Conference Proceedings

RAND Technical Expert Panel Medicare Advantage (MA) and Part D Contract Star Ratings, May 5, 2021 Meeting

MEETING SUMMARY

Cheryl L. Damberg and Maria DeYoreo

For more information on this publication, visit **www.rand.org/t/CFA1504-1**.

Published by the RAND Corporation, Santa Monica, Calif.

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

May 5, 2021

MEETING ATTENDEES

Technical Expert Panel Members

NAME	POSITION
Arlene Ash, Ph.D.	Professor and Division Chief for Biostatistics and Health Services Research in the Department of Quantitative Health Services, University of Massachusetts Medical School
Liza Assatourians, J.D.	Vice President of Federal Programs, America's Health Insurance Plans (AHIP)
Ann Burns, R.Ph.	American Pharmacists Association
Kim Caldwell, R.Ph., B.S.	Principal, Texas Star Healthcare Consulting
Lindsey Copeland, J.D.	Medicare Rights Center
Darrell Gaskin, Ph.D.	Professor and Director of the Center for Health Disparities Solutions, Johns Hopkins University
Emma Hoo, B.A.	Director, Pacific Business Group on Health
Deborah Paone, Dr.PH., M.H.S.A.	Performance Evaluation Lead for Quality Measurement, Social Determinants of Health, and Care Innovation, Special Needs Plans (SNP) Alliance
Ninez Ponce, M.M.P., Ph.D.	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, M.D., M.P.H.	Professor of Medicine and Pediatrics, University of California Davis School of Medicine
Allyson Schwartz, M.S.S.	President and CEO, the Better Medicare Alliance
Jane Sung, J.D.	Senior Strategic Policy Advisor, American Association of Retired Persons (AARP)
Dolores Yanagihara, M.P.H.	Vice President of Analytics & Performance Information, Integrated Healthcare Association (IHA)

Technical Expert Panel Member who was unable to attend:
Eve Kerr, M.D., M.P.H., University of Michigan

RAND Staff
Cheryl Damberg, Project Director
Maria DeYoreo, Project Codirector

Marc Elliott, Principal Senior Statistician
Marika Booth, Statistical Analyst
Melony Sorbero, Senior Policy Researcher
Anagha Tolpadi, Statistical Analyst
Chau Pham, Project Manager

MEETING SUMMARY

Welcome and Introductions

- The RAND Corporation project director, Cheryl Damberg, began the meeting by welcoming attendees and reviewing the meeting agenda.

Overview of Topics Discussed at the November 2020 Technical Expert Panel (TEP) Meetings Related to Social Risk Factors and Star Ratings

- RAND team member Melony Sorbero briefly reviewed items that were discussed at the two TEP meetings that occurred in November 2020, noting that the TEP members agreed that the Centers for Medicare and Medicaid Services (CMS) should proceed with confidential reporting of stratified performance to Medicare Advantage (MA) and Prescription Drug Plan (PDP) contracts.
- Since the November meeting, the RAND team conducted analyses using 13 Part C and two Part D measures to assess the number and percent of contracts whose performance can be reliably measured in social risk factor (SRF) strata. The strata examined were Low Income Subsidy (LIS) or dually eligible (DE) for Medicare and Medicaid versus non–LIS/DE, and disabled versus nondisabled. To report a measure by stratum, the contract would be required to meet the denominator criteria for individual stratum and have reliability of at least 0.60. Dr. Sorbero presented the findings to the TEP.
- A TEP member asked how many contracts there are in total.
    - RAND responded that the total number of contracts with data varies by measure (slide 10).
- A TEP member asked for clarification on whether contracts that did not meet criteria did not have enough people in the denominator to report by stratum.
    - RAND responded that not meeting the criteria for an individual stratum meant not having enough people in the denominator or having a score with low reliability.
- The TEP member then asked whether or not meeting the criteria was related to the distribution of disabled or LIS/DE people across contracts.
    - RAND confirmed that is true. If a contract had a large number of enrollees but a small number in a specific SRF stratum, it might not have a sufficient number in the denominator for that stratum or might have low reliability in that stratum.
- One TEP member commented (referring to slide 10) that it is notable that the measure with the lowest percent of contracts meeting the criteria (i.e., the Health Outcomes Survey [HOS] improving bladder control measure) was an improvement measure. The TEP member asked the RAND team if there were insights about why this was the case and how this could be addressed
    - RAND responded that having a small percentage of contracts meeting the criteria might be a function of the fact that this is a survey measure collected through the HOS, which has a modest sample size, and the sampling frame is not stratified to obtain reliable estimates for subgroups of beneficiaries.
- Another TEP member thought that the improving bladder control measure does not measure person-level improvement over time, but rather compares a contract's

performance year-over-year based on the sample in year one and the sample in year two to assess contract improvement.
    - RAND clarified that improving bladder control is not an improvement measure despite the measure's name. This measure assesses the percentage of plan members with a urine leakage problem in the past 6 months who discussed treatment options with a provider. This measure is based on cross-sectional HOS data from a single year (e.g., for Star Rating year 2021, data collected from Cohort 20 Follow-up [2019] and Cohort 22 Baseline [2019]).[1] Contracts must achieve a denominator of at least 100 to obtain a reportable result.
    - Other HOS measures assess improvement over time, but RAND did not include any HOS outcomes measures that assess improvement in the stratified performance reporting work. RAND focused only on cross-sectional measures.
- A TEP member asked, regarding the number shown on slide 11 for the disabled stratum, whether the 97 contracts represent a high proportion of the disabled population or whether they have a lot of disabled beneficiaries enrolled.
    - The RAND team did not look at this but indicated it could do so. As follow up, the team could calculate (1) the percentage of those contracts' beneficiaries who are disabled and (2) collectively what proportion of all disabled MA beneficiaries are in the contracts measured. Both could be informative.
    - RAND noted that with a few exceptions, in most contracts fewer than 50 percent of enrollees are disabled. There is not a concentration of disabled beneficiaries in a single contract like there is for LIS/DE (e.g., Dual Eligible Special Needs Plans or very high percentage LIS/DE contracts). Most contracts have a modest percentage of disabled enrollees.
- A TEP member wanted to understand if RAND was measuring disability using the reason the person first enters the Medicare program.
    - The RAND team confirmed it uses the variable OREC, which is the original reason for Medicare entitlement. RAND does not designate someone as disabled if the original reason for entitlement was End State Renal Disease or old age and the person later became disabled.

Discussion Questions

- RAND posed the following discussion questions to the TEP:
    - *Should there be a minimum number, or percentage, of contracts that must meet denominator and reliability criteria for performance to be confidentially reported to contracts for a given measure?*
    - *What should that minimum be or how should it be determined?*
    - *Are there any downsides to reporting results to plans if only a very small number of contracts have "reportable" data for a measure?*

[1] Each year a random sample of Medicare beneficiaries is drawn and surveyed from each participating MA contract with at least 500 enrollees (i.e., a survey is administered to a different baseline cohort, or group, each year). Starting in 2007, the MA sample size was 1,200.

