

PUBLIC COMMENT SUMMARY REPORT

Project Title:

Development of Inpatient Outcome Measures for the Merit-based Incentive Payment System

Dates:

The Call for Public Comment ran from November 29, 2018 to January 4, 2019.

The Public Comment Summary was created on February 1, 2019.

Project Overview:

The Centers for Medicare & Medicaid Services (CMS) has contracted with Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation (CORE) to adapt two claims-based hospital measures to assess the quality of care provided to Medicare beneficiaries by clinicians or groups who are eligible to participate under the Merit-based Incentive Payment System (MIPS). The contract name is Development, Reevaluation, and Implementation of Hospital Outcome/Efficiency Measures for Hospitals and Eligible Clinicians. The contract number is HHSM-500-2013-13018I, Task Order HHSM-500-T0001. As part of its measure development process, CORE has requested interested parties to submit comments on the candidate or concept measures that may be suitable for this project.

Project Objectives:

The primary goal of this project is to re-specify two hospital-level quality measures for potential future use in the MIPS. The two measures CORE is re-specifying are the:

1. Hospital-Wide All-Cause Unplanned Readmission Measure (hereafter “HWR measure”).
2. Risk-Standardized Complication Rate Following Elective Primary Total Hip Arthroplasty and/or Total Knee Arthroplasty (hereafter “THA/TKA complication measure”).

We consulted with clinical experts and a Technical Expert Panel (TEP) composed of multiple experts for input in the development of the two MIPS outcome measures. The re-specified measures will assess each clinician’s or group’s readmission or complication rate, respectively, relative to that of other MIPS participating clinicians or groups with similar patients. One measure, HWR, is already in use in the MIPS (referred to as the all-cause readmission or ACR measure #1789 in other communications); this is an updated re-specification. The quality measure scores will be calculated using patient characteristics and outcomes documented on routinely submitted Medicare administrative claims; therefore, the MIPS clinicians or groups whose performance will be assessed by the quality measures will not need to submit any additional data.

The primary goal of the comment period is to gather expert and stakeholder input to inform quality measure development for patients with a range of acute and/or chronic conditions, or patients undergoing elective procedures.

Information About the Comments Received:

Public comments were solicited through:

- Email notifications to CMS listserv groups;
- Email notifications to CORE listserv groups;
- Email notifications to the measure development TEP and THA/TKA Clinical Work Group;
- Measure specific listening sessions held during the public comment period;
- Presentation of the HWR Measure to clinicians, practice managers, and others to elicit feedback; and
- Web posts on the CMS Public Comment website.

CORE received the following number of responses on the two re-specified MIPS quality measures under development:

- 29 comments/comment letters received for the HWR measure;
- 10 comments/comment letters received for the THA/TKA complication measure; and
- 3 comments/comment letters that were out of scope.

Specifically, we received comments from 33 entities:

- 19 Individuals
- 2 Health systems
- 9 Professional associations
- 2 Hospital associations
- 1 Quality Improvement Organization

Additionally, we received one comment from the HWR Measure public comment listening session.

Stakeholder Comments—General and Measure-Specific

CORE received 34 comments on various aspects of the two MIPS quality measures under development. Please note that we are considering all suggested changes to the measure specifications summarized in this report. We have provided high-level summaries of comments and responses below, and the verbatim comments can be found in the attached Public Comment Verbatim Table document. Some of the comments received were general comments that applied to both measures under development. Others related to one or the other measure specifically. Most comments aligned with the specific questions posed with the public input materials for each measure:

HWR Measure

1. Does the measure identify the appropriate clinicians or groups responsible for 30-day unplanned readmissions following discharge from an acute care setting? Please explain your response as needed.

2. Do you agree with the recommendation to report this measure at the level of groups with at least 100 patients in this measure? Please explain your response as needed.
3. What, if any, additional validity testing would be meaningful for this measure?

THA/TKA Complication Measure

1. Does the measure identify the appropriate clinician or group responsible for complications following elective primary THA/TKA procedures?
2. What, if any, additional validity testing would be meaningful for this measure?

