Conference Proceedings

RAND Medicare Advantage (MA) and Part D Contract Star Ratings Technical Expert Panel May 31st 2018 Meeting

MEETING SUMMARY

Cheryl L. Damberg and Susan M. Paddock

For more information on this publication, visit www.rand.org/t/CF391

Published by the RAND Corporation, Santa Monica, Calif.

RAND Technical Expert Panel
Medicare Advantage (MA) and Part D Contract Star Ratings

May 31st, 2018

MEETING ATTENDEES

Technical Expert Panel Members

NAME	POSITION
Arlene Ash, PhD	Professor and Division Chief for Biostatistics and Health Services Research in the Department of Quantitative Health Services, University of Massachusetts Medical School
Liza Assatourians, JD	Vice President of Federal Programs, America's Health Insurance Plans (AHIP)
Anne Burns, RPh	Vice President, Professional Affairs at the American Pharmacists Association (APhA)
Lindsey Copeland, JD	Policy Director, the Medicare Rights Center
Jennifer Eames Huff, MPH	Independent consultant; Senior Advisor, Pacific Business Group on Health
Eve Kerr, MD, MPH	Louis Newburgh Research Professor of Internal Medicine, University of Michigan Medical School, Director of the Ann Arbor VA Center for Clinical Management Research, and Director of the Michigan Program on Value Enhancement
Elisa Munthali, MPH	Senior Vice President of Quality Measurement, National Quality Forum
Amy Nguyen Howell, MD, MBA	Chief Medical Officer, America's Physician Groups (APG)
Deborah Paone, DrPH, MHSA	Performance Evaluation Lead for Quality Measurement, Social Determinants of Health, and Care Innovation, Special Needs Plans (SNP) Alliance
Ninez Ponce, MMP, PhD	Professor, University of California Los Angeles Fielding School of Public Health's Department of Health Policy and Management, Director of the Center for Global and Immigrant Health
Patrick Romano, MD, MPH	Professor of Medicine and Pediatrics, University of California Davis School of Medicine
Marissa Schlaifer, MS, RPh	Consultant, Pharmaceutical Care Management Association
Allyson Schwartz, MSS	President and CEO, the Better Medicare Alliance
Nadine Shehab, PharmD, MPH	Senior Scientist, Medication Safety Program in the Division of Healthcare Quality Promotion, Centers for Disease Control and Prevention
Jane Sung, JD	Senior Strategic Policy Advisor, AARP Public Policy Institute
Dolores Yanagihara, MPH	Vice President of Analytics & Performance Information, Integrated Healthcare Association (IHA)

RAND Staff

Cheryl Damberg, PhD (PI/Project Director)
Susan Paddock, PhD (co-PI/Project Director)
Marc Elliott, PhD
Melony Sorbero, PhD
Justin Timbie, PhD
Rachel Reid, MD
Jessica Phillips, MS
Emily-Kate Chiusano

CMS Observers

Elizabeth Goldstein, PhD
Sarah Gaillot, PhD

RAND CORPORATION

MEETING SUMMARY

Welcome and Introductions

- The RAND Project Director, Cheryl Damberg, began the meeting by welcoming attendees and asking Technical Expert Panel (TEP) members to introduce themselves.
- Cheryl Damberg briefly described the purpose of forming a TEP and outlined the goals for the meeting.
    - The RAND Corporation is under contract to the Centers for Medicare & Medicaid Services (CMS) to conduct analyses to inform improvements to the Medicare Star Ratings for Medicare Advantage (MA) and free-standing prescription drug plans (PDPs).
    - RAND determined that it would benefit from obtaining input from a diverse group of individuals who have expertise in quality measurement, risk adjustment, the delivery of health care, and those who bring the perspectives of Medicare beneficiaries.
    - The purpose of the RAND Star Rating TEP is to provide input to RAND on various components of the rating system and analyses that RAND might consider performing to inform policy considerations. The TEP will meet twice per year for the duration of the contract, which ends 8/31/2019.
    - RAND stated the TEP does not replace existing mechanisms to obtain broad stakeholder input regarding enhancements to the Star Ratings program. CMS will continue to solicit stakeholder feedback on the Star Ratings program through the annual Call Letter and regulation processes (i.e., Notice of Proposed Rule Making).
- For the first TEP meeting, RAND sought input on analyses it should consider regarding three topics: 1) measure thresholds used to assign Star Ratings, 2) the utility and feasibility of constructing and reporting Star Ratings at the level of Plan Benefit Package (PBP) or geographic area, and 3) the measures included in the calculation of Star Ratings.

TEP Discussion

- The TEP's discussion is summarized below, including clarifying questions, comments about various issues being considered, and suggestions for analyses. The summary is organized by the main discussion topics of the meeting, which map to the content of the PowerPoint slide deck used to guide the discussion:
    - Overview of the MA and Part D Star Ratings program and how Star Ratings are computed
    - Measure Thresholds Used to Assign Stars
    - Utility and Feasibility of Constructing and Reporting Star Ratings at the Level of Plan Benefit Package (PBP) or Geographic area
    - Measures

Overview of the MA and PDP Star Ratings Program

- RAND co-Project Director, Susan Paddock, presented an overview on MA and Part D Star Ratings which included the goals of the Star Ratings program, key aspects of the

Star Ratings system, and the Call Letter and regulation processes. *The presentation can be found in the PowerPoint slide deck.*