- The RAND team noted that during the discussion on previous slides, a TEP member gave the team an additional criterion to consider, specifically the proportion of the population covered (e.g., disabled) by reportable contracts.
- A TEP member stated that the report is very valuable, especially because it is providing confidential information back to the plan. They did not see a downside to doing this, besides time and effort.
- A TEP member asked whether there were any questions or concerns about the accuracy or validity of the HOS survey measures. Is this meeting the goal of having good, stratified information on these populations that are at risk? This member agreed CMS needed to start somewhere; however, one would not want to start somewhere where it is questionable. This member wondered whether it will be taken seriously by the people that CMS wants to act on this information.
    - RAND clarified that most of the measures examined were HEDIS (Healthcare Effectiveness Data and Information Set) measures.
    - RAND commented that, although questions have been raised about various surveys and measures, some of the statements made in those contexts have not reflected the preponderance of evidence. The Star Ratings measures being discussed all have evidence of validity.
- A TEP member was strongly in favor of using the information on SRFs, but believes it will be a struggle to get information across clearly. It is probably worth continuing to push forward to see whether something will come out of using this information. The member cautioned that it is important to proceed slowly. The TEP member commented that there might not be full disclosure of information either from providers or patients.
    - RAND clarified that the focus is on the SRF information contained in CMS administrative data (enrollment information) regarding LIS, dual and disability status, and these SRF indicators should not be problematic.
- A TEP member added that when discussing burden, it is important to distinguish between adding additional data to the report or something that requires additional data collection. Once the report is programmed, there is not much effort required to calculate and populate the report as long as extra data collection is not needed. As long as it involves providing more information that will help put information into context, this TEP member was in favor of it. It would be useful to tell people: "Here's how your contract scored; this is how many people would be in the denominator, this is how many people we tried to survey and that the percentage we are giving you is based on a certain response rate." Having those data available to plans to dig into is a good thing.
- Another TEP member agreed that this is not extra data to be collected, so it is not an extra burden. The information is already being collected and it is just being reported out, but the contracts should be aware of the limitations associated with using these data. Getting feedback from the plans to see what they think about the usefulness and validity of the reports would be helpful.
    - Another TEP member added that if CMS's goal is to raise awareness of disparities, then there is not a downside to reporting measures with small numbers of contracts being able to report in strata.

Review Mock-up of Stratified Report Template

- Dr. Sorbero presented a draft report template containing artificial data to illustrate how stratified performance data could be shared confidentially with contracts through the CMS Health Plan Management System (HPMS). The report template shows artificial stratified performance scores for LIS/DE versus non–LIS/DE and for disabled versus nondisabled. Both included national comparisons.
- A TEP member noted that showing comparisons of contracts with other similar contracts (e.g., with a similar enrollment profile) would be better benchmarking. If CMS compares contracts to other contracts that are completely different in terms of their enrollment profile, it would be much harder for a given contract to understand and use the data.
    - Another TEP member added that they liked the idea of a contract being able to compare itself to other similar contracts. It is an interesting, nuanced question and is something worth thinking about it.
    - RAND asked how to appropriately define similar contracts. Are quartiles of the risk factor narrow enough? Or would smaller subgroups of contracts be important?
        - The TEP member stated that with other data, deciles are used. If not deciles, maybe quartiles are a good place to start.
    - Another TEP member concurred with the idea of comparing contracts with other, similar contracts, not based on performance, but based on the inputs to that performance, particularly if there is a different composition of the population served. This is similar to the Academic Performance Index (API) scores for schools in California, which compare schools with similar demographics (e.g., percent on free and reduced lunch).
- Another TEP member disliked comparison based on means. This TEP member said that it would be better to see a distribution of scores across contracts in the comparator group (e.g., similar contracts or nationally).
    - RAND asked whether it should be the full range of performance or limited to performance in the interquartile range?
        - The TEP member suggested that histograms are a good way to display the information.
    - The RAND team commented that it would need to investigate whether showing this type of display would be possible in HPMS.
- A TEP member asked if there is any information obtainable about fee-for-service (FFS) performance, which would be another interesting benchmark for comparison.
    - RAND shared that it had done work to compare FFS performance in contract service areas to MA or PDP performance on a subset of measures that can be created with claims and/or encounter data. The team had not discussed whether to include FFS as another benchmark for comparison.
- A TEP member asked RAND to comment on the confidence interval (CI) surrounding the difference. Are there any concerns about the width of the CI? This TEP member was slightly concerned about misinterpretation of relatively small differences.
    - RAND commented that it selected sample sizes and reliability thresholds to support comparisons, but could not say what the CIs are. The measurement approach was intended to provide reasonably precise comparisons.

- As follow-up, RAND indicated that it would look at the standard errors of the estimates to determine if CIs are very tight around the estimates and, as a result, small differences could be flagged as significantly different without being meaningfully different.
- Another TEP member pointed out that measures such as breast cancer and colorectal cancer screening have many exclusions for different clinical situations to ensure the measure is clinically appropriate for included individuals, so it gets down to how the measure is captured, how much data are left for analysis, and whether the comparisons allowed for noise (related to sampling, measure specification, etc.), so the CI would be important to examine.
    - RAND indicated that all of the measures proposed for the stratified reports use two years of data to provide more robust estimates.

Discussion Questions

- RAND posed the following questions to the TEP:
    - *Does the mock-up report capture important information? Should anything else be added?*
    - *A potential future enhancement for contract-level performance is an indicator for whether a contract's performance percentile for beneficiaries with SRFs is at least X points different than the contract's performance percentile for beneficiaries without SRFs. What is a meaningful difference that should be flagged?*
- RAND commented that based on the TEP discussion, there appears to be support for additional benchmarks (e.g., benchmarks based on FFS beneficiaries or contracts with similar profiles in terms of percent disabled, percent duals, low-income).
    - Several TEP members agreed.
- A TEP member stated that some of the column labels in the contract level performance data seem a bit awkward (slide 17). The type of information seems reasonable, but the labels could be streamlined. For example, the column "Contract LIS/DE Performance Substantially and Significantly Higher/Lower Compared to National LIS/DE Performance" might need rephrasing to be clearer. This TEP member remarked that you can interpret "significantly" to cover both the concept of clinical and statistical significance.
    - A TEP member stated that the contract-level performance data is helpful to display in the report.
    - Another TEP member stated that the presentation of the data in a format that is easily digested is important. Some real attention should be put into being clear, having good labels, and presentation of results in an illustrative rather than tabular format.
    - RAND stated that it would discuss with CMS what is possible to display in the HPMS environment. Generally, what has been shared is just data, as opposed to different types of displays to help with interpretation. RAND agreed that CMS should provide information that is easily digestible by plans.

- Another TEP member stated that providing data without comparisons could be hard to interpret and act on, especially without comparing like-to-like contracts using peer groupings.

- A TEP member wondered, for comparisons that classify performance as higher, lower, or similar, whether there would be value in flagging big differences (e.g., 6 points higher or lower) versus those just outside the 3-percentage-point range that was set for *lower* or *higher*. This would speak to things that plans should be alarmed about and work on or feel good about in terms of their performance.
  - Several TEP members agreed that it would be good to flag differences that are quite large.
- A TEP member mentioned that in the commercial sector, for some of the measures that are more mature, the distinction between high and low performance narrows. Would slicing performance into segments to define large versus smaller differences be more challenging when describing what is a significant difference?
  - RAND noted that some of these measures have been in Star Ratings for quite a while. At the contract level, some measures have larger variation than others in terms of high and low performance. RAND stated that it could investigate those distributions.
  - RAND agrees that on average the variation decreases and scores increase, but this is not uniformly the case, and there is still quite a lot of variation. It is worth thinking about that dynamic and trying to find meaningful differences.
  - Dr. Sorbero presented an approach to summarizing contracts' performance relative to national performance (i.e., higher, similar, lower) using a stacked bar for each strata for three hypothetical measures (slides 19 and 20).

Discussion Questions

- RAND posed the following questions to the TEP:
  - *Would contracts find this type of information summary helpful?*
  - *Who else might be interested in seeing this type of data summary?*
  - *Are there other data summaries that would be useful to contracts or others?*
- A TEP member commented that showing an individual contract where it falls in the distribution would provide helpful context to the contract for understanding the data. Having the bar graph indicate specifically where the contract falls in the distribution would be very helpful.
- Another TEP member added that contracts would like to be able to compare themselves to other contracts. Comparisons are important. They want to do their best and are competitive. This TEP member had incorrectly interpreted what the display was showing, thinking it was showing the average performance.
  - The RAND team said that they will be very careful and explicit in what the numbers in the bar are indicating.
- A TEP member asked for clarification on why these patterns appear in the "contracts' performance compared to national performance" graph and if it is because the thresholds used for comparison are not based on the sample but rather are based on data from all plans.
  - RAND commented that the national performance is based on all contracts that receive a Star Rating in the most recent Star Ratings year and is not

restricted to the contracts that met reliability and denominator criteria for that stratum. When the team restricts contracts based on meeting the denominator and reliability criteria, there is a tendency to lose the smaller contracts across the board. On average, smaller contracts do have a tendency to have lower performance than larger contracts, and that can lead to some of the differences observed for national performance versus performance among contracts that meet the criteria for the individual stratum.

- Another TEP member asked for clarification on the national performance statistics reported at the bottom of the graphic.
    - RAND clarified that it is showing the comparison of the contract performance to the national performance in that particular stratum. The national performance is shown for each stratum.
    - The TEP member stated it was not clear that there were four separate benchmarks.
    - RAND suggested that it could rearrange the national performance statistic so that national performance in strata is listed directly below each respective bar graph.
- RAND concluded that there is more work to be done to improve the clarity of the displays. RAND will continue to think about other ways to present the data and welcomes any additional comments from the TEP on how to display information.

Assessing Social Risk Factors

- Dr. Sorbero also presented slides on RAND's work on assessing the availability of SRF data and review of existing assessment tools.
- A TEP member asked which domain race was categorized under. The member noted that race is part of cultural context but could also signify discrimination. Having race separate from the other SRFs would be important because it embodies many things, not just culture. If race is a barrier, it is a risk factor. But race also is associated with housing discrimination, transportation discrimination, and food security.
- A TEP member asked what the goals are. Is the goal to better incentivize contracts to intervene? Is the goal to define what kinds of interventions are the best? Is the goal for contracts to have more information to recognize what the problems are and identify ways to address them? What are contracts expected to do with this information? Will providing contracts with more tools help contracts achieve these goals? What are the goals for CMS? Can questions be added to these tools to capture more information?
    - RAND agreed that these are great questions. RAND is in the process of taking the first step, trying to understand existing assessment tools, what information is captured by these tools, and whether the content of these tools maps to the greatest areas of concern. RAND is trying to determine whether it would be valuable for CMS and/or plans and their providers to get SRF information.

Discussion Questions

- RAND posed the following questions to the TEP:
    - *Should the Star Ratings program include a measure to incentivize the collection of SRFs by MA and Part D contracts?*
    - *Which SRFs are the highest priority to collect?*
    - *Should CMS leverage an existing assessment tool or provide contracts with flexibility on which tool to use if certain SRFs are collected?*
    - *Who should be responsible for collecting the data and how often?*
- A TEP member cautioned that some individuals might obtain care outside of the Medicare program, and consequently, the Medicare data will not capture the totality of care provided to beneficiaries. This person thought that to accurately represent disparities, it would be important to capture data about care received outside of Medicare.
- A TEP member commented that it is important to check that the SRF data collection tools have been tested among different populations (e.g., non-English, disabled, etc.) and vetted. Even if a preferred tool is identified, is it realistic to mandate a single tool to apply as a mandate for all MA contracts? Will data be rolled up from multiple tools into one platform that equalizes the results? Also, risk factors are not all equal; risk factors impact people differently therefore we might need to incorporate weighting. What risks are important relate to the person and their situation. The importance of the risk factor varies across individuals. Some of the tools had yes or no questions, while others had scales, which also complicates things.
- A TEP member stated that this requires a lot of thought. Many of these tools were developed to be administered in a primary care setting or by health care organizations, where people are in waiting rooms and have established relationships with providers. People's relationships with their plans are quite different and, often, distant. People generally do not want to have much to do with their plan; often, the relationship is hostile. This member worried about the additional burden of asking enrollees questions about SRFs, and depending on the context and setting in which the survey is administered. The devil is in the details on how to administer the survey question and how to assure that the data are protected and used for appropriate purposes.
    - RAND noted that this last TEP comment plays to the last question RAND posed, specifically, "Where is the right place to capture this information?"
- RAND concluded the meeting by stating that it is not finished discussing this topic with the TEP and will return to it in a future meeting. RAND asked TEP members to share any additional comments they might have postmeeting by sending an email to the RAND team.