1. Summary of Comments Common to Both Measures

1.1 Comments on the general utility of the measures

Four commenters expressed support for the development of additional outcome measures to assess and improve the quality of care provided by MIPS clinicians and groups.

Response: Thank you for your support. Many stakeholders have contributed to the final specifications of these measures; we appreciate their contributions to making these measures useful and meaningful to clinicians and patients.

Four commenters felt the measures lacked sufficient data and empirical evidence to demonstrate that clinicians can meaningfully influence these outcomes. Further, the commenters expressed reservations about the measures' reliability and/or validity based on the information provided. Additionally, one commenter did not believe the burden of the measures should be placed completely on the clinicians, arguing a patient's role in influencing outcomes should be taken into consideration.

Response: Thank you for your thoughtful feedback. Constructing meaningful, reliable, valid provider quality measures is challenging and requires balancing competing factors and values. The measures we have presented address the wide variety of clinician, technical, and patient feedback we received during development of these measures. In the methodology reports, we provide evidence that these measures do capture reliable and valid quality signals at the clinician and group level. At the same time, we recognize stakeholder concerns raised during public comment, and consider them a critical part of ongoing measure refinement.

We provide additional responses to the specific concerns about measure reliability and validity in our responses to the measure-specific comments below. Future analytic and reevaluation work by the contractor will consider all stakeholder concerns, including those regarding the evidence base, and will continue to work with stakeholders and experts to ensure these measures are scientifically robust and meaningfully contribute to healthcare improvement.

1.2 Extent to which face validity results sufficiently validate the two measures

Two commenters expressed concerns over the limitation of face validity evaluation to TEP members, suggesting that TEP members could be biased by their participation in development of the measure and supported further face validity testing beyond the TEP and/or additional validity testing such as construct or predictive validity.

- One commenter also argued the face validity survey should, “specifically assess the degree to which the experts agreed that the measure attributed to each accountable unit resulted in scores that were valid and useful.”

Response: We thank the commenters for sharing their concerns. We support the idea of extending the assessment of face validity to other stakeholders. Survey-based validity testing is most meaningful when respondents have a thorough familiarity with the measure, and we agree that this would ideally include those who did not participate in the development; however, it is challenging to identify an appropriate set of stakeholders, one which is both representative and familiar with the details of the measure, to survey. We agree that assessment of predictive and construct validity could be a valuable addition to validity testing, although technical and conceptual challenges will need to be considered first. We will continue to examine ways in which to incorporate this feedback into the measure testing.

1.3 General comments on attribution

Two commenters felt that attributing readmissions to clinicians was not at an appropriate level to influence change. They specifically questioned Principle #3: Clinician Quality Reflects Hospital Quality of the methodology reports and whether physicians had the ability to drive change at a hospital.

Response: We appreciate these comments and understand that many factors may influence healthcare outcomes. This is true of hospitals as well, in that hospital performance on quality measures reflects, in part, the quality of clinicians and factors in the community outside the hospital. Yet just as hospital measures are not adjusted for clinician quality or community factors, we elected to measure clinicians without reference to the environment in which they practice. This makes the measures more useful to patients, who are concerned about what their potential experience will be, rather than a clinician’s outcomes in an ‘average’ hospital. This design was supported by both empiric findings as well as through TEP and expert input. One advantage of measuring both clinicians and hospitals on the same outcome (without accounting for the other entity’s influence) is that this allows for greater identification and understanding of successes and areas needing attention. The high performing clinician at a poor performing institution can then be acknowledged and can better drive care improvements within their institution. Likewise, a poor performing clinician at a high performing hospital can adopt practices that improve the care they provide.

1.4 Summary of comments on Risk adjustment

One commenter expressed support for the HKC model C-statistic of 0.65, noting that while not ideal, it is probably the best that can be achieved using claims data. However, **four commenters** were concerned the reported C-statistics for the risk adjustment models for both measures were too low.

- One of the commenters suggested the C-statistic for the THA/TKA measure could be improved by using more orthopedic-specific co-morbidities.

Response: We appreciate these comments and suggestions. It is worth noting there is a tradeoff between provider signal and C-statistics; the more the provider contributes to the outcome, the lower the C-statistic. For example, a C-statistic of 1 would indicate the outcome is perfectly determined by patient factors, and thus there is no provider quality signal. Therefore, any expectation of a minimal “acceptable” C-statistic can only be in the context of an expected provider signal; yet, we have little independent information about what that signal should be. At the other extreme, consider a “model” which predicts for each patient the observed average outcome for the provider. If the provider has a certain observed complication rate, we use this to predict the probability the patient has a complication. While most agree this crude rate is a fair estimate of provider performance, the corresponding model has a formal C-statistic of only 0.5. Thus, we think a C-statistic of 0.65 is acceptable.

In addition, it is important to keep in mind that all risk factors were chosen for consideration in consultation with clinicians, and that any factor which may be considered a complication of care is not included in the risk adjustment model. Inpatient procedures and discharge diagnoses that are not clearly present on admission would increase the C-statistic for the model but would also adjust provider scores for factors that likely represent lower quality.

We are committed to constant refinement and improvement of risk adjustment models used in all measures. The successive contractor will reevaluate this model and available risk factors on an ongoing basis, with the goal of producing the most accurate and fair risk adjustment models for assessing provider performance.

1.5 Summary of Comments on the Addition of Social Risk Factors to the Risk Models

8 commenters were concerned the models did not account for social risk. They argued the measures should adjust for social risk factors to avoid unfairly penalizing clinicians who treat vulnerable populations.

- One commenter recommended using the American Community Survey to inform the risk adjustment approach to both measures.
- Five commenters cited that differences in socioeconomic status (SES) and other social risk factors may lead to poorer outcomes unrelated to quality of care provided to the patient and recommended these differences be resolved through further testing and analysis.
- Two commenters noted that prior decisions by NQF regarding the hospital-level HWR measure to not require the inclusion of social risk factors should not imply that social risk factors not be reexamined and accounted for in a clinician-/ group-level measure.

Response: Thank you for your feedback and suggestions. We appreciate these comments and agree that social risk can be an important contributor to adverse patient outcomes. The challenges of adjusting for social risk factors include a limited number of valid, patient-level factors in the available data and the difficulty in separating the risk inherent in the patient from the performance of providers who treat large numbers of such patients.

We have included social risk factor testing in materials submitted to NQF, and will be guided by discussions with NQF on decisions to include social risk factors in the final models. In those analyses we examined two available social risk factors: dual eligibility (“DE”) status and residence in a zip code with a lowest quartile score on the Agency for Healthcare Research and

Quality (AHRQ) SES Index (“low AHRQ SES”). After adjusting both measures for each of these factors, we calculated final risk-adjusted (HWR) or risk-standardized (THA/TKA) rates for both clinicians and groups and found the correlations between scores adjusted for risk factors and those reported here ranged from 0.9972 to 0.9992, indicating extremely high agreement. This shows that adding these social risk factors will have minimal impact on the measure scores.

We are committed to constructing measures that are reliable and valid, and which account for factors beyond providers’ control. Ongoing research aims to identify valid patient-level social risk factors for testing and inclusion, and to construct complementary measures for highlighting disparities related to social risks. In addition, measure implementation can also potentially account for social risk through alternate means such as stratification, which Congress recently mandated for other CMS measurement programs.¹

2. Summary of Comments on the MIPS Risk-Standardized Complication Rate Following Elective Primary Total Hip Arthroplasty and/or Total Knee Arthroplasty Measure

2.1 Summary of Comments on the General Utility of the Measure

One commenter specifically supported the measure as potentially incentivizing quality for THA/TKA procedures.

Response: We appreciate this expression of support for the MIPS THA/TKA measure. We also recognize that of the ten commenters on this measure, five recommended alterations or shared concerns about specific aspects of the measure specifications but did not state concerns about the measure in general, or its usefulness in driving improvement in care.

2.2 Summary of Comments on Attribution

One commenter specifically mentioned that they felt the attribution algorithm as described would reasonably identify the surgeon responsible for the surgery.

Response: We appreciate this comment. Of the ten commenters on this measure, only two expressed concern about the attribution.

2.3 Summary of Comments on Reliability and Volume

Three commenters expressed concerns over the minimum volume threshold of 25 cases over three years for clinicians or groups.

- Two commenters were concerned that requiring a minimum of 25 cases excludes low volume providers. One commenter added that using this cutoff excludes more providers than it includes and suggested that a surveillance program be developed for low volume providers. Another commenter questioned if the volume cutoffs for all MIPS measures, including this one, was based more on convenience than on research.

Response: We appreciate the concern that not measuring low volume providers will fail to incentivize optimal care for all patients, as low surgical case volume might be associated with

poorer quality of care. While a minimum reporting volume of 25 cases does decrease the number of clinicians and groups captured, it does not significantly decrease the number of patients captured. Even with a 25-case minimum, 93% to 98% of all patients are retained in the measure (for clinicians and groups, respectively). Further, this measure assesses only Medicare Fee-for-Service (FFS) beneficiaries and does not include patients with other insurance types. Therefore, we cannot accurately determine whether clinicians are truly ‘low volume’ as they may have performed cases on non-Medicare patients or Medicare patients <65 years of age not captured in our data.

To be consistent with the hospital-level measure and to reflect that we cannot truly identify low volume clinicians, we tested a minimum reporting volume of 25 cases for clinicians and groups and found moderate to good reliability. Further, we did not adjust for volume in the measure calculation. However, CMS and the reevaluation contractor may continue to explore and evaluate the impact of different reporting thresholds.

2.4 Summary of Comments on Risk Adjustment

One commenter did not support the data smoothing and shrinkage techniques used for this measure to address issues with noise and variance in data. They noted it is difficult for both providers and patients to interpret the results in a meaningful way and would prove challenging for providers to use the data to inform changes to their clinical practice, and for patients to determine when quality care is being provided.

Response: Thank you for your comment. The measure uses a hierarchical model to adjust the clinician or group score for the reliability of the score relative to that of other clinicians. Smoothing of provider scores has been recognized as an important feature of provider profiling and was endorsed by the Committee of the Presidents of Statistical Societies, among others.^{2,3}

It is always the case that if a patient attributed to a provider has a better outcome, the provider has a better score. In this sense, there is no ambiguity about how to interpret the results, for either providers or patients. We recognize that both measures provide limited information about lower volume providers; however, this is a consequence of those providers having limited data to make inferences about, not the techniques used to construct scores for those providers.

2.5 Summary of Comments on Unintended Consequences

Two commenters expressed concerns about unintended consequences of the measure.

- One commenter was concerned that as surgeons move towards treating their healthier patients in outpatient settings, the patients in this measure will be sicker.
- One commenter thought that surgeons may “cherry pick” healthier patients which may provide fewer options for less healthy patients, or that some surgeons would choose not to perform these procedures which could adversely affect patients in rural areas with limited surgeons.

Response: We thank the commenters for sharing these concerns. We are attentive to the possibilities for unintended adverse consequences for all CMS measures, not only in development but also through surveillance of practice patterns after implementation. One goal

of risk adjustment is to account for and avoid encouragement of any potential ‘cherry picking’ by adjusting for patient factors that vary across providers, thereby ‘leveling the field’. Of greater concern is that some patients will not be able to obtain treatment because of their perceived high risk, especially as this is an elective procedure. CMS, through its reevaluation contractor, will monitor measure use, including analyses such as evaluating changes in patient mix over time, as part of its surveillance efforts.

3. Summary of Comments on the MIPS Hospital-Wide All-Cause Unplanned Readmission Measure

3.1 Summary of Comments on the General Utility of the Measure

One commenter specifically supported the measure.

Response: We appreciate this support of the MIPS HWR measure. Of the 29 comments on this measure, many included recommendations or concerns about specific aspects of the measure but did not state concerns about the measure in general, or its usefulness in driving improvement in care.