- TEP members asked questions to clarify their understanding of the program and commented on the structure of the program, as follows:
    - ***Could you explain the Categorical Adjustment Index (CAI), what's the magnitude of that adjustment, and roughly what proportion of contracts receive an adjustment?*** All contracts receive some adjustment based on the contract's percentage of Low Income Subsidy (LIS)/Dual Eligible (DE) and disabled beneficiaries. Contracts with relatively low percentages of disabled and/or dual beneficiaries might receive a decrease while other contracts receive an increase. The magnitude of the adjustment varies by year. It is applied to unrounded stars. A small portion of the contracts (< 5%) receive a bump to another rating category when the CAI is applied; for most contracts, the CAI does not change their Star Rating. While the bump can happen anywhere over the range of performance, usually it happens in the 3.5 to 4.0 and 4.0 to 4.5 star range. The CAI does not move contracts up a full star or greater, so you don't see a contract shifting from a 3.0 Star Rating to a 5.0 Star Rating. The bump occurs among contracts with a large proportion of LIS/DE beneficiaries and which are close to a Star Rating category threshold. The CAI is similar to what you would get if you were to directly apply case-mix adjustment in a regression model to the measure scores. RAND builds the CAI based on running case-mix adjustment models. The adjustment is based on the standard way of calculating case-mix adjustment, which is looking at the mean within-contract difference associated with the characteristic (e.g., dual status). Only 9 measures are currently adjusted to determine the CAI values; CMS re-evaluates the set of measures included for adjustment each year by examining the LIS/DE and disabled disparities to select which measures will factor into the CAI. Social risk factors will be a discussion topic at the fall 2018 TEP meeting.
    - ***Regarding the clustering methodology –thresholds are made by the performance distribution for each measure but not necessarily based on what is a clinically meaningful threshold. Would CMS consider not having any plans fall into the 1-star category?*** RAND is interested in hearing the TEP's thoughts on how to set thresholds, including whether consideration should be given to setting substantive thresholds versus data-driven thresholds, as under the current approach used in Star Ratings.
    - ***Why are stars assigned at such a low level (i.e., at the measure level) instead of aggregating the measure scores? At the measure level, CMS computes a Star Rating, then averages the measure stars. Assigning stars at the lower level results in information loss.*** In terms of rolling up individual measure stars, this structure pre-dates RAND's involvement in the Star Ratings program. Most measures are reported on a 0-100 scale, such that averaging those scores makes sense. However, there are some measures that are not reported on the 0-100 scale. There's always some loss of information when you discretize an average. However, there are two motivations for this process--scaling and equal impact. Most measures are on the 0-100 scale, but the fact that they are on that scale does not mean

they have equal impact. Some measures use more of that range than others and would have a much greater weight. Some of the original motivation was to explicitly assign importance weights to each measure (e.g., 1.0, 1.5, 3.0) rather than have weights be determined by scaling—for example, outcome measures are assigned a weight of 3.0 while process measures are assigned a weight of 1.0. The RAND team agreed there are some tradeoffs.

- ***What constitutes improvement and how is it computed?*** The improvement measure is capturing whether, for a given contract, there is a significant change in performance on measures between the current year and prior year. Improvement is computed by counting the number of measures that had a significant year-over-year increase, the number of measures that had a significant decrease, and the number of measures that remained the same (i.e., no significant change). The improvement measure is currently weighted 5.0. There are hold harmless provisions for contracts with high ratings because they have less room for improvement. The Technical Notes for the Star Rating program have an appendix that details the methodology for calculating the improvement measure. See: https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.html
- ***At what point does the CAI come into the Star Rating calculation?*** Once the measure stars are computed, the stars are aggregated, taking a weighted average of the measures. A reward factor is then added to the overall score for contracts with higher or consistent performance over time. The final step is adding the CAI. The CAI is applied at the overall and/or Summary Star Ratings.
- ***It might be helpful for TEP members to know about measure specifications and exclusions, and to consider the data sources for measures***. A TEP member expressed concern about several measures; they commented that the Health Outcomes Survey measures (HOS) are self-reported and highly weighted in the calculation. They believe the methodology for the HOS is problematic for persons with high social health determinants.
- ***There is an assumption that the Star Rating measures are quality measures; however, a TEP member stated there are a subset of measures that focus on compliance with regulations, audits that are outside assessment of clinical quality. These are talked about as "quality measures" and a contract's score can be impacted by whether the contract had compliance issues. Does considering what measures should be included to measure contract quality fall within the scope of RAND's work?*** RAND commented that historically there has been a measure in the Star Ratings program that captures compliance issues (related to beneficiary access and performance problems). The 2018 Star Ratings year was the last year for that measure. During the 3rd agenda topic (measures), RAND sought the TEP's input on different measures and this is an opportunity to discuss whether compliance-type measures should be in the Star Ratings program. CMS commented it has received very conflicting comments on these types of measures, with plans commenting that compliance-type measures should

not be in the Star Ratings program and beneficiary advocacy groups and others saying these types of measures should be included in the program.

- ***There was a general observation that the rating system is complex regarding how all the different pieces are put together and adjusted. What is the conceptual rationale for the reward factor?*** The reward factor provides further recognition to contracts with relatively high performance as well as consistent performance (low variation). When the reward factor was initially developed, this was based on the interest of the industry. The reward factor was introduced prior to the CAI. The intent is to give a small bump-up for consistently high performance.
- ***Does this mean if there were 2 contracts with an average of 4 stars across measures, 1 contract with all 4 stars and the other with 3 stars and 5 stars, the contract with straight 4 stars would receive the reward?*** It is more likely that the contract with consistent 4 stars would receive the reward. The reward factor incentivizes contracts to bring up lagging performance areas.
- ***Historically has there been clustering around 4 and 5 stars? If so, has RAND or CMS looked at the clinical meaningfulness of those measures when calculating Star Ratings?*** RAND stated that not many contracts receive 5 stars (n~15); 4 stars and above represents ~40% of contracts and ~68% of enrollees. A large fraction of Medicare enrollees are enrolled in higher performing contracts. Not many contracts receive less than 3 stars, but each year there are some that do.
- ***Will the TEP have input on re-structuring of the Star Ratings methodology?*** The focus of the RAND TEP is to help RAND consider analyses around the different components of the system. For example, thresholds and social risk factors. The focus is on thinking about improvements to the scoring system, potentially ways to simplify the methodology and to ensure the scoring system is meeting the objectives set forth by CMS.
- ***If the goal is to improve quality, what is CMS' goal for the 5 star category? Is the goal to keep the number of contracts receiving 5 Stars limited and not allowing more plans to be 5 star plans? Or is it to have more contracts?*** CMS commented that it isn't a goal of the Star Rating program to try to limit the number of 4- and 5-star plans. CMS would be pleased if plans were moving in the 5 star direction because it is better for beneficiaries. There are some payment considerations because 4 and 5 star plans get Quality Bonus Payments, 3.5 star and above get rebates.
- ***Regarding measures, there are other clinical factors that affect whether you should perform recommended care for certain beneficiaries (e.g., not recommending mammography for people who are in palliative care, or not asking a patient with dementia about patient experience).*** This TEP will not be dealing with measure specifications and measure exclusions, but it is fine if TEP members want to flag problems with measures. Refining measure specifications is an issue for measure developers. RAND clarified that the CAHPS survey allows for proxy respondents from caregivers.

Measure Thresholds Used to Assign Stars

- Susan Paddock provided background information on measure thresholds, trends in performance for different measures, and reviewed potential enhancements to the threshold methodology, including approaches that would provide partial or full advance notice of thresholds and that would increase the stability of thresholds, such as through resampling, a process that involves randomly splitting the contract measure scores into subsets, estimating thresholds on each subset, and averaging the resulting thresholds to obtain a final set of thresholds. *The threshold presentation can be found in the corresponding slide deck.* RAND stated that CMS is revisiting how thresholds are used to assign measure stars for the clustering approach, which is applied to the majority of Star Ratings measures, and that RAND has been working with CMS on exploring potential ways the approach to setting thresholds might be revised. There is interest in focusing on changes in the methodology that provide greater stability of thresholds over time as well as providing advance notice to contracts about what the thresholds will be. The question is how advance notice might be provided. One approach might be to use prior Star Ratings data and a question, of course, is how should the data be used to develop thresholds to provide advance notice? Might thresholds be predicted using prior data or are there ways to reasonably limit threshold changes over time, like with a threshold or a guardrail? And which years should be used to set thresholds?
- **Summation of Threshold Discussion:** After reviewing the information contained in the slide deck related to improvements in measures over time and time lags involved with approaches that would predict thresholds using prior years' data, the TEP expressed strong support for exploring options for improving the stability of thresholds, through resampling, addressing outliers, and/or setting guardrails. TEP members did not support the advance notice approaches due to the long time lag between data used to set thresholds and the actual measurement and rating year (i.e., 2016 data used to set thresholds for 2021 Star Ratings). The TEP felt that if year-to-year stability in thresholds could occur, then there would not be a need for advance notice and incentives for improvement would be retained.
- The bullet points below summarize questions raised by the TEP and comments on RAND's options for addressing large year-over-year changes in thresholds:
    - ***Do the same number of contracts receive low Star Ratings each year?*** The number of contracts that receive low Star Ratings varies from year to year.
    - ***The 2 star thresholds look to be less stable over time. Do you have a sense for why 2 star thresholds have been less stable?*** There tend to be fewer contracts in the 1 and 5 star categories, especially the 1 star category, because all the plans are trying to have very good performance, and so the scores do tend to be clustered more on the higher end. Performance at the low end is less consistent than performance at the high end of the distribution. Another way to say this is, we think that the performance itself is less consistent at the low end than it is at the high end. At the high end, contracts are homing in on fairly consistent levels of performance in

contrast to the low end where there really is a lot of variability in the actual lowest performance each year.

- ***What is the denominator minimum for HOS measures?*** RAND stated that information can be found in the Technical Notes about denominator sizes required for each measure. TEP members thought that small denominators might be causing the swings in performance. Even though some contracts have a smaller number of cases than is allowed to be scored, the minimum sample sizes are enforced for scoring. For example, CMS adheres to the denominator requirements for scoring the measures. One TEP member commented that with small numbers of respondents on the HOS survey, 1 or 2 people might be influencing the results on the Improving or Maintaining Physical and Mental Health measures, which are used as standalone measures (weight of 3.0) and in the improvement measure. RAND clarified that the HOS measure already measures improvement and is not included in the improvement measure.
- ***To what extent is there real-world evidence that contracts plan for thresholds? Does knowing what a threshold is going to be in advance result in contracts stopping their quality improvement work once the threshold is met?***
    - TEP member comments included:
        - Having a sense on where you, as a contract, stand and where your deficit areas are is helpful in securing resources from plan leadership to close the gaps.
        - Knowing thresholds in advance helps contracts identify which measures they need to focus on; knowing the thresholds may change the contract's priorities of where to direct resources.
        - Contracts constantly have to work on medication adherence. If a contract knows they are not performing well, that information impacts where they go for outside help. Knowing the threshold does not mean you stop working on that measure.
        - Contracts pay a lot of attention to thresholds. Thresholds are dependent on how the industry as a whole is performing so, in effect, a contract can't stop working to improve. It might be better to go with the "guardrails" approach and give some predictability so contracts can continue to work on improvement with reassurance there won't be big swings. Contracts do plan more than a year in advance and huge swings in thresholds are detrimental.
        - It is important for contracts and their network providers and partners to get advance notice of the thresholds, the "goal lines", so they can work on needed improvements. It's a very critical issue right now, not knowing what the thresholds are going to be.
        - Stability is a big deal for plans as well as for providers. Providers are under a lot of pressure right now around a lot

of measurement (e.g., MACRA, ACOs). Hopefully, the TEP can also discuss alignment of measurement across programs.
        - Plans pay a lot of attention to improving and getting where they have to get to, to achieve the highest score possible. Because all the scores are based on relative performance, a plan must keep up. A plan can't ignore measures and say that they don't matter anymore.
        - There are some concerns about using old data and whether lagged performance thresholds will stall some of the progress for improvement. It may be better to look at the alternative, which is guardrails, and to talk about deviation and how much change happened. The guardrails would give some predictability so that plans continue to work on improvement and have some assurance that there won't be incredible swings one year to the next, because it affects how they plan.
    - One TEP member indicated that they previously wanted advance notice of thresholds, but after better understanding how this would work they have concerns. Specifically, they are concerned about using old data and whether it will stall progress for improvement, and they concluded a guardrail type approach would be better and give plans some predictability.
- ***CMS has asked RAND to think about the stability of thresholds under this clustering approach. Clustering can be sensitive to outliers.***
    - Comments from the TEP were as follows:
        - Has RAND considered using multiple year averages to set thresholds? A weighted average over two or three years could provide some stability.
        - The last row in the data timeline (per the PowerPoint presentation) showing which year of data would be used to set thresholds would allow CMS to do some prediction of thresholds that would alleviate some of the outlier effects. Per the data shown on the graphs, RAND would need several years of data from earlier years to make the predictions to provide advance notice. In response to a question from the TEP about feasibility of accelerating the data submission schedule, RAND noted that the current capture and submission of data would require reaching back to 2016 for generating thresholds for the 2021 ratings (to allow contracts to know thresholds prior to the start of 2020).
            - Given the lags in changes to thresholds over time, it is possible that more, if not all, contracts would move into the 4 and 5 star categories.
        - There was a comment about whether advance notice would provide an advantage to higher performing plans because they have more resources to dedicate to meeting those thresholds. Could it create a larger gap between high and low performing contracts?
    - ***RAND noted a need to consider the feasibility of any given prediction approach and the need to tailor it across different measures given their***

different improvement trajectories and the factors affecting improvement. One TEP member said that CMS could do predictions of where thresholds could be to alleviate outlier effects, and that several years of data would be needed to sort this out analytically due to rapid rise in performance and then leveling off. The implication is that if thresholds are set ahead, plans could get to thresholds and all plans could be 4 or 5 stars. When considering the scoring, CMS and the plans need to feel confident the measures reflect true quality. It may be there are important areas missing. It is important to consider the measures that are included and whether they the right measures. RAND commented that there is a tension between complexity and transparency in terms of how the thresholds could be set using prediction modeling approaches.

- ***From the perspective of the beneficiary, one TEP member noted there is a tension between rewarding contracts for improvement and the purpose of communicating information to beneficiaries and distinguishing performance***. If all contracts receive a 4 or higher Star Ratings, this would signal to beneficiaries that all contracts are doing well. CMS should not lose sight of identifying meaningful performance differences to help beneficiaries discern quality among contracts.
- ***Several TEP members commented that there seems to be a need to stabilize thresholds***. If CMS can stabilize thresholds, then the program could achieve predictability in threshold changes for contracts without needing to provide advance notice of thresholds. There was consensus that setting thresholds for 2021 using 2016 data to calculate thresholds did not make sense. Prioritization should be on creating more stability, particularly stabilizing lower end thresholds that are more outlier-sensitive.
- ***What is the real intent of the thresholds? Is it to drive what the plan chooses to weight in terms of their quality improvement efforts?*** It is important to consider what the plan's population health burden is for its beneficiaries? One TEP member commented that quality improvement efforts should be driven by clinical and epidemiological factors, and focus on clinically meaningful measures.
- ***Why does CMS take each measure and determine thresholds instead of determining pre-set thresholds?*** There was a comment about why CMS might want there to be more 5 star plans, since CMS shouldn't want to make distinctions without a difference. If the movement to improve has been successful and lots of plans are up towards the top end of performance, then we don't want to grade on the curve. If half the plans deserve a 5 Star Rating, then we shouldn't be trying to turn some of them into 4 star plans. To have achieved a 5 Star Rating, a contract would have to have achieved a certain level of accomplishment that is tied to the reality of what the contract has done rather than how the contract compares with its peers. It was noted that thresholds are the "original sin" and that if there was a smooth function, a contract would know that's the extent that it did better. A contract would be graded a little bit better and wouldn't be scurrying around like crazy to try to jump over a hurdle not being sure where the

hurdle is going to be. And even if it is known where the hurdle is, it's kind of crazy to scurry to jump over the hurdle.

- ***CMS does publish industry trends on thresholds and plans are using that information to plan their quality improvement areas of focus.*** Some of the thresholds are so close it's very difficult for a plan to tell where they will land. This is a universal ask – knowing thresholds in advance.
  - Several TEP members commented that partial advance notice is not helpful because the contracts are already so far along in the measurement year. You'd have to set thresholds early for people to plan.
  - One member commented that there are some dramatic spikes with Part D measures, and some of this is related to very small contracts. One suggestion was that CMS consider eliminating contracts with small denominators from the threshold calculations to stabilize thresholds.
  - RAND stated in its exploratory analyses that some of the measures that are EHR-enabled tend to have a distinct trend relative to others. That's basically indicative of structural changes in healthcare provision in terms of performance and what that might mean to these trends. For some measures, there are likely reasons why there have been large increase in scores year-over-year; some measures are about having an EHR system in place or not (e.g., BMI), which is a "check the box" measure that the EHR prompts for. Some of the measures seem topped out and shouldn't be in the Star Ratings rubric (e.g., they just differentiate between those plans with and without the "program"—meaning EHRs). CMS may need something different than 5 stars for these type of measures, perhaps fewer stars or take them out and make them display measures, and not force the 5-star scheme.
- ***How important is consistency across all measurement and rating programs at CMS? Could CMS give advance notice for this one program?***
  - CMS responded that, in general, CMS likes consistency across programs; however, there are different purposes for the measurement programs. Given the link to the Quality Bonus Payments, the MA Star Ratings program looks different than other programs. CMS does not receive a lot of questions from hospitals regarding thresholds compared to the MA Star Ratings side. There is a large difference in the amount of performance-based payments between the CMS programs.
  - One member commented it takes a couple of years for a plan to reach a threshold of performance on a new measure. It takes time for providers to pay attention and to do something about the measure. This needs to be taken into account when looking at past years of data. Would CMS be going back and holding plans accountable to something old if using prediction approaches with lagged data?
- ***Variation across measures in trends in performance.***

- RAND stated that that there are improvements in net performance for most measures and that the pattern of improvement varies across the measures. There is decreasing variation in the measure scores over time.
- There are some distributions and measure scores that are quite stable over time. Not everything is rapidly changing in all sorts of directions. The screening measures have fairly stable distributions as well as the HOS measures for improving/maintaining physical health and mental health. The HEDIS diabetes care and arthritis measures are quite stable; reducing the risk of falling, call center, appeals upheld, a measure of whether members are choosing to leave the plan, as well as medication adherence measures are stable.
- RAND also noted the mean performance is improving over time for most measures and that improvements tend to be greater on the lower end versus the higher end.

- ***It would be of interest to beneficiaries to compare similar populations, such as special needs plans, so that the comparison is meaningful for "people like me***." CMS should consider displaying stratified scores based on population subgroups.
- ***How should RAND and CMS consider structuring "guardrails"?***
  - One member of the TEP commented that some measures are topping out at 90% or 50% performance. Is that a hard stop or not? The performance rate may be signaling something about that clinical content area and patient preferences for treatment. Should we expect to get to 100% performance on every measure? Some of the recommended care embodied in the measures is preference sensitive and forcing performance higher may lead to over-treatment. This type of exclusion isn't found in the measure specifications. For example, colorectal cancer screening is topping off at 70%; to suggest that we're going to keep going up is folly and potentially harmful. Maximum thresholds should be set based on something from literature (e.g., in clinical trial only X% can get there under optimal circumstances) to reduce the risk of over treating patients. CMS should examine the clinical trial data to set the upper limit and incorporate into measurement. This doesn't get to where to set the other thresholds, but does set the upper limit.
  - Are there situations when it would be acceptable to have fewer than 5 star categories? As performance improves and scores are very closely clustered, are there situations when the 4 and 5 star categories should be combined?
    - One TEP member was uncomfortable with the constraint of 5 rating levels. Is it reasonable to have a smaller number of rating levels? When there are only 2 contracts at 1.0 star level, this category is under populated. At the upper end—as performance improves—some of the 4-5 star thresholds are extremely close, with only a 2-point difference, potentially leading to a distinction without a difference. The upper limit may lead to unintended consequences (e.g., strokes) of treatment that is too aggressive.
    - Another member highlighted the flip side of this where all contracts are high (e.g., 92% on nephropathy); are there instances where all contracts would receive a 5 star or the measure would not be

measured anymore? Also, there are situations where no contract should get 5 stars on a measure that is clinically important but no one is achieving. This commenter thought there should be some flexibility in the system.

- One TEP member observed that the solutions are not simple. It is fundamentally impossible to measure everyone with the same ruler when they have different populations, and CMS has made a heroic effort to do it reasonably. The challenge is developing specifics for addressing some of the issues discussed during this meeting. Some things can be fixed. If CMS is going to have thresholds (which this member opposed), then CMS shouldn't allow thresholds to bounce around, such as due to one new plan entering the program with initially low performance. The other problem is having so much money ride on this. Information on performance in and of itself is valuable. CMS will never get it perfect. There are a lot of differences between plans being evaluated.
- It was noted that some of the issues raised (e.g., exclusions such as palliative care, cognitive function) are better addressed by measure developers earlier in the process, not the Star Rating calculations.
    - In addition to looking at the measure attributes, it is important to look at the program attributes, and how to evaluate a measurement system or a program. This is an area where NQF is doing work and hopes in the future to have criteria to evaluate the measurement system/program.

Utility and Feasibility of Constructing and Reporting Star Ratings at the Level of Plan Benefit Package (PBP) or Geographic Area

- RAND team members, Justin Timbie and Marc Elliott, presented on the utility and feasibility of constructing and reporting Star Ratings at the level of Plan Benefit Package (PBP) or geographic area. *A copy of this presentation can be found in the corresponding slide deck.* Current measurement and reporting is at the contract level. Many contracts cover multiple states or regions. Furthermore, a contract might have really different plan benefit packages (e.g., one plan benefit package might be a dual SNP while another might be very different within the same contract). Some benefit packages might have higher or lower premiums, and so forth. Contract-level reporting is essentially a weighted average across those things. If a beneficiary is interested in a particular benefit package, they might want to compare options across specific benefit packages rather than looking at a mixture of benefit packages.
    - ***Expanding a survey to obtain more data at a more granular level is very expensive. Has CMS considered doing some sort of small area estimation as an alternative to collecting more data?***
        - RAND commented that if the goal is to make inferences at a smaller level, a model-based approach like small area estimation might allow CMS to reach its goal at a lower cost.

- One TEP asked whether, in the case of small contracts, their information would be combined with information from others in the same area or other contracts; plans may argue that the rating they would receive is not based on the performance of their contract and so isn't legitimate. The question is how would people react to that type of approach for generating estimates of performance at the PBP or geographic level? From a statistical point of view, you could get additional information without additional data collection. Is it acceptable to have information in that form? There was concern expressed about the administrative burden on providers of having to supply more data.

- ***TEP members agreed that having more information by geography would be beneficial to beneficiaries when making choices but wanted to understand the consequences.***
  - A TEP member stated that beneficiaries cannot compare MA and FFS contracts and that geographic-level reporting makes more sense once comparisons to FFS are possible. When will CMS have MA data that is comparable to FFS?
    - RAND commented that ongoing quality checks of the MA encounter data are occurring, which is being considered as a means for constructing measures.
  - If it is not feasible to change data collection, one option is to keep reporting units the same (i.e., contract-level) but provide supplemental reporting where more granular data exist.
  - A TEP member asked what would be the benefit or incentive to the plan of supplemental reporting? RAND commented that the only incentive would be public reporting absent payment; the value would be to allow consumers to make apples-to-apples comparisons—such as comparing two PBPs.
    - There was a comment that the additional information may be helpful to plans but may overwhelm consumers.
  - RAND asked how much additional burden it is to dump data for half the patients versus all patients for most measures, since NCQA has pushed to capture measures through administrative data and many plans submit on the universe of eligible patients (not a sample). One TEP member stated that if there could be guidance provided to contracts who have the resources to pull these data that could produce needed information to inform the decision process.
  - One TEP member asked whether PBPs change often, making reporting not relevant? RAND responded that PBPs do change from year to year but was not sure how dramatic those changes are, so that would need to be a consideration.
- There was support expressed for having RAND model some of these approaches so that plans could see how they would fare under geographic and PBP-level reporting.
  - RAND stated that based on preliminary analyses it had done with CAHPS and HEDIS measures examining the feasibility of PBP-level

reporting, PBPs are about as different from one another within a contract as contracts are different from one another; so that provides a signal. Then a reliability threshold of .8 is often recommended for things that involve high-stakes applications, and that means that of the scores that are put out there, at least 80% of their variance is due to true variation in performance as opposed to being a function of a small sample size. In general, 100 survey completes would get you that level of reliability for the currently reported CAHPS and HEDIS measures (drawn from CAHPS) were CMS to do this at the PBP level. The sample size that you need to get at least decent reliability isn't that large necessarily. A potential consideration is that in small contracts, beneficiaries may complete surveys in multiple years as the sampling pool is small. RAND also stated that most contracts currently have at least one PBP that would be reportable without changing things. The median contract has four PBPs and the mean contract has seven or eight PBPs, so that's often only going to be covering the biggest option or the biggest option or two. Only about one in five contracts has two PBPs that would be reportable for most measures under the way that CAHPS and HEDIS is collected now without any change to how things are done. Another metric is informativeness to evaluate the merits of PBP-level reporting; the idea here is that you learn more for some measures by going down to the PBP level and that you learn generally more by going down to the PBP level in CAHPS measures than you do in HEDIS.

- One TEP member stated that it would be helpful to use existing data to model the approaches and to allow plans to see the output before moving forward.
- RAND asked for input on how modeling might be done, as there are a lot of methodological details that need to be fleshed out if we did proceed.
    - What kind of analysis should RAND consider to inform whether it makes sense to pursue reporting either at geographic intersections of contracts or at the PBP level?
    - If RAND modeled something that was geographic, do you have recommendations about the geographic areas to use?
    - If RAND were trying to set things up to inform PBP level comparisons, is it particularly important that, for example, we allow somebody who's shopping for a dual Special Needs Plan to see different dual Special Needs Plans across different contracts or is it more important that we allow them to compare within a single contract? Or is it to assess whether the more expensive option within that contract is producing the same quality as the less expensive option within the same contract or are both of these comparisons potentially interesting?

- Related to service areas, some of these are non-contiguous and unintuitive contract service areas. Is there some way to address that issue without necessarily going all the way to stratifying by small areas where data might be sparse?
- Also, since several of these options would involve a lot more data collection (either patient surveys or HEDIS data collection), how should we think about the tradeoff? Is the added value worth it?
- Should RAND focus on PBP ratings as a replacement for contract ratings or as a supplemental source of information?
- If one were to pursue one of these approaches that drill down into smaller geographic areas or that try to drill down into PBPs within contracts, what would be the approach that you would recommend for dealing with the inevitable units that are smaller than we could really say anything reasonable about? Should pool them together? Should we refer back to the contract level information or some other option?

Measures

- RAND team member, Rachel Reid, presented on measure domains and Part C and Part D measures included in the calculation of Star Ratings. *A copy of this presentation can be found in the corresponding slide deck.*
    - ***How many display measures are there?*** There are around 30 display measures across Part C and Part D. Display measures include measures being considered for inclusion in Star Ratings calculations and measures removed from Star Ratings calculations.
    - ***How are measures being considered for inclusion differentiated?*** Through the Call Letter process, CMS indicates which measures are being considered for inclusion in the Star Ratings.
    - ***What are the gaps in the current measure set?***
        - One TEP member expressed concern that there are not enough outcome measures, putting in a plug for PROMIS. This member also favored measures of opioid and Rx-based measures, though cautioned we don't know enough about opioid measures (i.e., rapid deintensification of opioids and possible unintended consequences).
        - There were issues flagged related to measures. For some measures higher/lower isn't always better. The plan all-cause readmission measure was flagged as problematic given an arbitrary 30-day marker. CMS should get away from readmission measures to total use of inpatient care through more effective management of folks with chronic diseases (in ambulatory care).
        - One member stated that the current medication measures depend on prescription drug data. There is a long wish list about medication measures, but there are limitations to the data sources.
        - Another noted the measure set is lacking measures that take into account populations with behavioral health and substance abuse issues.

- There was a comment about making sure there is alignment across programs and measures to try to reduce burden on providers.
- A population subgroup that doesn't often get thought of are 18-64 years of age who are dually eligible beneficiaries and disabled. There is a need to review the measures and make sure this population is considered.
- One TEP member commented that there are only 30 standalone PDPs left (RAND corrected this figure and stated the number is around 60 PDPs). More and more plans are integrating drug and health benefits, such that medication standalone measures do not reflect the drive towards more integration and a focus on outcome measures.
    - CMS responded that while integrated care is the trend, the number of PDPs is decreasing because there have been consolidations. The Part D standalone measures continue to be needed.
- One TEP member expressed support for keeping some compliance measures. It doesn't compute for a consumer when they've just selected a 5 star plan and then they find out the plan has been sanctioned for major compliance issues, while other TEP members thought compliance measures should be dropped from Star Ratings as they are not quality measures.
- One TEP member cautioned that CMS should not make the Star Ratings system more complex by adding more measures if it goes beyond what beneficiaries want.
- TEP members expressed interest in patient experience and outcome measures.
- Tobacco use is another measure that is not currently reflected in stars.
- For complex patients, care coordination is very important and this is an area that is missing.
- If CMS is going to move towards more value-based measurement, there is a wide range in cost for a given rating. We know that cost and quality are often not related. Cost is a component missing from the ratings.
- CDC is looking at measures of ED visits and hospital admissions. CDC is validating ICD codes in administrative data. These measures have been relatively easy to construct with administrative data.
- Regarding measures of high dosage of opioids, CMS should coordinate with the work PQA is doing in this space. To caution, incentives for rapid decrease of opioid use could have unintended consequences.
- It would be interesting to assess attempts to reduce medications. We have little information on how older patients feel about medication overuse. In one survey, only 14% of older adults agreed that more medical care is always better. This is an area that would be important in this population.

- One TEP member felt the measure list is not currently balanced and is incongruent.

- ***Other measurement-related issues the TEP identified were as follows:***
  - There was interest in using administrative data to construct measures and patient reported measures, for example patient-reported measures regarding function. CMS should consider stratifying measures by population subgroups.
  - One TEP member indicated that providers and plans are not happy with the CAHPS' mode of survey administration.
    - CMS is testing web-based and mixed method surveys across different settings and populations. Web-based surveys is something that CMS is committed to exploring. There are some populations that will complete web surveys while some older populations do not have access or the ability to complete a web survey. CMS needs to make sure everyone is able to respond.
  - Concern was expressed about the time lag between measurement and the rating year. The example provided was a contract could have a compliance ding based on older data which has since been corrected but is still reflected in the Star Rating. Beneficiaries are not getting up-to-date information on a contract's performance.
  - CMS should create better measurement alignment across programs; particularly with MACRA and ACOs moving forward given it is the same providers reporting measures across various CMS programs, with potentially different specifications for similar measures.
  - CMS needs to make sure the measures are meaningful and actionable.
  - It would be great if the MA program would be a leader in disparities measurement, reporting, and reduction.
  - Disenrollment is a current Star Rating measure. Do we know why people disenrolled, which could be helpful information? RAND responded that it implements the annual disenrollment survey on CMS's behalf to find out why beneficiaries voluntarily disenroll from contracts and this information is posted on Medicare Plan Finder.
  - Risk-adjustment matters, and CMS cannot compare plans when they have a large portion of beneficiaries in high risk or high use groups. This is a field that has been resisting risk adjustment, which it shouldn't because it matters.
    - Both NCQA and PQA as measure developers have been working on risk-adjustment and will be building this into measure specifications starting next year.
    - RAND has already been exploring adjustments for social risk factors and this will be topic of the October TEP meeting. The Star Ratings adjusts for the CAI, which proxies case-mix adjustment.
  - There were several comments on reducing burden and improving alignment:

- Could CMS use a smaller pool of patient experience questions across all settings? PBGH has developed a provider-level survey that has 12 questions that is reflective of the full CAHPS survey. More streamlined measurement should be considered.
- There is possible duplication across HOS and CAHPS; might some of the HOS questions be folded into CAHPS?
- There are also measures used in different programs that don't use the same specifications.

- ***Beneficiaries should be able to weight components of the overall composite measure differently.*** One TEP member commented that there is a problem with composite measures—you throw a lot in, weights are inherently arbitrary, and beneficiaries would likely weight differently. Ideally, there would be a personalized Star Rating scheme, based on beneficiary priorities. Beneficiaries may be willing to input some of their personal information to get personalized information in return. Allowing beneficiaries to enter their own weights and preferences would be important. CMS could make Plan Finder more interactive to help beneficiaries. This might be easier to implement on the Part D side versus on the Part C side. There are ways to improve Plan Finder. Also, CMS should support FFS performance being described on Plan Finder so that beneficiaries can make comparisons.
    - There is concern about HOS and diversity of patient populations across plans. The HOS survey is only mailed or available over the phone in 3 languages. This puts some plans at a disadvantage.
- ***The discussion shifted to considering how to define when a measure is topped out and whether to remove the measure. Comments included:***
    - If a measure has plateaued but the measure is important and we think plans can do better, we should leave it and not give anyone 5 stars.
    - There are implications related to pulling out a "topped out" measure. There is some literature that shows that when a measure is removed, it stops being incentivized and there is back-sliding on performance.
        - RAND asked whether it would be helpful to see what has happened to performance of measures moved to the display page, and the TEP expressed interest.
    - Should CMS consider some maintenance level that a plan gets credit for?
    - Retiring a measure is an opportunity for new measures. For example, retire BMI and add a meaningful measure (e.g., obesity rates).
    - Some TEP members supported removal of topped out measures but noted it would impact contract Star Ratings.
- ***Throughout the day, the term "meaningful" measures has come up***. RAND asked the TEP if it could identify which measures are "meaningful" and should be included in the Star Ratings and which should be dropped?
    - One TEP member pointed to a study published recently in the *New England Journal of Medicine*, which rated measures in the MIPS program against several criteria to determine which measures were meaningful. The criteria used paralleled the NQF criteria such as validity and eligibility but used a

different process for rating—it used a modified-Delphi panel process to rate the measures.

- There was interest among TEP members to continue the measures discussion at future meetings.

SUMMARY OF ANALYTIC IDEAS THAT SURFACED DURING THE TEP ON STAR RATINGS

Thresholds (Advance Notice, Guardrails, Increasing Stability)

1. Merging 4 & 5 stars sometimes and/or having absolute ceilings was an area identified as worth exploring. RAND could simulate versions of some of these approaches if desired. For example:
    a. One could set a maximum upper threshold based on clinical considerations, a level of performance not distinguishable from 100% at standard sample sizes or some similar criterion. This might be something like 95 or might not always correspond to a fixed 0-100 value, above which all scores are 5 star.
    b. One could set a minimum difference between 4 and 5 stars where you merge if it's not met (could be 1-2 points on 0-100 or could be determined by something at the patient level).
    c. One could set a lower bound that merged 1-2 or 1-2-3 into a higher category based on an absolute threshold (maybe above 80% can never be less than 2-stars, above 90% can never be less than 3 stars, or base this on something at patient level)
    d. All of these "amounts" (upper and lower bounds, minimum differences) would be different for binary measures on the one hand and CAHPS/continuous measures on the other hand. And some of these might be specifically clinically informed.
2. There was support in the TEP for reducing the effect of outliers on thresholds.
3. There was support for using the most up-to-date data in estimating thresholds rather than using older data to predict thresholds in order to allow advance notice.
4. If CMS were to eliminate any big consequences of thresholds by moving from "stair-step" to "continuous" approach, then advance notice would not be needed because nothing bad happens to a contract when their score is 1 point below a threshold. Guardrails matter less when this approach is combined with resampling to reduce noise. For this to work, incentive payments have to be continuous/smooth—not just the Star Ratings. One approach to modeling this would be:
    a. What if measure scores were transformed to a continuous star score 1.0-5.0, with "thresholds" setting the location of 1.0, 2.0, 3.0, 4.0, and 5.0, with piecewise linear connections between them (so different slopes).
    b. These would then be averaged to get continuous summary stars 1.0-5.0; payment/bonuses would smoothly be attached to these.
    c. This result is that thresholds would matter less, so advance notice would perhaps not be needed. It is possible that this would improve measurement in several ways.
    d. Such an approach could still use stars to display and summarize as well as to perform the function of converting from unlike scales to a common scale.

Alternative Levels of Reporting (Geographic and PBP-level Reporting)

1. There was some support for PBP-level reporting, not to replace contract-level reporting but to provide supplementary information for consumers, without extra data collection.
    a. It was suggested as an approach to use small area estimation/modeling adding borrowed strength from models to generate PBP-level estimates.
    b. Additional work is required to identify how geographic units would be defined.

Measures

1. There was support for identifying which measures have topped out (low variation, high performance), and running Star Ratings with those measures removed and determining how that would affect contract ratings.
2. There was support for developing criteria on what is clinically meaningful and running a Delphi process with experts to rate measures against the criteria to winnow the measure set to those that are clinically meaningful. Could simulate the effects of Star Ratings if measures that aren't meaningful are removed.
3. There was support for examining what happened to previously incentivized measures that were moved to the display page, in order to assess whether backsliding has occurred in the absence of incentives.