues that might arise in direct physician-to-physician comparisons that are unique to individual clinician-level reporting.
Composite Scores	• A TEP member commented that the best strategy for presenting disease-specific measures is to layer reporting with a summary score and then present the detailed measure components and stated that CMS should move to the use of composites and relative scoring as soon as possible.
Risk Adjustment	• A TEP member suggested including broad measures of prognosis (Charlson co-morbidity score) in the predictive model. • A TEP member recommended a report about risk adjustment to inform future measures selection.[6] • A TEP member commented that without risk adjustment, the DM outcome measures could be misleading to patients. • A TEP member stated that from the patient's lens, you may want to see if there's a variation – need to be able to identify physicians and health care professionals treating patients with sicker than average profiles so that these health care professionals can receive targeted assistance. • A TEP member opposed controlling for race/ethnicity, since risk adjusting away these factors sends the signal that the low performance rates are justifiable. • A TEP member stated an argument for adjusting for socioeconomic status was that "five-star" providers are needed in tough environments. Already there exists a whole class of patients facing health care access problems, and we don't want good providers getting closed in tough neighborhoods.
Physician Effects	• A TEP member asked about the impact of "physician effects" on these measures, and whether we'll get to a point where can look at these effects. VA has done research on physician-level versus group-level effects on patient outcomes. Another TEP member pointed out that publications from the past few years suggest there is a relatively low level of physician effect on the diabetes measures.

[6] See DM Shahian, X He, JP Jacobs, JS Rankin, ED Peterson, KF Welke, G Filardo, CM Shewan, and SM O'Brien. "Issues in quality measurement: target population, risk adjustment, and ratings." *The Annals of Thoracic Surgery* 96, no. 2 (2013): 718-26.

Topic Area	TEP Comments
Data Accuracy & Reliability	• A TEP member highlighted the need to be transparent with health care consumers about the reliability and accuracy of the data published on the web site. • Two TEP members found inaccurate information about themselves on the web site. A TEP member commented that if there is a third party source that has inaccuracies, CMS should find out where the third party gets the information from.
Hospital Affiliation	• It was commented that the hospital affiliation definition should be improved.
Process Measures	• A TEP member said that they have done tests and found that only a tiny percentage of consumers are interested in clinical process measures.
Education & Outreach	• A TEP member stated that there should be an educational component to train consumers to use the performance data published on Physician Compare.
Measure Display	• A TEP member stated that reporting the results using blocks seems to be the best option. Using the star ratings for Physician Compare may be confusing, since other star ratings provide comparative information about relative performance whereas the results to be published on Physician Compare are absolute data. • A TEP member stated that the 5 star graphic representation of quality is a consumer favorite, used in reviews of restaurants, hotels, goods, etc. In essence the star icons (i.e. 5 of 5 stars) are the gold standard of consumer satisfaction. • A TEP member supported using the bar (number of squares) with percentage numbers, rather than stars, because the stars have an emotional value to consumers which is inconsistent with proxy measures. • A TEP member stated that there may be value in displaying individual physician (and by extension physician group) quality measures with an indicator different from that used to rate institutional (hospital, health plan, ACO) quality. The value is that a different graphic communicates to consumers that choices of physicians are in fact different than choices of institutions. • A TEP member stated that the display should include a statement to clarify that the performance data are based on a patient sample. Lack of risk adjustment and benchmarks should also be mentioned. • A TEP member stated that it is important to note that performance on a set of condition-specific measures variably correlates with other condition measure sets. However, the total number of measures in the current set proposed for public reporting is quite modest.

REFERENCES

Shahian, DM, X He, JP Jacobs, JS Rankin, ED Peterson, KF Welke, G Filardo, CM Shewan, and SM O'Brien. "Issues in quality measurement: target population, risk adjustment, and ratings." *The Annals of Thoracic Surgery* 96, no. 2 (2013): 718-26.

John Adams, RAND. "Reliability of Provider Profiling: A Tutorial." Accessed October 16, 2013. http://www.rand.org/pubs/technical_reports/TR653.html.

APPENDIX A: QUALITY MEASURE/MEASURE DISPLAY SCORECARD

Name: _____

(1) For each quality measure, in the Rating column please score the measure according to how appropriate it is for public reporting on Physician Compare. (1=Not Appropriate 5=Very Appropriate)

(2) In the comments section, please note why the measure is suitable for public reporting. For measures rated less than 3, please describe changes to the measure specification, data collection, or proposed display that would make the measure more suitable for public reporting.

Outcome	Measure	Rating	Comments
Diabetes Mellitus	DM-3: Blood Pressure Control		
	DM-5: Cholesterol Control		
	DM-2: High Blood Sugar		
	DM-10: Low Blood Sugar		
	DM-7: Eye Exam*		
	DM-8: Foot Exam*		
	DM-11: Daily Aspirin Use		
	DM-12: Tobacco Non-Use		
Coronary Artery Disease	CAD-1: Antiplatelet Therapy*		
	CAD-2: Lipid Control		
	CAD-7: ACE Inhibitor or ARB Therapy		

*Data collected only in 2012 and not available for future years

(3) For the measures available for public reporting in 2014, please rate how appropriate each measure display is for reporting on Physician Compare. (1=Not Appropriate 5=Very Appropriate). Feel free to add comments as you see fit.

Display	Example	Rating	Comments
Percent	80%		
Stars			
Stars + %			
Pie Charts			
Blocks + %			

(4) If you have any additional comments, please write them here: