



WHITE PAPER

Special Needs Plans under Medicare Advantage Quality Measurement: A Focused Look at the Medicare Health Outcomes Survey (HOS)

© December 2018

*By: Deborah Paone, DrPH, MHSA
Performance Evaluation Lead, SNP Alliance*

Focus on the Medicare Health Outcomes Survey

The Special Needs Plan Alliance (SNP Alliance) has been surveying health plans, analyzing peer-reviewed literature, and gathering information from researchers and experts to better understand the strengths and limitations around the use of the Medicare Health Outcome Survey (HOS) in the Medicare Stars rating and quality measurement system for Medicare Advantage plans. There is compelling evidence indicating need for improvement. The challenges are particularly pronounced for special needs plans as they serve diverse, low-income, disabled, and chronic care, complex, or advanced illness populations. These populations and plans are more likely to be negatively impacted by the limitations in HOS instrument, methods, and scoring. This White Paper describes the limitations of HOS and offers recommendations for improvement.

For more information contact Performance Evaluation Lead, Deborah Paone, or President & CEO, Cheryl Phillips at the SNP Alliance. They can be reached at dpaone@snpalliance.org or cphillips@snpalliance.org See also: www.snpalliance.org for additional publications and reports.

Table of Contents

Executive Summary	1
Background	2
Challenges and Limitations of HOS for Special Populations	7
<i>Diversity of Medicare Population not fully Considered</i>	7
<i>Instrument Limitations as Applied to People with Special Needs</i>	8
<i>Inadequate Methods and Administration</i>	9
<i>Information Decay, Time Lag</i>	11
<i>Proxy Bias</i>	11
<i>Results May be Affected by Underlying Characteristics of Enrollment</i>	12
<i>Attribution, Context Issues</i>	14
<i>Inadequate Models and Adjustments</i>	15
<i>Further Data and Cautions from the Field</i>	16
Options for Addressing Limitations	18
1. <i>Level the measurement field</i>	18
2. <i>Use additional exceptions and exclusions</i>	19
3. <i>Discontinue the longitudinal design</i>	20
4. <i>Use better predictive models and data sources</i>	21
5. <i>Use alternative instrument(s)</i>	21
6. <i>Re-test HOS with disabled and diverse groups</i>	22
7. <i>Improve methods of administration and accommodation</i>	22
8. <i>Conduct analysis at a finer level and report on results</i>	23
9. <i>Provide greater transparency and consumer education</i>	23
10. <i>Set up feedback groups to advise and inform</i>	23
Conclusion	25
Appendix A: Medicare Advantage Quality Measurement	27
Appendix B: Medicare Health Outcomes Survey Instrument	31
Endnotes	37
References	38

Executive Summary

The Special Needs Plan Alliance (SNP Alliance) recognizes the challenges in identifying, developing, and using appropriate self-reported health outcome measures in performance evaluation. Measures that meet necessary standards around accuracy, validity, reliability, feasibility, and utility must be coupled with systems that deploy robust methods for data collection, administration, and analysis to ensure a fair, accurate, and robust quality measurement system. This is particularly hard to achieve when the characteristics of the population being reviewed are diverse and the units of analysis are very different.

Over the last few years the SNP Alliance has surveyed health plans, analyzed peer-reviewed literature, and gathered information from researchers and experts to better understand the strengths and limitations around use of the Medicare Health Outcome Survey (HOS) in the Medicare Stars rating and quality measurement system for Medicare Advantage plans. There is compelling evidence indicating need for improvement. The challenges are particularly pronounced for special needs plans as they serve diverse, low-income, disabled, and chronic care, complex, or advanced-illness populations. These populations and plans are more likely to be negatively impacted by the limitations in HOS instrument, methods, and scoring. This White Paper describes the limitations of HOS and offers recommendations for improvement.

Limitations examined include:

- *Diversity of Medicare Population not fully Considered*
- *Instrument Limitations as Applied to People with Special Needs*
- *Inadequate Methods and Administration*
- *Information Decay, Time Lag*
- *Proxy Bias*
- *Results May be Affected by Underlying Characteristics of Enrollment*
- *Attribution, Context Issues*
- *Inadequate Models and Adjustments*
- *Cautions Regarding Analysis, Interpretation, and Reporting*

The SNP Alliance offers potential solutions and strategies for addressing limitations, from those that would require only a change in analysis and reporting to those that would require more extensive modification. These include:

1. Level the measurement field
2. Use additional exceptions and exclusions
3. Discontinue the longitudinal design
4. Use better predictive models and data sources
5. Use alternative instrument(s)
6. Re-test HOS with disabled and diverse groups
7. Improve methods of administration and accommodation
8. Conduct analysis at a finer level and report on results
9. Provide greater transparency and consumer education
10. Convene stakeholder and expert panels

The SNP Alliance is working with a coalition of stakeholders to promote awareness of these issues and engage in discussion to work toward solutions.

Background

Special Needs Plans & Population Groups

Special Needs Plans (SNPs) and Medicare Medicaid Plans (MMPs) are a subset of Medicare Advantage (MA) plans specifically authorized and designed to meet special care needs of Medicare beneficiary populations groups. Special population groups include: **younger people with physical disabilities** (age 18-64), people with **severe complex and disabling conditions** (e.g., ALS, Parkinson's, advanced renal disease, COPD, AIDs-HIV, etc.), and **frail elderly persons** with cognitive, functional, and disease-related impairments. The plan types and subgroups include:

- **Chronic condition SNPs** (C-SNPs): serving persons with certain severe or disabling chronic conditions (e.g., HIV-AIDS, chronic heart failure, COPD, mental illness, etc.).
- **Institutional SNPs** (I-SNPs): serving persons residing in nursing homes or with comparable care needs in the community.
- **Dual eligible SNPs** (D-SNPs): serving persons covered by both Medicare and Medicaid.
- **Fully Integrated Dual Eligible SNPs** (FIDESNPs) and **Medicare-Medicaid Plans** (MMPs) – which are a specific type of D-SNP and provide both Medicare and Medicaid benefits, including long-term services and support.

SNPs and MMPs work to coordinate an extensive service array with and for those who are **dually-eligible** for both Medicare and Medicaid. These individuals often require community long-term services and supports, behavioral health services, specialty medical, pharmaceutical, and condition-focused care, as well as other assistance to address their complex needs. The health plan works to integrate and coordinate the two separate programs — Medicare and Medicaid — each with different rules governing how plans and providers may interact with the beneficiary. While SNPs are regulated, evaluated, and paid on the same basis as other MA plans, they are required to provide additional benefits and services to their target populations and to implement tailored care management according to unique Models of Care that serve these distinct enrollees. There were over 2.6 million people enrolled in SNPs or MMPs nationally in 2018.

Medicare Advantage Quality Measurement

Brief Overview of Medicare Star Ratings¹ - The Centers for Medicare and Medicaid Services (CMS) established and oversees the Medicare Advantage Quality Management System including Part C and Part D Star Ratings. CMS has described the goals² of its Quality Management System as follows:

- *To display quality information on Medicare Plan Finder for public accountability and to help beneficiaries, families, and caregivers make informed choices by being able to consider a plan's quality, cost, and coverage;*
- *To incentivize quality improvement;*
- *To provide information to oversee and monitor quality;*
- *To accurately measure and calculate scores and stars to reflect true performance;*
- *To recognize the challenges of serving high risk, high needs populations, while continuing the focus on improving health care for these important groups.*

Medicare Star Ratings & Measures³- There were 48 measures within the Medicare Advantage Part C and Part D Star measure set in 2018, and these were divided into five categories:

- 1) Outcomes Measures
- 2) Intermediate Outcomes Measures
- 3) Patient Experience Measures
- 4) Access Measures
- 5) Process Measures

Measures rely on data from four sources:

- Health Plans and Drug Plans
- Survey of Enrollees
- Data Collected by CMS Contractors
- CMS Administrative Data

Each measure has specifications to define the numerator, denominator, exclusions/exceptions, and any adjustments that should be made. CMS has assigned a weight to each measure from 1 to 5. The improvement, outcome, intermediate outcome and patient experience measures are weighted highest. Process measures are weighted the lowest. Health plans are reviewed and scored at the contract level. Many health plans have multiple plan products within a single contract. After all the measure-specific scores are tabulated, the summary and overall Star rating for the health plan is calculated as a weighted average of the measures.

SNP Specific Measures- Special Needs Plans, including (1) Chronic Conditions SNPs, (2) Dual Eligible SNPs, and (3) Institutional SNPs have four more Star measures to collect/report than general MAOs. These are all process measures with a weight of 1:

ID	Measure Name	Brief Description – % SNP enrollees within the measurement year:
C08	SNP Care Management	. . .who receive a Health Risk Assessment
C09	Care for Older Adults – Medication Review	. . .of older adults who had a medication review
C10	Care for Older Adults – Functional Assessment	. . .of older adults who had a functional status assessment
C11	Care for Older Adults – Pain Assessment	. . .of older adults who had a pain assessment

Star ratings are carefully watched by plans, advocates, and industry experts. This is because they impact MAOs in reputation, marketing, enrollment, and through bonuses. Star ratings are a presumed marker of performance in care management and care provision, as well as an indicator of internal processes. Star ratings are used by consumers in selecting among health plans when considering their health insurance options. Plans with 4.5 or 5 stars are considered “high performing plans” and have a special star icon on the Medicare Plan Finder. Star ratings affect plans financially. Plans with 4 stars or more can receive quality bonus payments which can be significant. Plans with less than 3 stars as an overall rating triggers the potential loss of their entire contract with CMS given poor performance.

Medicare Health Outcomes Survey (HOS) Instrument

The Medicare Health Outcomes Survey (HOS) evolved from other surveys. The Medical Outcomes Trust and Short Form-36 (SF-36) instruments were first developed and tested in a Veterans population in the late 1980's/early 1990's by John E. Ware and colleagues at Tufts University.⁴ These instruments were designed to be brief, generic, self-administered surveys for individuals to report/rate themselves on health-related quality of life, conditions, function, and disease burden. The HOS was derived from a shorter version of these early instruments—the Veteran/RAND-12 survey. The HOS was first applied to the Medicare Advantage population in 1998—with a baseline (Cohort 1) survey of enrollees that year and a follow-up survey on the same individuals two years later in 2000. The instrument has gone through several revisions since 1998, however it retains the Veteran/RAND-12 items—which are central in calculating several of the Medicare Stars measures. Items on the Medicare HOS include many of the eight domains from the original instrument, such as: general health, physical functioning, mental health, bodily pain, energy/fatigue, social functioning, and role limitations.

HOS is currently comprised of 68 items. The two HOS-derived *physical health* and *mental health component summary scores* are calculated from twelve separate items contained within HOS—all from the VR-12. There are another three measures derived from HOS responses and included in the Medicare Star Ratings measure list for Medicare Advantage Organizations (see Table 2).

Measure #	Measure Name	Primary Data Source	Improvement Measure?	Weight in Stars
C04	Improving or Maintaining Physical Health (PCS)	HOS	No	3
C05	Improving or Maintaining Mental Health (MCS)	HOS	No	3
C06	Monitoring Physical Activity	HEDIS/HOS	Yes	1
C18	Reducing the Risk of Falling	HEDIS/HOS	Yes	1
C19	Improving Bladder Control	HEDIS/HOS	No	1

There are additional condition-specific questions contained in HOS, such as self-reported functional status, disability status, and conditions/diagnoses (that a doctor has told the beneficiary he/she has). There are also questions on other beneficiary characteristics such as: race, primary language, income and education level, marital status, living arrangements, and home ownership. Not all questions are answered by the beneficiary. For example, the question on income is most often skipped. When the respondent provides this information, CMS can use this information to make some adjustments to the responses given—more about these adjustments are discussed further elsewhere in this White Paper.

HOS Cohort Sample - Medicare Advantage plans must participate in the HOS self-report survey process as a condition of their contract with the Centers for Medicare and Medicaid Services (CMS). Each spring a random sample (baseline) of Medicare beneficiaries is drawn from Medicare Advantage Organizations' enrollment lists (MAOs with a minimum of 500 enrollees). Independent survey vendor companies approved by CMS and contracted by the MAOs then receive a list of the beneficiaries CMS. Health plans pay for the survey, but the beneficiary/enrollment mailing list generated from the health plan enrollment file is not made known to the plans. The HOS survey vendor makes multiple attempts to contact these individuals by telephone or mailed request to conduct the survey (Baseline/Time 1). Two years

WHITE PAPER - A Focused Look at the Medicare Health Outcomes Survey

later, the same people are surveyed again (Follow-up/Time 2). Responses from baseline are compared to responses at follow-up to see if there are differences. The original list of names given to the vendor for each Medicare Advantage plan is 1,200 people. However, response rates to the request to participate in the HOS survey result in much lower actual sample sizes.

HOS-derived HEDIS measures in Stars: *Maintaining or Improving Physical or Mental Health*

The two most heavily weighted Star measures that are considered patient reported outcomes are the *physical health* and *mental health* component summary measures. Each is a composite score measure derived from six separate questions from HOS that the beneficiary answers about his/her physical or mental health status. The beneficiary responds based on his/her perspective and status as of the time the survey is conducted. An unexpected decline from the answers the person provided two years prior on these same questions will result in a poorer score. The twelve PCS and MCS questions are:

C04 – Improving or Maintaining Physical health (PCS) – “physical component summary”

Measure Description - This measure is derived from six items on HOS to derive a **physical health** composite score called the “Physical Component Summary.” This is defined as: *The percent of plan members whose physical health was the same or better than expected after two years.*

1. *In general, would you say your health is: Excellent, Very Good, Good, Fair, Poor*

*Does your health **now** limit you in these activities? If so, how much? (Scale: limited a lot, limited a little, not limited at all)*

2. *Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf?*

3. *Climbing **several** flights of stairs?*

*During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities **as a result of your physical health**? (Scale: none of the time, a little of the time, some of the time, most of the time, all of the time)*

4. **Accomplished less** than you would like.

5. *Were limited in the **kind** of work or other activities.*

6. *During the past 4 weeks, how much did pain interfere with your normal work (including work outside the home and housework)? (Scale: not at all, a little bit, moderately, quite a bit, extremely)*

WHITE PAPER - A Focused Look at the Medicare Health Outcomes Survey

C05 – Improving or Maintaining Mental health (MCS) – “mental health component summary”

Measure Description- This measure is derived from six items on HOS to derive a **mental health** composite score called the “Mental Health Component Summary.” This is defined as: *The percent of plan members whose mental health was the same or better than expected after two years.*

1. **During the past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)? (Scale: none of the time, a little of the time, some of the time, most of the time, all of the time)

Accomplished less than you would like

Didn't do work or other activities as **carefully** as usual

2. How much of the time during the past 4 weeks: (Scale: all of the time, most of the time, a good bit of the time, some of the time, a little of the time, none of the time)

3. Have you felt **calm and peaceful**?

4. Did you have **a lot of energy**?

5. Have you felt **downhearted and blue**?

6. During the past 4 weeks, how much of the time has your **physical health or emotional problems** interfered with your social activities (like visiting friends, relatives, etc.)? (Scale: not at all, a little bit, moderately, quite a bit, extremely)

Interpretation of an individual's PCS or MCS score as described by CMS in their Technical Notes:

Very high scores indicate absence of health concerns, and no limitations in usual social and role activities due to physical or mental health problems. A positive change (i.e. follow-up score higher than baseline) or no change in score from baseline to follow-up indicates an individual has improved or maintained their current physical or mental health.

Interpretation of un-adjusted aggregate plan results: *A higher rate is indicative of a larger number of beneficiaries whose physical or mental health was the same or better from baseline to follow-up on the PCS component of the HOS.*

Interpretation of adjusted observed-to-expected rate: *A higher rate indicates the plan had a higher proportion of beneficiaries whose physical health was same or better from baseline to follow-up than would be expected based on the adjustments made. A rate of 1.0 indicates the plan is performing exactly as expected based on the case-mix. A rate below 1.0 indicates the plan is performing worse than expected. A rate above 1.0 indicates the plan is performing better than expected.*

Challenges and Limitations of HOS with Special Populations

We have described the HOS instrument and current methods of administration, analysis and reporting in the previous section and more detail is found in the Appendix. In this section we explore the challenges and limitations of HOS, particularly focusing on special needs populations. These are:

- *Diversity of Medicare Population not fully Considered*
- *Instrument Limitations as Applied to People with Special Needs*
- *Inadequate Methods and Administration*
- *Information Decay, Time Lag*
- *Proxy Bias*
- *Results May be Affected by Underlying Characteristics of Enrollment*
- *Attribution, Context Issues*
- *Inadequate Models and Adjustments*
- *Cautions Regarding Analysis, Interpretation, and Reporting*

Diversity of Medicare Population not fully Considered - A fundamental issue with HOS is that this instrument may not be valid or reliable for a subset of the Medicare population which is increasingly diverse. The original respondent data set which was used to test/validate HOS

Limitation of HOS:

Insufficient dataset – Younger, disabled, diverse, female, dually eligible people may not be adequately represented

The original respondent data set which was used to test/validate HOS items and subsequent instruments (e.g., Veterans SF-36) primarily represents older (65+) men who served in the military. These data are 20+ years old.

items and subsequent instruments (e.g., Medical Outcomes Trust Survey, Veterans SF-36, Veteran-RAND VR-12) may have had insufficient representation of younger, physically disabled, dually-eligible individuals, ethnically/linguistically diverse non-US born persons, and women. The original testing population was primarily older (65+) men who served in the military. The MOS, VR-36 and VR-12 are based on these data, which are over 20 years old.

Many of the subsequent studies to adjust the instrument and validate its rigor continue to be based on the veteran population. While this is convenient, it is not a population that matches the demographic or other characteristics of the Medicare-Medicaid dually eligible sub-population.

Without adequate representation of these diverse subgroups, subsequent adjustment applied to control for population characteristic variables is unlikely to have much effect if it relies on distribution and data patterns in the original dataset to generate coefficients. Very small numbers of non-English speaking, younger people with disabilities, or frail elderly

women, for example, in the original dataset would not allow for robust development of adjustment coefficients for these groups, or tests to determine if such sub-population characteristics were associated with statistically significant differences in responses.

An example of the lack of heterogeneity of the underlying population is found in a 2004 study comparing the floor and ceiling effects of items on the Medical Outcomes Survey (precursor to HOS) with the Veterans SF-36 survey. The sample of individuals in this study who completed both surveys (therefore allowing the two responses to be compared) had the following

WHITE PAPER - A Focused Look at the Medicare Health Outcomes Survey

characteristics: 98% were male, 90% were over age 65, 81% were white, and 72% were married (Kasiz et al., 2004).

Instrument Limitations as Applied to People with Special Needs - HOS items and the wording, format, and length of the instrument can be challenging for people with special needs.

Wording - For example, HOS questions about climbing stairs, golfing, and performing household tasks such as vacuuming may not be perceived as relevant to persons with physical disabilities or frail elderly people. Such individuals may receive daily assistance for activities of daily living. People who use a wheelchair and other medical adaptive equipment due to a spinal cord or brain injury or degenerative neuromuscular disease may read these questions as not pertaining to them. The respondent may skip the item or may feel that the instrument was not thoroughly vetted for persons like them. This can affect not only the specific item response but elicit other negative reactions which can affect additional responses in the remainder of the survey, including the person's decision to end the survey prior to completion.

Wording and format issues may be particularly important in non-native born and non-English speaking persons. Bias can arise from language, wording, or methods of administration which do not properly accommodate the intended respondent. For example, non-native born people may not understand that the word "blue" is used to mean "sad." They may not have any experience with golfing or bowling as hobbies and therefore mis-interpret the question or conclude it does not pertain to them. The format and response scales can also be confusing.

In addition, question order has been shown to be important in non-English speaking respondents. Studies show there can be a difference in response for non-English speakers based on where the self-rated health question appears in the instrument—for example among Spanish-speaking persons (Sunghee and Grant, 2009).

Some Medicare Advantage health plan report that within their enrollment, beneficiaries speak over 100 primary languages. However, HOS is only available in English, Chinese, and Spanish. Special needs plans report a high level of linguistic diversity and low health literacy.

Limitation of HOS:

Some Medicare Advantage health plan report that within their enrollment beneficiaries speak over 100 primary languages.

However, HOS is only available in English, Chinese, and Spanish.

One health plan recently described their SNP enrollment profile as follows: *We have an average age of SNP members of 82 years, 76% have a high school degree or less, 7% speak a language other than English, 72% are single, 85% live in a rural area, and 41% have an income of \$10,000/year or less.*

In the 2017 *SNP Alliance Annual Member Survey*, member health plans were asked about the top social determinant of health risk factors observed in their enrolled SNP and MMP populations.

Plans were asked to involve their care managers and examine available internal data to respond to this question. The results corroborate many of the factors identified in other published studies.

WHITE PAPER - A Focused Look at the Medicare Health Outcomes Survey

The top SDOH factors reported by this sample of special needs plans were:

- low income/poverty status (80% reporting),
 - low health literacy or education level (73%),
 - lives alone or has few social supports (67%),
 - lack of available mental health services and supports in the community (60%),
 - housing instability/transience (53%),
 - transportation challenges (47%), and
 - food insecurity (40%).
- Additional SDOH factors mentioned included history of trauma, violence, and abuse.

We have an average age of SNP members of 82 years, 76% have a high school degree or less, 7% speak a language other than English, 72% are single, 85% live in a rural area, and 41% have an income of \$10,000/year or less.

(Source: *SNP Alliance Annual Member Survey 2017*, unpublished data).

As many who work with non-English speaking and non-native born populations, the issue is not only with translating an existing instrument/document, but in ensuring that the questions asked—in the way and format they are posed and with the response scales offered—are understood by these linguistically diverse respondents. This language barrier is a substantial impediment to beneficiaries being counted and included.

Another important issue is how the health beliefs, ethnic culture, and health literacy impact both the understanding of the item by the beneficiary and their willingness to admit their condition or limitations to a stranger. Some cultures frown on admitting challenges with mental or emotional health and may attribute difficulties to physical health problems, thereby indicating more problems on these questions, rather than the mental or emotional health questions. This introduces bias. The more that the respondent sample includes these diverse groups, the less we can be certain that the respondents accurately interpreted or responded to the question. This goes back to the challenges with the original population and dataset upon which the survey was designed and tested.

Inadequate Methods and Administration – Current methods of data collection used in the two-year longitudinal design of HOS may be inadequate for a special needs population.

Limited accommodation - Response and participation in HOS depends on having a mailing address, or regular telephone number and telephone/cell phone access. Survey vendors must reach the right person (beneficiary) within the window of time offered for administration. The survey (written in English, Spanish, or Chinese) is first mailed and then telephone attempt is made (for English or Spanish-speaking individuals). For these two languages the vendor may try to reach the beneficiary by phone and conduct the interview over the telephone. Telephone interviews are not available in any other language. Thus, current methods of data collection (telephone or mail) would make it very difficult for some people to participate. The limited accommodation in the survey methods serve as a barrier for persons with spotty/intermittent access to communication devices, who have housing instability, must move frequently, or do not read/speak English, Chinese, or Spanish. These are more likely to be low-income, ethnically diverse, and dually-eligible individuals.

Non-representative respondent sample - Health plans having a high proportion of their enrolled population as non-US born are more likely to have a significant number of beneficiaries who fall

WHITE PAPER - *A Focused Look at the Medicare Health Outcomes Survey*

in these categories. Since the health plan is prohibited from assisting the beneficiary in understanding the HOS instrument and real-time translation services are not available to potential respondents, these individuals may be unable to participate in the survey at all. The fewer people who participate in the survey, the less they are represented in the respondent sample. In this way, their participation in this aspect of the Medicare Stars rating and quality measurement process is restricted, and the dataset of responses is unlikely to include them, nor the resulting scores reflect their input.

Weak sampling methods/response rate- currently all plan members age 65+ are included (with a proposal to also include younger dually eligible persons 18-64) in the random sampling process to identify 1200 beneficiaries for contact. There is no oversampling of certain groups in the baseline sample—groups that would be less likely to respond. It is likely that the baseline and especially the follow-up final group of HOS respondents is skewed to English-speaking individuals, people who feel well enough to speak on the phone or read and complete a written survey, people who do not have significant cognitive impairment, and those with some health literacy. We do not know how well the sample mirrors the characteristics of the full enrollment profile of the health plan.

Longitudinal Design - The longitudinal “look-back” design of HOS is especially problematic for plans that serve a high proportion of very frail or medically compromised individuals. These individuals may experience significant health declines due to their medical conditions between the two measurement years. The HOS longitudinal design requires the same beneficiary be surveyed two years after the baseline survey. In plans with a high proportion of medically-compromised, advanced illness, and advanced-age populations, the drop-off or refusal rate, (including rate of death) can be quite high.

Special needs plans report that their HOS respondent sample is very low, with some plans reporting below 4%. For example, **Health Plan A** reports that of the 400 people returning the baseline sample (33% response rate), only 181 of these individuals returned the follow-up survey. This represents less than 3% of the plan’s total enrollment and an overall response rate of 15%.

Health Plan B reports that given the frailty, medical conditions, advanced age, and disease course of their I-SNP enrolled members, about 25% of the membership passes away each year. Thus, it is expected that half of the respondent sample will be lost to death between the two measurement periods. Furthermore, since in the PCS measure, a member who dies--where death was not predicted/expected by the model--is included in the “worse than expected” outcome group. Therefore, this plan is likely to score lower than plans with a healthier, younger membership profile on this PCS measure.

Health Plan C reports that 30% of their special needs plan enrollment is comprised of people with disabilities who are dually eligible given their low-

Focus Example: Sampling Issues

Health Plan A reports that of the 400 people returning the baseline HOS, only 181 of these individuals returned the follow-up survey. This represents less than 3% of total enrollment in this plan’s contract.

Health Plan B reports that given the frailty, medical conditions, and disease course of their I-SNP enrolled members, about 25% of the membership passes away annually. Thus, it is expected that half of the respondent sample will be lost to death between the two measurement years

Health Plan C reports that 30% of their special needs plan enrollment is comprised of people with disabilities who are dually-eligible given their low income (Medicaid) and significant physical disabilities (making them eligible for Medicare).

income status and physical conditions making them eligible for Medicare. This proportion of persons who are dually eligible with physical disabilities is much higher than general Medicare Advantage plans typical enrollment profile.

Information Decay-Time Lag Impacts Utility - There is a two-year lag time between the HOS survey years. After the second (follow-up) survey, the HOS survey results are scored for each health plan and across all MAOs. Following this final step, health plans are sent their individual HOS survey result reports. By the time the health plan receives respondent information, several years have gone by. The person may have left the plan, moved, or passed away (given that some plans have a high proportion of individuals of advanced age or illness). Because of this time lag, HOS results do not provide much current/useable information for plans to act or intervene to help the individual improve his/her health status or address functional, pain, emotional, or physical issues which were reported. Plans have indicated that they seek and use more timely alternative information sources, such as member outreach surveys, care management processes and notes, or other methods to help their members identify and report on current needs and engage members in self-care and health improvement opportunities. In describing the HOS reports received, health plans state they have not routinely compared the respondent sample characteristics of the HOS sample to the enrollment characteristics of their members from 3 to 4 years ago. Given the retrospective nature of the information, they are unsure what the value of that exercise would provide. They would rather attend to current member needs and characteristics and help craft quality improvement, care management, and provider responses using more recent information within the enrollment year.

Proxy Bias – Use of proxy respondents is likely to introduce bias. Situations where proxies are likely to be involved include:

- When the member's condition status prohibits participation directly. Persons with intellectual and developmental disabilities, moderate to severe cognitive impairment, or other neurological disorders, as well as advanced illness or frailty would have difficulty participating in HOS.
- When a family member or paid caregiver completes the verbal or written survey without the input of the respondent in order to reduce burden on the respondent.
- When a person who does not speak or read English, Spanish, or Chinese wants to participate in HOS they may rely on bilingual family members, including children, to translate.

Reliance on bilingual family members, including children, as translators is not recommended, as the individual's level of health literacy and expertise in translation may be inadequate.

Studies of proxy responses (by those other than the respondent individual) show that the proxies often have different opinions from the intended respondent. The direction of the proxy bias can be higher or lower (better/worse)—although often family and paid caregivers rate the person's ability lower than the individual respondent would when answering questions on self-rated health and ability. This is subjective and unpredictable.

Given that there is interest in extending the use of HOS to younger people with physical disabilities, it is important to recognize that research on self-report among persons with disabilities finds more variability in responses between time periods by the individual versus a proxy respondent (Sunghee, L., Mathiowetz, N, and Tourangeau, R., 2004). In other words, proxy responses are not capturing the beneficiary's self-reported outcomes or perspective.

The issue of proxy response is especially significant for health plans that specialize in serving people who are certified by state definition as “requiring a nursing facility level of care” (NFLOC). These individuals make up the entire enrollment of I-SNPs. Among these persons, cognitive impairment, functional limitations, and frailty is very high. Some I-SNPs report an average age of 86 within their enrollment. These individuals have multiple chronic conditions. Many may also have a limited life prognosis based on these conditions. It is not unusual for this population to experience high death rates in two years—from 25 to 50% of these individuals may die due to the progression of their diseases in-between the two measurement periods. This significantly impacts the follow-up survey rate. In addition, the HOS is often completed by a professional caregiver (e.g., personal care assistant, home health nurse) or a family member or significant other. The proxy respondent at baseline (T1) may not be the same person responding at follow-up (T2). Thus, the two time periods would be surveying different people—neither of whom is the beneficiary.

Results May be Affected by Underlying Characteristics of Enrolled Population – There is evidence that self-report of health status, such as in the HOS questions, can be impacted by a person’s current living situation and health conditions. Respondents may reflect on aspects their lives which have affected them recently as they answer questions about their health status. Things which are not under the control of the health plan or provider may influence their response, such as loss of housing, job, significant life partner, or living in a poor/unsafe neighborhood. These events are shown to cause stress and impact health status.

Associations with Lifestyle Choices - Several studies of the physical health and mental health measures (PCS, MCS) have documented perceived change of health and health-related quality of life in response to poor health behaviors, self-care, and lifestyle choices. For example, significant negative associations have been shown between smoking, obesity, and alcohol use—and the resulting PCS or MCS scores (Iqbal, et al., 2007). In other research among persons with marginal living circumstances, studies show that self-rated health status is strongly associated with their situation and living conditions (Pedersen, Groebaek, and Curtis, 2010).

Persons with Disabilities - Rowland et al. (2014) conducted a review of the literature to examine health outcome disparities within persons with disabilities. The research team found that the population of persons with disabilities is quite diverse and that there was little published research on key disparity factors, such as race and income, to document or address health outcome differences. According to research reported by the American Journal of Public Health, people with disabilities fare far worse across a broad range of health indicators and social determinants of health than their non-disabled counterparts (Krahn, G., Walker, D., and Correa-De-Araujo, R. 2015).

Characteristics that may affect health status and outcomes of younger persons with disabilities:

- Co-morbidities; including behavioral health and mental health disorders as well as complex chronic and medical/physical conditions
- Substance abuse disorder/chemical dependency
- Presence/absence of social support – social isolation
- Housing insecurity
- Living in a poor neighborhood
- Access to mental health and behavioral health services in the geographic area—such as in rural/remote regions where providers are few and far between

- Continuity of personal care assistant—long-term support/relationship
- Adapted transportation availability and accessibility – particularly after daytime hours and on weekends

To explore issues with applying HOS to younger people with disabilities further, we discussed the potential challenges related to certain mental health and other disorders which are more common among younger persons who are dually eligible for Medicare and Medicaid. HOS is being proposed to apply to these younger dually-eligible individuals, and therefore consideration of these issues is especially important and timely.

Clinician Perspectives on Mental Health – Some clinicians note that the mental/emotional health items on HOS pertaining to: “accomplishing less than you would like” or “didn’t do work or other activities as usual” could be affected by a change in the individual’s personal care assistant or family caregiver, or the loss of a close friend which could drive anxiety and affect usual daily routine.

The clinicians we spoke also to noted challenges in collecting this kind of information from persons who may have impaired self-awareness arising from their conditions. They posit that among the elderly on which the HOS was tested where there may have been greater stability across time periods, at least for the MCS measure. Among the elderly, geriatricians remark that changes in self-reported emotional or mental health are often impacted by depression and anxiety. These conditions may be more amenable to comparisons over two time periods through self-report, provided there is cognitive stability.

However, applying the HOS items that comprise the MCS measure to younger dually eligible people may be problematic. In this population conditions affecting emotional and mental health may be more prevalent, and there can be wider swings in condition (e.g. bipolar, psychosis, schizophrenia, and other related disorders). Comparing responses about self-reported mental health status over two time periods (the longitudinal design of HOS) may be especially problematic for these individuals.

Clinicians explained to us that a respondent with these conditions may not have the necessary self-awareness about his/her current status at a point in time. When they are euphoric, they report excellent health. Alternatively, they report very poor health when depressed. Extreme mood swings are one of the most common symptoms of bipolar disorder. Each phase can last 1 to 2 weeks or more. Life stresses are very likely to impact these individuals. Such stresses have been reported as one of the most common triggers of these severe mood swings. Unfortunately, individuals with this disorder may delay or avoid bringing information about recent life events or stresses to the attention of their providers. These disorders can affect mood and mental awareness in substantial and unpredictable ways, even when under treatment.

Severity and Types of Conditions - There are additional condition complexities which can impact physical and mental health functioning over time. Changes in the individual’s condition due to disease—for example degenerative neuromuscular diseases such as Multiple Sclerosis, ALS, and Parkinson’s disease—would also affect an individual’s ability to accomplish what they would

Limitation of HOS:

Personal and condition characteristics affect responses on health status.

Life events and disease impacts health. Some characteristics which influence a person’s perception of health status may not be able to be addressed by providers or health plan, despite offering the best treatment and care.

like or “interfere with social activities.” These diseases, and the stage of their severity are not listed on the HOS. However, they are major conditions and should be considered.

Similarly, regarding the physical health items, such degenerative conditions would likely affect the individual’s ability to perform moderate activities such as “moving a table, pushing a vacuum cleaner, bowling, or playing golf,” and “climbing several flights of stairs,” particularly over the two time periods for data collection. Such neuromuscular and degenerative diseases as well as other conditions such as Alzheimer’s disease, can also affect recall and emotional health status. This is vital information for understanding responses provided by the individual.

Palliative or End of Life Goals - For some people, maintaining or improving health status is not their goal. Individuals with a terminal disease, on palliative care, or in hospice may be focusing on the quality of their interactions with their loved ones, not on maintaining or improving their health status or performance on activities of daily living. They have reconciled themselves to the reality of their medical condition. They might find the implications or suggestions given by the wording of the questions to be judgmental or focused on the wrong things. For these individuals, HOS may not adequately capture their perspective or accommodate them in the survey or methods of analysis. There are not exclusions from the measure application for people who are receiving palliative or hospice care. *Furthermore, persons who die between the two measurement periods are still included in the final HOS measurement results.* The lack of exclusions or exceptions for this measure due to these end-of-life and palliative care considerations seems particularly disadvantageous to health plans that serve a high proportion of people with these considerations.

Attribution, Context Issues - People with complex chronic and degenerative, behavioral or mental health conditions should receive the best standard of treatment and support. Plans and providers serving these individuals need to coordinate care and support across settings and over time.

Attribution - When people with these conditions respond to questions about their physical and mental health status, they may reflect on the progression of their condition(s) rather than what a plan or provider has done or is able to do. These individuals and the health plans and providers serving them may be limited in how much they can overcome or counteract the degenerative and cumulative effects of these diseases. Individual engagement and attitude are two factors which can make a difference, in addition to seamless access to services, care, and treatment. With questions about health status, it is difficult to know with certainty what can/should be attributed to excellent or poor care by the provider team and health plan care management and what should not. This issue of attribution is critical to evaluate further.

Context - If all populations enrolled across all Medicare Advantage plans were similar, the lack of information (context) regarding what might be affecting physical or mental health status independent of provider or plan action would not be an issue—since the effects of these kinds of conditions or characteristics on self-report of physical or mental health status would be spread equally across all plans within the data received. If, however, health plans have very different populations of persons enrolled—in terms of disease burden, social determinants of health risk

Limitation of HOS:

Certain major complex degenerative conditions and condition severity are not included in HOS

Major degenerative neuromuscular and degenerative diseases such as Multiple Sclerosis, ALS, Parkinson’s disease, Alzheimer’s disease, and condition stage or severity are not listed.

factors, care complexity, or other characteristics—then the effect of these issues on self-reported health status may be unequal. It is reasonable to presume that there are more substantial negative impacts on health plans serving a high proportion of people with complex needs when compared to health plans serving healthy or higher income individuals without such condition complexity or social risk burdens.

Inadequate Models and Adjustments - The steps used by CMS to adjust HOS survey information received from beneficiaries are described in detail in the Appendix. As mentioned, for the PCS and MCS measures, the health plan is considered accountable for the person’s status compared to two years prior based on the predictive models. The more people who report improvement, the better the plan’s overall score. An unexpected decline from answers given by the beneficiary compared to two years prior around these physical and health mental health questions will result in a poorer score for the PCS and MCS measures, unless further adjustments are performed. The greater number of respondents who report decline, the worse the score for the health plan on these two measures.

Some of the limitations from the methods used include:

- lack of sufficient or complete data
- missing variables known to impact health outcomes (not in model)
- best fit models based on outdated information
- small sample sizes
- lack of inclusion – sample of final respondents may not be representative of the enrollment profile

Limitation of HOS:

The predictive “best fit” models may be inadequate because they:

- *Are missing information from the survey (skipped items)*
- *Do not capture severity of condition*
- *Do not consider additive effect of multiple characteristics*
- *Miss important additional variables impacting health outcomes*

A person who has a degenerative, progressive health condition (e.g., ALS, Alzheimer’s disease, Multiple Sclerosis) or a behavioral health condition (e.g., psychosis, bipolar disease) which affects physical or mental functioning—may decline more in the two years than predicted by the models used. This could be because the models are missing data elements (skipped), they do not adequately capture severity, they do not consider the additive effect of multiple characteristics, or they miss important factors (additional variables not in the model) which have an impact on decline or improvement.

Limitation of HOS:

Many of the sociodemographic characteristics which are used to “case mix adjust” responses and are used to calculate predicted health status are at the end of the instrument and are least likely to be answered.

Questions on race, marital status, home ownership, language, education, income are all at the end of the survey.

A notable limitation of both the PCS and the MCS “best fit” models is that *calculations to categorize the beneficiary are based on those case-mix variables for which the beneficiary has provided data in the baseline HOS survey.* Unfortunately, *many of the sociodemographic questions are at the end of the survey.* These are the items least likely to be answered. The fewer questions the person answers on these items, the less data available to make these case-mix adjusted predictions. The less the data available, the fewer case-mix adjustments made, the less likely that the predictions will match actual results and that the scoring will be accurate. This is a potentially serious flaw.

WHITE PAPER - A Focused Look at the Medicare Health Outcomes Survey

Another concern about the predictive models is that the self-report surveys do not provide information about the extent/severity/stage of the person's condition(s). The conditions listed simply have Yes/No responses, even for cancer.

The best fit models used by CMS are to accurately predict health status outcomes for people from baseline to two years later. An underlying fundamental assumption is that these health status outcomes can be addressed by actions of the person's providers and health plan. That is, what the provider or health plan does can mitigate negative impact or condition progression. Mounting evidence refutes this presumption, particularly when social risk factors are considered. Evidence suggests that life events, such as the loss of a job, loss of a life partner, loss of home, or a severe medical event of a close family member may also impact physical and mental health status over two years. Information on these significant life events which may have occurred in the past two years is not captured by the current instrument.

Further Data and Cautions from the Field

In the most recent SNP Alliance Annual Survey, a set of health plans reported their HOS baseline sample, follow-up sample size, and the percentage of their total enrollment within that contract that the HOS respondent sample represented. Data provided shows most plans' HOS respondent sample sizes (which were used to calculate the PCS and MCS scores) were very low—most well below 10% of total enrollment. Response rates from the original random sample (usually 1200 persons) to the final respondent sample ranged from a high of 30% to a low of 9%. The drop off from baseline to follow-up was also very substantial; often more than 50% of persons did not respond to the follow-up survey. This underscores the points made earlier in this White Paper about the two-year longitudinal design and the data collection methods being inadequate to ensure a valid, and representative sample needed to support accurate scoring. Table 3 provides more detail.

Total H# Enrollment	Baseline Respondents (sample size is usually 1200)	Follow-up Respondents (% of original sample)	% of Total plan/contract enrollment that the final PCS and MCS scores were based upon for these plans
8488	413	204 (17%)	2.4%
3139	1042	227 (19%)	7.2%
11159	418	262 (22%)	2.3%
1813	941	362 (30%)	20.0%
11085	144	104 (9%)	0.9%
8379	399	181 (15%)	2.2%
10699	537	123 (10%)	1.1%
7645	187	133 (11%)	1.7%
12514	225	168 (14%)	1.3%

Source: SNP Alliance 2018 Annual Survey, unpublished data.

SNP Alliance health plans provided additional qualitative information and comments about the challenges and limitations they see with the PCS and MCS measures generated through HOS:

In addition to lack of representativeness of the sample used and the different characteristics of the population from one health plan to another which may not

WHITE PAPER - A Focused Look at the Medicare Health Outcomes Survey

be addressed by the case mix adjustment as applied, we have several additional concerns about the validity of the HOS methods. We suspect the survey methods may not be valid for year-to-year comparisons in this population. Evidence and concern arise from:

- 1) Values for both the PCS and MCS scores fluctuate by up to 8 percentage points from one year to another, though we have maintained a high degree of care management and service and have even enhanced our interventions over these periods.*
- 2) The risk adjustment process for SNP plans may not be computing a fair expected rate of death or of getting worse over time when calculating the PCS and MCS scores for our population which is of advanced age.*
- 3) Proxy participation in HOS is not included in the risk adjustment models for the longitudinal health status measures. Given that 30-40% of our member surveys are completed by proxy for our SNP population, we are concerned about residual confounding. We understand that the HOS cohort design allows for members to serve as their own controls and that HOS protocols emphasize completion of the survey by the same respondent/respondent type (member or proxy) at baseline and follow-up. However, 30% of our SNP sample has a different respondent at baseline compared to follow-up; and the current methodology makes no apparent adjustment.*
- 4) Assessing for improvement in physical and mental health over two years may lack face validity even with a strong risk adjustment model when population complexity—conditions, social risk factors, and interaction between mental and physical health issues such as exists in the dually eligible SNP population—are not fully taken into account.*
- 5) The very small cohort samples we observe increases potential for bias, outliers, and inaccuracy.*

With evidence of the limitations in the HOS instrument, sampling, predictive modeling, measure weighting, and scoring, special needs plans and providers serving vulnerable populations are more likely to receive negative scores driven by the characteristics of the population they serve, rather than the quality of care they provide.

Other experts have cautioned about the potential negative and unintended consequences of limitations in quality measurement and scoring for Medicare providers and health plans. In their 2016 Report to Congress, the ASPE technical expert committee found:

In every care setting examined, providers that disproportionately cared for beneficiaries with social risk factors tended to perform worse than their peers on quality measures. Some of these differences were driven by beneficiary mix, but some of the difference persisted even after adjusting for beneficiary characteristics. As a result, safety-net providers were more likely to face financial penalties across all five operational Medicare value-based purchasing programs in which penalties are assessed, including programs in the hospital, physician group, and dialysis facility settings. They were also less likely to receive bonuses in Medicare Advantage.⁵

Options to Address Limitations

There are potential strategies which would address some of the limitations of HOS. Options range from those that require only a change in analysis and reporting to those requiring more extensive modification and investment. Strategies and options should be considered for modeling and testing to improve accuracy, fairness, utility, and reliability. Options include:

1. Level the measurement field
2. Use additional exceptions and exclusions
3. Discontinue the longitudinal design
4. Use better predictive models and data sources
5. Use alternative instrument(s)
6. Re-test HOS with disabled and diverse groups
7. Improve methods of administration and accommodation
8. Conduct analysis at a finer level and report on results
9. Provide greater transparency and consumer education
10. Convene ongoing stakeholder and expert panels

The first three options on this list maintain the use of HOS but make modest adjustments. These three steps could be implemented quickly, at least on an interim basis. They offer promise for relief to address unequal burden and inequities arising from the instrument, methods, scoring, and adjustment methods.

The next two options (#4 and #5) would require more time and testing, however there are robust alternative data sources and instruments already available and experts in measurement science to help guide testing on these alternative approaches.

The rest of the options would continue with use of HOS but deploy greater analysis and effort to adjust limitations that do not adequately reflect/consider the diversity within the Medicare population and population differences across MAOs. These actions #6 through #10 also offer opportunity for stakeholder and expert involvement in making measurement, method, and scoring improvements, sharing information, increasing transparency, and improving value (utility and relevance) of the measurement results for consumers.

#1 Level the Measurement Field

To level the measurement burden on MAOs and recognize the unique characteristics of SNP populations, CMS could discontinue application of the two (PCS and MCS) HOS Star measures for Special Needs Plans, given the characteristics of the SNP population which are different than general Medicare population.

In exchange, CMS could substitute 2 of the 4 SNP-specific measures for the PCS and MCS measures that relate to the issue of health status and care management (e.g., C10 *Care for Older Adults-Functional Status Assessment* and C09 *Care for Older Adults-Medication Review*).


General MAOs (non-SNPs) do not have to report on these two SNP-only measures; currently, the burden of measurement for 4 extra Star measures only falls on SNPs. Thus, by eliminating the PCS and MCS from SNP rating and substituting the C09 and C10 measures instead, SNPs would have only two extra measures under Stars rather than four.

WHITE PAPER - A Focused Look at the Medicare Health Outcomes Survey

The HOS survey could still be conducted—information on results could still be provided to SNPs as it is with other MAOs. However, SNPs would not receive a Star rating on the PCS or MCS measure. This would even the measurement burden across MAOs and still offer some information to all MAOs on how a sample of their enrollment self-reports on their health status

This could be implemented by CMS as an interim step while improving HOS or identifying more robust patient-reported outcome measures. In a survey of special needs plans by the SNP Alliance, this is the top recommended action step requested.

- *Additional Option:* Stratify MAOs by quintiles based on proportion of enrollment LIS/DE/Disabled. Separate the MAOs into these 5 cohorts for 5 cut-point thresholds to be set for each measure; reporting on the PCS and MCS measures and assigning Star levels would be by quintile cohort. Table 4. below provides an illustration.

	PLAN Cohort A	PLAN Cohort B	PLAN Cohort C	PLAN Cohort D	PLAN Cohort E
	Low Vulnerability				High Vulnerability
Proportion of Contract (H#) DE/LIS/Disabled:	0-20%	21-40%	41-60%	61-80%	81 – 100%
C04 (Physical health composite)	Range, avg scores in this cohort of plans	Range, avg scores in this cohort of plans	Range, avg scores in this cohort of plans	Range, avg scores in this cohort of plans	Range, avg scores in this cohort of plans
C05 (Mental health composite)	Range, avg scores in this cohort of plans	Range, avg scores in this cohort of plans	Range, avg scores in this cohort of plans	Range, avg scores in this cohort of plans	Range, avg scores in this cohort of plans

Using the plan cohort groups illustrated above, CMS could **set 1 through 5-star thresholds for each cohort**, rather than have one national rating distribution. This approach would provide a much cleaner picture of performance to account for important beneficiary characteristics known to impact health outcomes. In addition, this approach would provide more meaningful benchmark comparisons for plans in each cohort group.

- *Alternate Option:* Revise the weights down to “1” for PCS and MCS measures for SNPs, given the characteristics of the population which are different than general Medicare population.

#2 Use Additional Exceptions and Exclusions

With any outcome measure on health status there are medical condition issues which will clearly impact self-report of health. If HOS continues to be used in Stars with PCS and MCS as outcome measures where an unpredicted change in status is attributed to plan or provider action, then we strongly recommend CMS add exceptions and exclusions in the measure specification of the PCS and MCS measures.

WHITE PAPER - *A Focused Look at the Medicare Health Outcomes Survey*

For example, persons could be excluded from the denominator and therefore removed from the list of names to be considered for HOS baseline or follow-up survey if the following is true:

- The person is on/in hospice care or palliative care, where a physician has indicated terminal prognosis and death within one year is a reasonable prognosis.
- A physician has diagnosed a degenerative neurological or neuro-muscular disease at a stage which is advanced or expected to advance thereby severely limiting physical or mental health functioning within one year. Specific condition staging or ICD-10 codes could be set that are indicators of this criterion to guide identification for people matching this criterion.
- A physician has diagnosed a physical or mental condition with a stage indicating very poor or severe loss of physical functioning and/or cognitive impairment which is not expected to reverse course within one year. This would include, for example, late stage cancers, late stage Alzheimer's disease. Specific condition staging criteria could be set that are indicators of this criterion to guide identification for people matching this criterion.
- The person is in intensive care, therefore would be removed from the survey roster. If a follow-up survey was to be mailed, this person's baseline data would be removed from PCS or MCS measure scoring.

We urge CMS to confer with clinical and other experts to review measure specifications for the PCS and MCS measures and make recommendations for additional exclusion and exception criteria with clear indicators and criterion.

#3 Discontinue the longitudinal design

The longitudinal two-year look-back design using HOS to create the PCS and MCS measures is particularly problematic with a frail, complex, chronic care population and those with significant SDOH risk issues. We recommend dropping the two-year design.

A single year survey with a representative random sample of current beneficiaries reporting on current health status would accomplish several objectives and address some of the limitations or weaknesses in how HOS is being used in the Medicare Advantage Star ratings. The single year approach would:

- Improve/increase the sample size
- Improve the timeliness issue
- Allow for more accurate comparisons of the HOS survey group to the total enrollment (e.g., age, gender, race, LIS/DE/Disabled status, language, etc.) to ensure that the sample is representative of the total enrollment population
- Allow for improvement in case mix adjustment across MAOs so that population differences (e.g. # of chronic conditions, medical or behavioral health complexity, age, income status, etc.) that can strongly influence health outcomes are considered. This single year survey data could be combined with other MAO

WHITE PAPER - *A Focused Look at the Medicare Health Outcomes Survey*

enrollment data that is more recent and comprehensive to improve the predictive modeling and case mix adjustment conducted.

- Provide for the opportunity to use HOS self-report health status results for quality improvement, identifying gaps and strengths

#4 Use Better Predictive Models and Variables/Data Sources

To correct for limitations in the current case mix adjustment methods, CMS should consider using additional data sources (other than the HOS instrument) to have sufficient data and additional variables for adjustment. This would best be combined with eliminating the 2-year longitudinal design. Several steps are offered for consideration:

- Rather than relying on the HOS instrument for the source on data elements for case-mix adjustment, other more complete source data with additional data elements could be used to create a baseline respondent profile. Key characteristics such as number and type of diagnoses, presence of major chronic conditions, disease/condition severity, utilization including use of LTSS services, claims, diagnostic codes, administrative data, enrollment data, and neighborhood level census data could be used to determine indicators of poverty, social risk characteristics and care complexity. The algorithm/regression model would be adjusted to include these additional variables with data on all baseline respondents as a group. Testing on diverse subgroups would be very important to determine if there are significant differences across subgroups that need to be taken into account for accuracy and predictive validity. The model could be refined in various ways to yield the best predictive model and this could be tested with each follow-up survey to determine which variables are most predictive, which variables are highly correlated, and how this may differ across different population subgroups. The current “best fit” models using HOS-generated data could be compared to these alternative regression best fit predictive models. Such comparison will provide important information about the adequacy of the current model and how it could be revised/improved.
- Another step is to use more recent US national population data. Rather than the 1990 US national population data, CMS would be advised to update the model with the last decade Census data to get closer to current population diversity characteristics in order to re-set averages/midpoints in their regression model(s)
- Annually, with each HOS cohort, CMS could publish the results of the predictive “best fit” models (comparing actual to predicted) for distinct beneficiary groups, starting with the DE/LIS/Disabled beneficiaries. This is an important action step toward transparency and accountability to demonstrate the predictive validity of the model on an ongoing basis.

#5 Use Alternative Instrument

Alternate physical and mental health status self-report items, currently in use and validated, could be used as substitute questions to the HOS 12 items used in the PCS and MCS. We urge CMS to consider this.

WHITE PAPER - *A Focused Look at the Medicare Health Outcomes Survey*

For example, the PROMIS Global Health measure V.1.2 (Fixed Length Short Form) is a robust measure set with 8 items that has several advantages, including wider use, robust testing and ongoing instrument maintenance with a team of measurement science experts under contract with the NIH to expand and automate use. There is an active and growing user community using PROMIS®, Neuro-QoL, ASCQ0Me®, and NIH Toolbox® to assess physical, mental, and social health along with sensory, motor, and cognitive function. The instrument is being widely used in intervention studies and comparative effectiveness research.

The Global Health instrument is already translated into 16 languages and is free to download.⁶ Furthermore testing to compare and crosswalk the PROMIS Global Health instrument to the VR-12 has already been done (Schalet et al, 2015). The researchers found the two instruments examined similar items. The PROMIS Global Health instrument produces two scores—the Global Health Physical Health (GPH) and the Global Mental Health (GMH) which are each based on four items.

A precursor to the global health items used is the Dartmouth COOP assessment tool, which is a pictorial version of scales for self-reporting physical health, functional health, mental health, and pain assessment such as developed and tested/validated (COOP charts) have been shown to be very useful in primary care settings.⁷ The COOP is a simple, reliable health related quality of life tool (HRQL) which can be used together with other patient-reported outcome measures. Studies of chronic care populations comparing outcomes using the SF-36 and the COOP chart showed high correlation in the constructs of health status across the two instruments, but the COOP charts were faster to complete and score.

These are just two examples of valid, reliable instruments measuring similar constructs which are widely in use and could replace the HOS instrument. We have not been involved in any testing of these instruments. Any alternative chosen would have to be piloted within a representative group of MAOs to fully understand the feasibility, utility, and accuracy of the alternate instrument and methods tested, the means for population difference case mix adjustments that the alternative instrument/method would support, and how superior the alternative would be to the current longitudinal design using HOS.

#6 Re-test HOS with Diverse and Disabled Groups

The HOS instrument is due for retesting to consider need for instrument items and scaling modifications among more diverse and disabled beneficiaries. Characteristics of beneficiary sub-groups that appear to be important include: low health literacy, linguistic diversity, non-US born, frail elderly, younger people with physical disabilities, persons with high social risk factors, such as low-income, and people with significant mental or behavioral health conditions. The PCS and MCS questions need to be retested among these population sub-groups. Results on the such testing should be published. We need to know how HOS could be improved in terms of item wording, response scale, question order, and to better understand how methods of administration/data collection could be improved to better accommodate this diversity. This would inform how to modify HOS and administration to ensure understanding, meaning, accuracy, validity, and reliability for these diverse subgroups.

#7 Improve Methods of Administration and Accommodation

CMS could improve methods of administration and better accommodate non-English speakers and those with limited health literacy. For example, vendors administering the survey could be

required to provide translation services to persons receiving the survey when the beneficiary requests such assistance. Instrument administration could be conducted through a series of 1:1 secure online facilitated and interactive sessions.

#8 Conduct Additional Analysis at a Finer level; Collect information on successful improvement efforts to build capacity for targeted quality improvement

CMS could conduct additional analysis around beneficiary subgroups and specifically examine health plans that showed steady improvement of PCS and MCS scores over several years. This would provide additional information around beneficiary subgroups (characteristics of the beneficiaries) who are achieving better health status and around health plan actions taken that are showing quality improvement performance results. The goal would be to provide guidance on what can positively influence physical and mental health status of different beneficiary subgroups—to engender meaningful quality improvement for these different groups and support a learning system.

#9 Provide Greater Transparency and Consumer Education

CMS could improve the distribution of information around HOS to reach more beneficiaries, improve consumer understanding, provide greater transparency and offer more meaningful/robust information to the field. For example:

- Translate the survey into other languages. If no additional translation is done, then make it clear when reporting measure results that the HOS survey is only in writing for English and Spanish speaking people and is only verbally offered in English, Spanish, and Chinese.
- Request every health plan to send all enrollees a brief information sheet explaining HOS, translated into the language spoken by the individual.
- Provide a brief information pamphlet about the HOS survey, in several languages, for physicians to give to their patients to be distributed the month prior to the survey administration and kept in waiting rooms.
- Report the sample size of the respondent data used in the PCS and MCS measure and show the full enrollment size (baseline and follow-up) and resulting response rate.
- Offer webcasts and podcasts to the general public in many languages describing HOS, the purpose, and walking individuals through the survey (video). Provide a Q&A or chat box function.
- Offer webcasts and podcasts with links on Medicare Plan Finder to walk the general public through the HOS survey and guidance on how to interpret the score results

#10. Set up Feedback and Expert Groups to continue to Advise and Inform

CMS could:

- *Stakeholder Focus Groups* – We urge CMS to consider conducting focus groups to capture stakeholder input. This would include engagement of younger dually eligible

WHITE PAPER - *A Focused Look at the Medicare Health Outcomes Survey*

persons, their providers and personal care attendants or families, special needs health plans which have a high proportion of these individuals enrolled, researchers who are familiar with this target population, providers serving persons with care complexity and severe, ongoing, and disabling chronic conditions.

- *Consumer Advisory Panel* - An annual convening of consumers which includes DE/LIS/Disabled persons, frail elderly, linguistically diverse/non-US born persons could serve as an Advisory Panel. This would help inform an annual review of methods, instrument, and results and provide feedback on information sheets, FAQs, podcasts, webcasts, and other educational/ informational consumer-focused materials as well as HOS surveyor training, translation services, and guidance around HOS. Such direct consumer guidance is often very informative.

This Panel or focus groups could offer invaluable feedback to tailor reporting for diverse subgroups who may not find the national information sufficiently tailored or specific to their local area, their situation, or the plan product in which they are enrolled. Questions consumers feel is important when evaluating options and information on quality measurement offered include:

Is this information for my region? Consumers are likely to prefer having quality measurement ratings and information from smaller reporting units—which would assist them in understanding the local/regional experience.

Characteristics of the people enrolled – Are they like me? - Consumers with specific complex chronic conditions or other care complexity that involves multiple treatments and services might wish to review quality measurement information on the plan product that is most meaningful/relevant to them based on their conditions, rather than a composite of several plan products of different types and sizes including people who are primarily healthy. Consumers with special conditions or characteristics want to find and compare quality outcomes achieved for people like them—to accurately assess if a health plan and provider network is doing a good job serving these individuals.

This also pertains to people with other defining characteristics that affect care seeking and delivery—such as being non-native born, speaking a primary language other than English, or having a low level of education/literacy. These individuals may wish to have more information about how the health plan quality measurement ratings differ by these characteristics.

- *Expert Measurement Panel* - Create an ongoing panel of researchers and industry experts in measurement and chronic care populations to review results and confirm validity/accuracy.

Conclusion

The SNP Alliance recognizes challenges in identifying, developing and using appropriate self-report measures as part of performance evaluation for organizations. The gold standard for a quality measurement, evaluation, and reporting system is to have measures that meet high standards around accuracy, validity, reliability, feasibility, and utility and are deployed within systems with robust data collection and administration that ensures equity and fairness. This is especially hard to achieve when the characteristics of the populations served by the organizations are quite different. While assessing self-rated health status is very important and a useful indicator to collect, there are challenges in comparing two populations that are significantly different. These challenges are multiplied when trying to make these comparisons over years.

Through our analysis we've identified shortcomings with the use of the HOS instrument in the Medicare Stars ratings for Medicare Advantage health plans in the way it is currently deployed. We've also identified and offered options for addressing some of the limitations.

Though the HOS instrument was well-developed within the 65+ White U.S. veteran population and has been shown to be valid and reliable for this population, the instrument needs retesting in diverse population subgroups. The demographic characteristics of the original and subsequent study populations do not match well the characteristics of diverse subgroups enrolled in Medicare Advantage plans.

Health plans with a high proportion of individuals enrolled with the following characteristics are more likely to experience the limitations in the HOS measures and methods:

- Dually-eligible, disabled, multiple major chronic or significant degenerative conditions, such as ALS, Parkinson's disease, or advanced Alzheimer's disease;
- Significant social determinant of health risk factors such as: Non-English speaking, non-native born, low-income, housing transient, food insecure, lacking a regular source of transportation, communication device variability, low education/low health literacy;
- Significant life events, such as loss of a partner, home, job, or a new diagnosis of a significant disease over the last two years;
- Persons who have a terminal diagnosis or are on palliative care where goals and expectations differ around maintaining physical health status

From sociological, public health, and health services research, we know that life events and personal characteristics can affect an individual's health as well as his/her perception of health status. The person's perception of health status is what is being measured in the PCS and MCS HOS items. Without adequately taking into account contextual issues, beneficiary-specific condition and complexity characteristics, and social determinant of health risk factors which impact outcomes, bias or inaccuracy is introduced. The veracity of the resulting quality scores may be questioned.

The lack of accommodation in methods and in the instrument, as well as the lack of stratification in scoring and reporting results contributes to the problems. The lack of translated surveys and the format, wording, and order of the items included in the HOS instrument may affect response—especially where the enrolled population is non-US born and speaks a language other than English, Chinese, or Spanish.

WHITE PAPER - *A Focused Look at the Medicare Health Outcomes Survey*

Survey methods/administration, analysis and reporting, are as important as the instrument itself. Sampling, data collection, verification, aggregation, stratification into meaningful cohorts, adjustment, analysis, interpretation and reporting have areas for improvement. The two-year longitudinal design is especially problematic.

We find some limitations in the predictive models and in data used for case mix adjustment. Predicting decline, maintenance or improvement in health status for diverse populations demands robust models and adequate data as well as periodic retesting to confirm accuracy and validity. HOS case mix adjustment and predictive best fit models may not have sufficient data or the full set of variables to adequately predict health status change across two time periods for very different populations.

Inequity in scoring could have unintended consequences. One consequence of this could be that the plans who are most focused on at-risk, vulnerable, and diverse individuals receive lower quality ratings and no quality bonus.

For these reasons, the SNP Alliance believes that further foundational work is needed to ensure that the instrument and measures are relevant, valid, accurate, and reliable for the intended population and that the methods adequately accommodate the diversity and characteristics of the beneficiary respondents.

We have offered potential solutions and strategies—from those that would require modest changes in administration, analysis and reporting, to those that would require more extensive modification or new measure and methods development. There are many viable options for improvement.

The SNP Alliance is working with a coalition of stakeholders around some of these issues. The participation of consumers, plans, providers, researchers, regulatory agencies, and policy makers is very important. This work has important implications to improve the value of performance measurement and reporting under Medicare Advantage particularly to diverse and special needs beneficiary groups.

APPENDIX A: Medicare Advantage Quality Measurement

Brief Overview of Medicare Star Ratings⁸ - The Centers for Medicare and Medicaid Services (CMS) established and oversees the Medicare Advantage Quality Management System including Part C and Part D Star Ratings. CMS has described the goals⁹ of its Quality Management System as follows:

- To display quality information on Medicare Plan Finder for public accountability and to help beneficiaries, families, and caregivers make informed choices by being able to consider a plan's quality, cost, and coverage;
- To incentivize quality improvement;
- To provide information to oversee and monitor quality;
- To accurately measure and calculate scores and stars to reflect true performance;
- To recognize the challenges of serving high risk, high needs populations, while continuing the focus on improving health care for these important groups.

Medicare Star Ratings & Measures¹⁰ - There were 48 measures within the Medicare Advantage Part C and Part D Star measure set in 2018, and these were divided into five categories:

- 1) Outcomes Measures
- 2) Intermediate Outcomes Measures
- 3) Patient Experience Measures
- 4) Access Measures
- 5) Process Measures

Measures rely on data from four sources:

- Health Plans and Drug Plans
- Survey of Enrollees
- Data Collected by CMS Contractors
- CMS Administrative Data

Each measure has specifications to define the numerator, denominator, exclusions/exceptions, and any adjustments that should be made. CMS has assigned a weight to each measure from 1 to 5. The improvement, outcome, intermediate outcome and patient experience measures are weighted highest. Process measures are weighted the lowest. Half star points are also identified for the overall Star rating for the health plan.

Scoring- Health plans are reviewed and scored at the contract level. Many health plans have multiple plan products within a single contract. The whole contract is the unit of analysis and reporting. The contract can span regions and even multiple states.

CMS assigns a numeric score for each measure. Then this numeric score is converted to a Star score for that measure for each health plan. This is done by applying one of two methods—so that the plan (contract)

Limitation of Medicare Stars Information

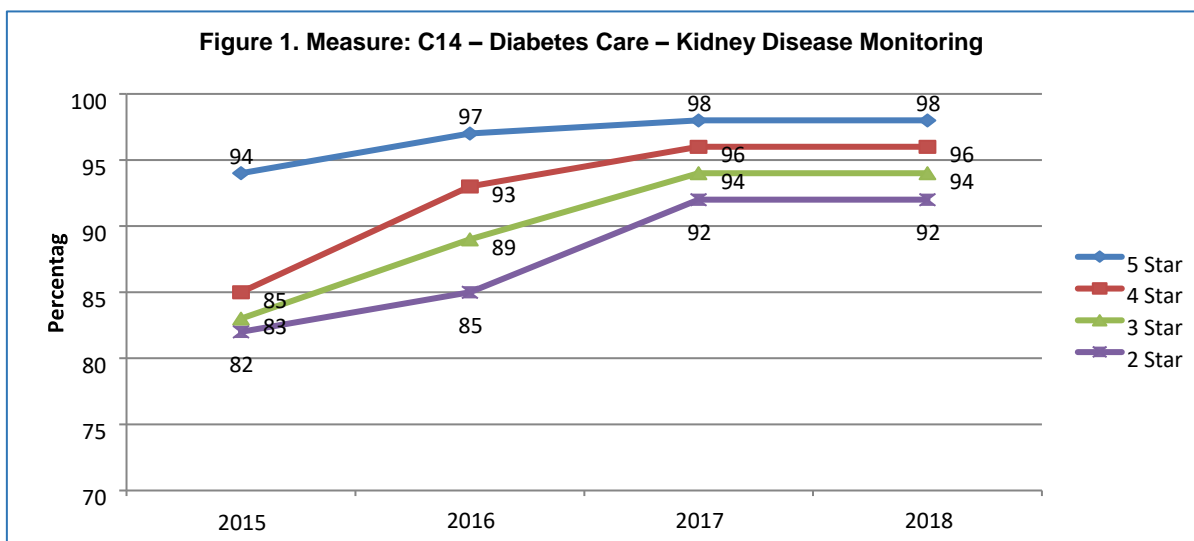
There is one national average of scores. This hampers utility; subgroup analysis and information are lacking.

Such information is necessary for meaningful benchmarks and quality outcome comparisons.

score will be compared to the measure scores from all other plans (contracts) under Medicare Advantage.

The two methods are: (1) clustering, and (2) relative distribution and significance testing. In the first method, CMS uses a clustering algorithm that identifies gaps in the distribution of scores for specific measures across reporting MAOs. The contracts are grouped such that scores within the same Star Rating category are as similar as possible (i.e., the 2-star, 3, star, 4-star, 5-star plan groups should have similar scores). CMS states that with this algorithm, scores in different star level categories should be as different as possible.

Despite this methodology, some measures show very little variation across all scores—which means that the gaps in scores are very small between Star levels and across all health plans. An example shown in Figure 1 depicts the 4-year trend of the differences in measure C14 “Diabetes Care – Kidney Disease Monitoring.”



Source: CMS Cut point Trends Report 2018.

This figure shows the spread between the best (5-star) plan and the lowest (2-star) plan was only 6 percentage points—almost all plans had achieved more than 90% effectiveness on this measure in 2017 and 2018. This is called a “topped out” measure. Topped out measures have caused concern among advocates, quality measurement experts, and plans. This is because when almost all organizations are achieving a very high rate of effectiveness, the differentiation between organizations scoring the lowest (2 stars) and the highest (5 stars) is very small and the meaningfulness of this difference is questioned.

Shifting Targets for Improvement- From 1999 to through 2015, CMS used set “cut-points” or thresholds for the star levels 1 through 5. This helped set targets at the beginning of the year for what rate of effectiveness was needed to achieve a given star level for each measure. Beginning in 2016, however, this was changed. Now health plans do not know the targets for star rating cut-points at the beginning of the measurement year. The star level thresholds can and do go up or down from one year to the next. This is because the Star rating score for each measure is not based on set absolute thresholds or cut-points, but on the value observed in comparison to others. The thresholds for 2, 3, 4, or 5 Star performance change year to year

WHITE PAPER - A Focused Look at the Medicare Health Outcomes Survey

and are based on the total performance across all health plans. There are no set cut-points for each measure's star levels.

Quality management personnel report that this makes setting quality improvement targets more difficult. Improvement in quality of care and service, in patient health outcomes, and in access and satisfaction is always a goal. Resources for quality improvement efforts may be expended throughout the year and be effective in improving a specific measure result. However, the threshold or cut-point for that measure (when released after the end of the year) may have also moved up. The net effect of this method can be that no star level improvement is achieved, even when the plan's effort did result in improvement on a measure. If enough plans also improved in that measure score, the plan will show no change or even decline. This can be interpreted by the public that the plan has not worked for or accomplished any improvement in that measure score. The shifting targets can dampen efforts across provider and plans trying to work together to demonstrate improvement. The reverse can also happen--where no improvement is made by a plan in a particular measurement area, but since the aggregate of all other MAO plan scores decline, the result can look like improvement.

As an illustration of the effects of this methodology, in 2015, a 5-star level of performance (the highest level) for the "Improving or Maintaining Physical Health" measure was at 68% of members self-reporting that they had achieved the same or better physical health than two years ago. A 4-star level was 60% in that same year. Then the 5-star level threshold went up to 72% in 2016, and up to 84% in 2017. The 4-star level went back down to 72% in 2018. Plans that achieved a consistent score of 67% would have been rated 4 stars in 2015 (good performance), but only 3 stars in 2016, 2017, and 2018—which is considered average performance. This measure is weighted a "3" and therefore has a greater effect on a plan's overall Star rating than the measures weighted a "1."

Final Adjustments- After all the measure-specific scores are tabulated, the summary and overall Star rating for the health plan is calculated as a weighted average of the measures. Finally, there are two adjustments made to the results. First, CMS rewards plans with consistently high performance. Second, CMS applies a Categorical Adjustment Factor (CAI) to nine measures. The CAI is meant to assist plans that have a high proportion of enrollees with low income, dual eligibility status, or who are disabled. This adjustment has not affected many contracts over the three years of its use (since 2016). The net effect of this adjustment has been minimal. Industry experts explain this may be due to the few measures included (9 out of 48) in CAI or the limited scope of the adjustment factor.

SNP Specific Measures- Special Needs Plans have four more Star measures to collect/report compared to general MAOs. These are all process measures with a weight of 1 and include:

ID	Measure Name	Brief Description – % SNP enrollees within the measurement year:
C08	SNP Care Management	. . .who receive a Health Risk Assessment
C09	Care for Older Adults – Medication Review	. . .of older adults who had a medication review
C10	Care for Older Adults – Functional Assessment	. . .of older adults who had a functional status assessment
C11	Care for Older Adults – Pain Assessment	. . .of older adults who had a pain assessment

Impact of Star Ratings- Star ratings are carefully watched by plans, advocates, and industry experts. This is because they impact MAOs in reputation, marketing, enrollment, and financially. Star ratings are a presumed marker of performance in care management and care provision, as well as an indicator of internal processes. Star ratings are used by consumers in selecting among health plans when considering their health insurance options. Plans with 4.5 or 5 stars are considered “high performing plans” and have a special star icon on the Medicare Plan Finder. Star ratings affect plans financially. Plans with 4 stars or more can receive quality bonus payments which can be significant. Plans with less than 3 stars as an overall rating triggers the potential loss of their entire contract with CMS given poor performance.

For more information, see the current Part C & D Star Rating Technical Notes including specifications and methodology for all measures which are available at: <http://go.cms.gov/partcanddstarratings>.

APPENDIX B: Medicare Health Outcomes Survey Instrument

Development - The Medicare Health Outcomes Survey (HOS) evolved from other surveys, beginning with the Medical Outcomes Trust and Short Form-36 (SF-36) instruments which were first developed and tested in a Veterans population in the late 1980's/early 1990's by John E. Ware and colleagues at Tufts University.¹¹ These instruments were designed to be brief, generic, self-administered surveys for individuals to report/rate themselves on health-related quality of life, conditions, function, and disease burden. Subsequent modifications of the MOS and SF-36 resulted in the Veterans RAND 36 Item Health Survey (VR-36) and shorter instruments (VR-12 and VR-6). These instruments continue to be used in health services research and practice in addition to being contained within the Medicare HOS.¹²

Applied to MAOs - The Medicare Health Outcomes Survey was first applied to the Medicare Advantage population in 1998—with a baseline (Cohort 1) survey of enrollees that year and a follow-up survey on the same individuals two years later in 2000. The instrument has gone through several revisions since 1998, however it retains the Veteran/RAND-12 items—which are central in calculating several of the Medicare Stars measures. Items on the Medicare HOS include many of the eight domains from the original instrument, such as: general health, physical functioning, mental health, bodily pain, energy/fatigue, social functioning, and role limitations.

Current Survey - The Medicare HOS is now 68 items in length. The two HOS-derived *physical health* and *mental health component summary scores* are calculated from twelve separate items contained within HOS—all from the VR-12. There are another three measures derived from HOS responses and included in the Medicare Star Ratings measure list for Medicare Advantage Organizations (see Table 2).

Measure #	Measure Name	Primary Data Source	Improvement Measure?	Weight in Stars
C04	Improving or Maintaining Physical Health (PCS)	HOS	No	3
C05	Improving or Maintaining Mental Health (MCS)	HOS	No	3
C06	Monitoring Physical Activity	HEDIS/HOS	Yes	1
C18	Reducing the Risk of Falling	HEDIS/HOS	Yes	1
C19	Improving Bladder Control	HEDIS/HOS	No	1

As shown in Table 2, several measures rely on the self-reported HOS survey data as the source information to derive the Star measure. Two of the most heavily weighted measures (weight of 3) derived from HOS are considered patient outcome measures. The two measures are: Physical Component Summary (C04 – PCS) on *improving or maintaining physical health* and Mental Component Summary (C05 - MCS) on *improving or maintaining mental health*. These two measures are used as indicators of how well the health plan is addressing physical and mental health needs of the beneficiary. (More detail on the derivation—what goes into creating the composite measure scores—is provided further in this White Paper).

There are additional condition-specific questions contained in HOS, such as self-reported functional status, disability status, and conditions/diagnoses (that a doctor has told the beneficiary he/she has). There are also questions on other beneficiary characteristic such as:

WHITE PAPER - *A Focused Look at the Medicare Health Outcomes Survey*

race, primary language, income and education level, marital status, living arrangements, and home ownership. Not all questions are answered by the beneficiary. For example, the question on income is most often skipped. When the respondent provides this information, CMS can use this information to make some adjustments to the responses given—more about these adjustments are discussed further elsewhere in this White Paper.

HOS Cohort Sample - Medicare Advantage plans must participate in the HOS self-report survey process as a condition of their contract with the Centers for Medicare and Medicaid Services (CMS). The National Committee on Quality Assurance (NCQA) is under contract with CMS to oversee the administration of the Medicare Health Outcomes Survey as applied to Medicare Advantage Organizations. NCQA is to ensure consistency and oversee the HOS survey methodology.¹³ Each spring a random sample (baseline) of Medicare beneficiaries is drawn from Medicare Advantage Organizations' enrollment lists (MAOs with a minimum of 500 enrollees). Independent survey vendor companies approved by CMS and contracted by the MAOs then receive a list of the beneficiaries CMS. Health plans pay for the survey, but the beneficiary/enrollment mailing/contact list generated from the health plan enrollment file is not made known to the plans. Therefore, the plans are blinded as to who is being surveyed during the measurement year. The survey also has a space for the respondent to indicate if another person (not the beneficiary) completed the form—this may be a family member, friend, or professional caregiver. This person is called a “proxy” respondent.

HOS Survey Process - The HOS survey vendor makes multiple attempts to contact these individuals by telephone or mailed request to conduct the survey (Baseline/Time 1). Two years later, the same people are surveyed again (Follow-up/Time 2). Responses from baseline are compared to responses at follow-up to see if there are differences. The original list of names given to the vendor for each Medicare Advantage plan is 1,200 people. However, response rates to the request to participate in the HOS survey result in much lower actual sample sizes.

NCQA reports that in 2017, the average response to the HOS survey (across all MA plans) was 38% (Baseline), or 456 people. The average follow-up (Time 2) response rate was 65%. This yields a final sample of 296 individuals of the original list of 1,200. Unfortunately, some plans report their final sample sizes are even smaller—at less than 100 individuals completing both HOS surveys. There is evidence that response rates vary substantially across Medicare Advantage Organizations (MAOs) and populations.

The responses of the individuals completing both the baseline and follow-up HOS survey are considered by CMS as being representative of the responses of the entire group of consumers enrolled and served by that health plan. Currently, there is no process to compare the characteristics of the respondents in the final sample to the characteristics of the entire plan enrollment (e.g., to see if the two groups have a similar profile across age, gender, conditions, education level, income, language, functional status, etc.). Given the time lag between survey years and the possibility of changing enrollment profile, this would have to be done twice – in the baseline and follow-up years.

HOS-derived HEDIS measures: *Maintaining or Improving Physical or Mental Health*

The two most heavily weighted Star measures that are considered patient reported outcomes are the *physical health* and *mental health* component summary measures. Each is a composite score measure derived from six separate questions that the beneficiary answers about his/her physical or mental health status. The beneficiary responds based on his/her perspective and

status as of the time the survey is conducted. A decline from the answers the person provided two years prior on these same questions will result in a poorer score.

Metric - The percentage of sampled Medicare enrollees **65 years of age or older** (denominator) whose physical health status or mental health status was the same or better than expected (numerator). [Note that in 2018 CMS proposed that this measure also include **younger people age 18-64 who are dually eligible**. This is being considered and is under testing by Mathematica, under contract with CMS.

What responses are included - All beneficiaries who responded to both baseline and follow-up HOS and are continuously enrolled in a MA plan for 24 months (baseline to follow-up survey) are included in the calculation of these composite scores, as long as there is sufficient data. CMS has a method for computing the scores if some of the data are missing in the second survey. The twelve PCS and MCS questions are:

C04 – Improving or Maintaining Physical health (PCS) – “physical component summary”

Measure Description - This measure is derived from six items on HOS to derive a **physical health** composite score called the “Physical Component Summary.” This is defined as: *The percent of plan members whose physical health was the same or better than expected after two years.* Beneficiaries’ response to the surveyor (by phone) or in the mailed survey on these items are used to derive the PCS:

7. In general, would you say your health is: Excellent, Very Good, Good, Fair, Poor

Does your health **now** limit you in these activities? If so, how much? (Scale: limited a lot, limited a little, not limited at all)

8. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf?

9. Climbing **several** flights of stairs?

During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities **as a result of your physical health**? (Scale: none of the time, a little of the time, some of the time, most of the time, all of the time)

10. **Accomplished less** than you would like.

11. Were limited in the **kind** of work or other activities.

12. During the past 4 weeks, how much did pain interfere with your normal work (including work outside the home and housework)? (Scale: not at all, a little bit, moderately, quite a bit, extremely)

C05 – Improving or Maintaining Mental health (MCS) – “mental health component summary”

Measure Description- This measure is derived from six items on HOS to derive a **mental health** composite score called the “Mental Health Component Summary.” This is defined as: *The percent of plan members whose mental health was the same or better than expected after two years.*

During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)? (Scale: none of the time, a little of the time, some of the time, most of the time, all of the time)

1. **Accomplished less** than you would like
2. Didn't do work or other activities as **carefully** as usual

How much of the time during the past 4 weeks: (Scale: all of the time, most of the time, a good bit of the time, some of the time, a little of the time, none of the time)

3. Have you felt **calm and peaceful**?
4. Did you have a **lot of energy**?
5. Have you felt **downhearted and blue**?
6. During the past 4 weeks, how much of the time has your **physical health or emotional problems** interfered with your social activities (like visiting friends, relatives, etc.)? (Scale: not at all, a little bit, moderately, quite a bit, extremely)

Responses to these twelve questions are used to derive the two measures on maintaining or improving physical or mental health. The results are then compared to expected or predicted responses. The steps in calculating the expected or predicted follow up responses and scoring the plan's performance are described further below.

6 Steps¹⁴ to Calculate the Predicted PCS and MCS:

Step 1: There are 59 total response options across the 12 questions used in HOS on physical and mental health. From these 59 options, **47 “indicator variables” are created.**

Step 2: From the full dataset, **weights are created for each of the 47 indicator variables.** The weights are derived from the Veterans Short Form Health Survey (SF-36) PCS and MCS Scales using the *1999 Large Health Survey of Veteran Enrollees*.

Step 3: Using the normative values that were obtained from a *1990 survey of the U.S. general population*, NCQA/CMS determines where 50% falls. A mean score of 50 represents the national average of all scores received across all plans. A 10-point difference above or below the mean score is considered one standard deviation.

Step 4: NCQA/CMS computes expected data for any respondent who did not complete all of the VR-12 items. This is done by using a modified regression estimate (MRE). With the MRE algorithm, scores can be calculated in as many as 90% of the cases in which one or more VR-12 responses are missing. Depending on the pattern of missing item responses for a beneficiary, a different set of regression weights is used to compute that individual's scores.

Step 5: There are two adjustments made in Step 5 for: (1) mode of administration and (2) case-mix factors.

Mode - Beneficiary results are adjusted for the impact of telephone administration compared to mail administration. That is, telephone results are generally higher, so if a respondent answered by phone, it is assumed that these scores are higher (better) than if he/she answered the mailed survey. To do this, the average difference between telephone and mail survey is subtracted from the phone survey results.

Case-mix – The responses from the beneficiaries sampled who provided answers in both the baseline and follow-up surveys are also “case-mix” adjusted – that is results are adjusted based on factors known to impact self-reported health status to try to account for the effects of these factors. This is also done to try to control for pre-existing differences across MA plans. The adjustments utilize the baseline data provided by the individual on the questions in HOS around characteristics such as: age, gender, race/ethnicity, income, education, home ownership, and their self-reported chronic medical conditions, self-reported limitations in activities of daily living, and their baseline physical and mental health scores. However, these adjustments can only be made *if the beneficiary responds to the specific questions which provide the data/information for such case mix adjustment in the baseline HOS survey.*

Creating “best fit” models - For the PCS and MCS, CMS classifies every beneficiary into one of the two predicted outcome categories: (1) maintained or improved, OR (2) declined. For the PCS, in order to calculate expected outcomes, CMS uses 3 different “physical health” case mix models and 6 different “death case mix” models. These models were created to compute “best fit” predictions, using the data provided at baseline. There are no predictions made for missing data. If there is incomplete data on two or more of the key variables (income, education, home ownership, self-reported medical conditions, self-reported limitations in activities of daily living) then only a few factors (e.g., age, gender, race and region of the country and survey vendor) are used. CMS uses these “best fit” models to categorize the beneficiary as either (1) expected to improve/maintain OR (2) expected to decline. The results for that beneficiary are then used to compare to the predicted expected results.

For the MCS, CMS uses a series of three different mental health models to categorize the respondent into expected maintain/improve or decline. Again, this is because CMS may not have data for all independent variables that could be used to calculate an expected score. Therefore, like the PCS expected score, the MCS predicted or expected score for each beneficiary is derived from a best fit model.

Step 6: The result for each respondent indicates “better,” “same,” or “worse than expected” values for the PCS and MCS. All respondent results for each health plan are aggregated for the PCS and MCS measures. Deceased members are included in the calculation of HOS results (generally, as “worse than expected”).

WHITE PAPER - *A Focused Look at the Medicare Health Outcomes Survey*

This final step produces the dataset for each health plan which is used to compare to all other health plans' results. The responses for all health plans create a national data set. CMS then sets the cut-points or thresholds for what specific values indicate "better/same," or "worse than expected" health status results. In this way, CMS derives the MAO (plan-specific), state, and national HOS results for the measurement period for these two measures.

What do the scores mean? - Interpretation of an individual's PCS or MCS score is described by CMS as follows: *Very high scores indicate absence of health concerns, and no limitations in usual social and role activities due to physical or mental health problems. A positive change (i.e. follow-up score higher than baseline) or no change in score from baseline to follow-up indicates an individual has improved or maintained their current physical or mental health.*

Interpretation of un-adjusted aggregate plan results: *A higher rate is indicative of a larger number of beneficiaries whose physical or mental health was the same or better from baseline to follow-up on the PCS component of the HOS.*

Interpretation of adjusted observed-to-expected rate: *A higher rate indicates the plan had a higher proportion of beneficiaries whose physical health was same or better from baseline to follow-up than would be expected based on the adjustments made. A rate of 1.0 indicates the plan is performing exactly as expected based on the case-mix. A rate below 1.0 indicates the plan is performing worse than expected. A rate above 1.0 indicates the plan is performing better than expected.*

Endnotes

¹ For more information on the Medicare MAO Quality and Performance Rating Management System, see: <https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.html>

² CMS *Notice of Proposed Rulemaking*, November 28, 2017. See Federal Register, Vol 82, No. 227.

³ The list of 2018 Medicare MAO Star Measures found at: <file:///C:/Users/debpa/Documents/SNPA/Quality/CMS/2018MeasureList.pdf> . See also the 2019 CMS Technical Notes at cms.gov.

⁴ For more information about Dr. Ware and his work, see: <https://www.jwrginc.com/about> Additional resources and studies: <https://news.brown.edu/articles/2018/07/rankings>

⁵ ASPE (2016). Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Payment Programs”

⁶ See HealthMeasures Person-Centered Assessment Resource for more information and to download the Global Health Instrument at http://www.healthmeasures.net/administrator/components/com_instruments/uploads/Global%20Health%20Scale%20v1.2%2013Apr2018.pdf

⁷ See: <http://www.fvfiles.com/521475.pdf>

⁸ For more information on the Medicare MAO Quality and Performance Rating Management System, see: <https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/PerformanceData.html>

⁹ CMS *Notice of Proposed Rulemaking*, November 28, 2017. See Federal Register, Vol 82, No. 227.

¹⁰ The list of 2018 Medicare MAO Star Measures found at: <file:///C:/Users/debpa/Documents/SNPA/Quality/CMS/2018MeasureList.pdf> . See also the 2019 CMS Technical Notes at cms.gov.

¹¹ For more information about Dr. Ware and his work, see: <https://www.jwrginc.com/about>

¹² For more information about the VR-36, VR-12 and VR-6D – See Boston University School of Public Health at: <http://www.bu.edu/sph/about/departments/health-law-policy-and-management/research/vr-36-vr-12-and-vr-6d/about-the-vr-36-vr-12-and-vr-6d/>

¹³ For more detail on the Medicare Health Outcomes Survey, see: <http://www.hosonline.org/>. For more detail on NCQA and HEDIS, see: <http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2017-table-of-contents/sohc-hedis-measures-of-care>

¹⁴ From the *Measurement Information Forms* for PCS and MCS published by NCQA.

⁸ The Assistant Secretary for Planning and Evaluation (ASPE) published in late December 2016. The entire report can be found at: <https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicare-value-based-purchasing-programs>

References

- Assistant Secretary for Planning and Evaluation. Report to Congress. (2016). *Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs A Report Required by the Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014*. U.S. Department of Health and Human Services. December.
- Department of Veterans Affairs, Veterans Health Administration (2003). *Survey of Veteran Enrollees' Health and Reliance Upon VA 2002 & 1999, Executive Summary*. December.
- Institute of Medicine. *Survey Measurement of Work Disability: Summary of a Workshop*. (2000). Nancy Mathiowetz and Gooloo Wunderlich, editors. Washington DC: National Academy Press. Available through NIH at https://www.ncbi.nlm.nih.gov/books/NBK225420/pdf/Bookshelf_NBK225420.pdf
- Iqbal, Sheikh Usman, Rogers, W., Selim, A., Qian, S., Lee, A., Ren, X., Rothendler, J., Miller, D., and Kazis, L. (2007). *The Veterans RAND 12 Item Health Survey (VR-12): What it is and How it is Used*. Paper by the Center for Health Quality, Outcomes, and Economic Research, Veterans Administration Medical Center (Bedford, MA), and Center for the Assessment of Pharmaceutical Practices, Boston University School of Public Health.
- Kazis, L., Lee, A., Spiro, A., et al. (2004). Measurement Comparisons of the Medical Outcomes Study and the Veterans SF-36 Health Survey. *Health Care Financing Review*. Vol 25(4). Summer.
- Krahn, G., Walker, D., and Correa-De-Araujo, R. (2015). Persons with Disabilities as an Unrecognized Health Disparity Population. *American Journal of Public Health*. Supplement 2. Vol 105(S2).
- Livermore, G, Whalen, D., Prenovitz, S., Aggarwal, R., Bardos, M. (2011). *Disability Data in National Surveys*. A report by Mathematica Policy Research, Inc. submitted to the Assistant Secretary of Planning and Evaluation, U.S. Department of Human Services. August.
- Nelson, E., Landgraf, J., Hays, R., Wasson, J. and Kirk, J. (1990). The Functional Status of Patients: Can it be measured in physicians' offices? *Medical Care*. 28(12): 111-1126. December.
- Pedersen, PV., Gronbaek, M. and Curtis, T. (2012). Associations between deprived life circumstances, well-being, and self-rated health in a socially marginalized population. *European Journal of Public Health*. October: 22(5): 647-52.
- Rowland, M., Peterson-Besse, J., Dobberlin, K., Walsh, ES., Horner-Johnson, W. (2014). Health outcome disparities among subgroups of people with disabilities: A scoping review. *Disability and Health Journal*. April 7(2): 136-150. DOI: 10.1016/j.dhjo.2013.09.003.
- Schalet, B., Rothrock, N., Hays, R., Kazis, L., Cook, K., Rustsohn, J., and Cella, D., (2015). Linking Physical and Mental Health Summary Scores from the Veterans RAND 12-Item Health Survey (VR-12) to the PROMIS® Global Health Scale. *Journal of General Internal Medicine*. 30(10): 1524-1530. October.

Sunghee Lee, Norbert Schwarz. (2014). Question Context and Priming Meaning of Health: Effect on Differences in Self-Rated Health Between Hispanics and Non-Hispanic Whites.” *American Journal of Public Health* 104, no. 1 (January 1): pp. 179-185. DOI: 10.2105/AJPH.2012.301055 PMID: [23678900](https://pubmed.ncbi.nlm.nih.gov/23678900/)

Sunghee Lee and David Grant (2009). The Effect of Question Order on Self-rated General Health Status in a Multilingual Survey Context. *American Journal of Epidemiology*. Vol. 169, No. 12. DOI: 10.1093/aje/kwp070.

Sunghee Lee, Nancy A. Mathiowetz and Roger Tourangeau. (2004). Perceptions of Disability: The Effect of Self- and Proxy Response. *Journal of Official Statistics*. Vol.20, No.4. pp. 671–686