3.2 Summary of Comments on Attribution

Five commenters agreed with the attribution approach to multiple providers.

- There was support for spreading attribution to include inpatient clinicians, and for including clinicians in the perioperative period.
- One commenter thought shared attribution would apply pressure/incentive for improved care coordination and communication.
- There was support for the idea that a multi-attribution approach encourages team-based care and prioritizes handoffs between providers during hospitalization and at discharge.

Three commenters were concerned that multiple attribution was inappropriate or not evidence based.

- Two commenters were concerned multiple attribution dilutes responsibility, while one thought there was no evidence to support attribution to multiple clinicians.

Seven commenters disagreed with attribution to Outpatient Primary Care Providers.

- Six commenters thought that outpatient providers could not influence the outcome; that there was no established relationship between outpatient care and readmissions.
- One commenter thought the outpatient attribution should not prioritize primary care physicians but rather should treat all specialties the same.

Two commenters were concerned about attributing patients to the discharging clinician.

- One commenter thought it might have unintended consequences, with on-call clinicians being reluctant to discharge patients with whom they had little contact, while the other commenter thought it would be unclear or inappropriate for hospitalist groups.

Response: We thank the many commenters for their thoughts and suggestions regarding attribution. The TEP felt very strongly that it was most appropriate to attribute readmissions to multiple clinicians to encourage coordination and shared accountability. Though it may be possible to determine empirically which clinicians have the greatest influence on readmissions, we have developed an attribution algorithm based on a conceptual model of which providers should, over time, have the most influence on the outcome. This approach is more transparent to clinicians and more likely to drive higher quality care, in that attributed clinicians are those who should theoretically work together on the transition of the patient from the hospital.

Attribution to an outpatient PCP is consistent with the existing MIPS hospital wide readmission (“ACR”) measure, which currently only attributes the outcome to the outpatient PCP. Once a patient has been discharged, the outpatient PCP is conceptually the clinician most likely to have contact with the patient. CMS is exploring measures designed to assess outpatient provider performance and will continue to consider how these measures complement each other. One potential option discussed with the TEP, which they deferred in favor of the current attribution approach, was to not attribute readmissions in the HWR measure to the outpatient PCP and instead focus on capturing readmissions through measures designed to attribute admission rates (which include some readmissions) to outpatient providers. CMS will continue to consider all approaches and input as it regularly reevaluates its measures and measurement programs.

3.3 Summary of Comments on Reliability and Volume

Two commenters were concerned about the reliability of the measure.

- Both commenters suggested additional reliability and validity testing and recommended a threshold of at least 0.7 for reliability. Further, they argued that if the measure has low reliability, it cannot have high validity.
- One commenter was also concerned the measure had low levels of signal-to-noise reliability at the minimum and maximum values.

3 commenters were concerned about the exclusion of low volume providers.

- One commenter argued that statistical approaches applied to low volume providers need independent review before being applied to these types of measures and was concerned that excluding groups and clinicians based on volume dilutes the results and can possibly exclude more providers than it includes.
- Two commenters urged CMS to consider adopting a threshold similar to the existing ACR measure that includes only to groups of 15 or more clinicians who have 200 or more readmissions.
- Two commenters were concerned about the number of providers excluded at the 100-patient threshold.
- One commenter noted that attributing this measure at the clinician level would result in small sample sizes subject to large swings in performance and low levels of reliability and validity, whereas applying the measure at the Taxpayer Identification Number (TIN) level would result in a larger patient population which would ensure higher reliability, encourage team-based care, and support our principles to align this measure closely with the hospital level measure.

Response: We thank the commenters for these remarks and suggestions. Because there are many ways of measuring reliability, there is no single threshold value representing ‘adequate’ reliability. Test-retest reliability (or ‘reliability over time’) as measured by intra-class correlation (ICC) [2,1] is typically much lower than signal-to-noise reliability because it is measuring agreement between nominally independent quantities. In the literature, test-retest reliabilities of subjective diagnoses or chart abstraction are below 0.5. Thus, we think 0.4 is a reasonable threshold for test-retest reliability. In contrast, signal-to-noise reliability measures the ratio of variation between providers to total variation, and values over 0.9 are not uncommon for provider measures.

The proposed reporting threshold of 100 patients for a group was based on initial reliability analyses. CMS has yet to determine the final reporting approach.

The exclusion of low volume providers has been a long-standing concern. For both hospitals and MIPS clinicians, it is methodologically challenging to construct a reliable, valid measurement which reflects true provider outcomes. NQF has currently convened a workgroup to address this challenge; we will explore any methods they endorse for measuring small volume providers. Currently, however, there are few alternatives to excluding them from reporting.

3.4 Summary of Comments on Risk Adjustment

One commenter thought in addition to adjusting for case-mix, CMS should consider accounting for the total number of conditions each patient has, which has been proven to impact outcomes.

Response: We appreciate this suggestion. For the MIPS HWR model we found that adding the number of comorbidities as a risk adjuster did not materially improve the model discrimination; of the five cohort models, the biggest change was for the cardiovascular cohort, where the C-statistic decreased from 0.6639 without this factor, to 0.6637 with the factor; the C-statistic declined for 3 of the 5 cohorts, and increased by 0.0001 for the other 2.

3.5 Other Concerns

One commenter recommended a shorter readmission window be applied to this measure.

- The commenter suggested a readmission window of 7 days would better reflect between-hospital variation in performance and may more accurately capture the impact of discharge care coordination and follow up, where as a longer window allows for confounding factors, such as patient’s social and community impacts, which are beyond the providers’ or hospitals’ control. A shorter window would provide the most effective means to detect true variation, providing more opportunity to develop effective solutions to reduce preventable readmissions.

Response: Thank you for your comments. The MIPS HWR measure is designed to align with the hospital HWR and all of CMS’s publicly reported 30-day readmission measures. While 30-days is an arbitrary cut off, there is increasing evidence supporting a range of interventions, both hospital- and provider-based, that demonstrate the reduction of 30-day readmission rates. For this reason, the outcome definition was not reevaluated in the re-specification of the measure.

One commenter was concerned that if the same physician plays 2 to 3 of the roles, the readmission would be counted 2 to 3 times against that physician.

Response: To confirm, each patient is counted, at most, only once for any physician. Please see the Methodology Report for more details on attribution.

Two commenters suggested changes to the outcome.

- One commenter noted that some planned readmissions for epilepsy surgery or intracerebral electrodes are not included in the list of planned readmissions that would be excluded from this measure.
- Another commenter noted that patients with medically refractory epilepsy readmitted for planned epilepsy surgery after a recent (30 days or less) pre-surgical evaluation in the hospital are not included in the list of planned readmissions that would be excluded from this measure.

Response: We thank the commenters for their detailed input. The MIPS HWR measure uses the Planned Admission Algorithm that was developed independently and applied across all CMS readmission measures. We will ensure this input is incorporated into the annual reevaluation of Planned Admission Algorithm.

Overall Analysis of the Comments and Preliminary Recommendations

We appreciate all the comments submitted on the two outcome measures re-specified for the MIPS. While the above document provides basic topic summaries, all comments are being considered in their entirety. For both measures, our goal was to accurately and consistently capture outcomes important to patients, ensure visibility to patients, providers, and policy makers, and incentivize care improvement. We value all concerns and will take all suggestions under consideration. We especially value the comments regarding small volumes and social risk factors. Our objective is to create useful measures that will minimize any unintended consequences and hope any changes to the measures will make them more valuable to stakeholders. The measures have been submitted to the NQF for endorsement, but this does not preclude additional changes prior to implementation.

References:

1. Office of the Assistant Secretary for Planning and Evaluation (ASPE). Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. 2016; <https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs>. Accessed September 1, 2018.
2. Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. *Circulation*. 2006;113(3):456-462.
3. Ash AS, Fienberg SF, Louis TA, Normand S-LT, Stukel TA, Utts J. Statistical issues in assessing hospital performance. 2012.