


ORIGINAL RESEARCH

Incentivizing Excellent Care to At-Risk Groups with a Health Equity Summary Score



Denis Agniel, PhD¹, Steven C. Martino, PhD^{1,2}, Q Burkhardt, MS¹,
 Katrin Hambarsoomian, MS¹, Nate Orr, MA¹, Megan K. Beckett, PhD¹,
 Cara James, PhD³, Sarah Hudson Scholle, MPH, DrPH⁴,
 Shondelle Wilson-Frederick, PhD³, Judy Ng, PhD⁴, and Marc N. Elliott, PhD¹ 

¹RAND Corporation, Santa Monica, CA, USA; ²RAND Corporation, Pittsburgh, CA, USA; ³Centers for Medicare & Medicaid Services, Baltimore, MD, USA; ⁴National Committee for Quality Assurance, Washington, DC, USA.

BACKGROUND: Social risk factors (SRFs) such as minority race-and-ethnicity or low income are associated with quality-of-care, health, and healthcare outcomes. Organizations might prioritize improving care for easier-to-treat groups over those with SRFs, but measuring, reporting, and further incentivizing quality-of-care for SRF groups may improve their care.

OBJECTIVE: To develop, as a proof-of-concept, a Health Equity Summary Score (HESS): a succinct, easy-to-understand score that could be used to promote high-quality care to those with SRFs in Medicare Advantage (MA) health plans, which provide care for almost twenty million older and disabled Americans and collect extensive quality measure and SRF data.

DESIGN: We estimated, standardized, and combined performance scores for two sets of quality measures for enrollees in 2013–2016 MA health plans, considering both current levels of care, within-plan improvement, and nationally benchmarked improvement for those with SRFs (specifically, racial-and-ethnic minority status and dual-eligibility for Medicare and Medicaid).

PARTICIPANTS: All MA plans with publicly reported quality scores and 500 or more 2016 enrollees.

MAIN MEASURES: Publicly reported clinical quality and patient experience measures.

KEY RESULTS: Almost 90% of plans measured for MA Star Ratings received a HESS; plans serving few patients with SRFs were excluded. The summary score was moderately positively correlated with publicly reported overall Star Ratings ($r = 0.66$ – 0.67). High-scoring plans typically had sizable enrollment of both racial-and-ethnic minorities (38–42%) and dually eligible beneficiaries (29–38%).

CONCLUSIONS: We demonstrated the feasibility of developing and estimating a HESS that is intended to promote and incentivize excellent care for racial-and-ethnic minorities and dually eligible MA enrollees. The HESS measures SRF-specific performance and does not simply duplicate overall plan Star Ratings. It also identifies plans that provide excellent care to large numbers of those with SRFs. Our methodology could be extended to other SRFs, quality measures, and settings.

KEY WORDS: disparities; social risk factors; quality-of-care; public reporting; Medicare Advantage health plans.

J Gen Intern Med 36(7):1847–57

DOI: 10.1007/s11606-019-05473-x

© Society of General Internal Medicine 2019

INTRODUCTION

Patients with social risk factors (SRFs)—sometimes called *social determinants of health* [1]—(e.g., patients with low income and racial-and-ethnic minorities) have less access to material and social resources and lower status than more-advantaged patients, leading to worse healthcare outcomes, including hospital readmission [2–5] and in-hospital mortality [6–8], independent of quality-of-care received [9, 10].

We propose a strategy of identifying and incentivizing delivery of high-quality care to patients with SRFs [11–13]. Because providing high-quality care to at-risk patients may be more expensive than increasing overall quality-of-care [14–16], organizations might adopt a one-size-fits-all-quality-improvement approach focused on the average patient rather than the linguistic, cultural, or educational needs of patient subgroups [16]. Further, organizations may avoid disadvantaged patients who might worsen quality scores [17, 18]. Thus, mechanisms are needed to monitor and mitigate such unintended consequences of public reporting and pay-for-performance schemes based solely on quality-of-care overall [17, 19]. A publicly reported measure highlighting quality-of-care for patients with SRFs could inform patients, quality-improvement staff, payers, and other stakeholders and be linked to incentives to improve care for those with SRFs.

Here, we present proof-of-concept data on the development of a new summative score to characterize the quality-of-care delivered to Medicare patients with SRFs in Medicare Advantage (MA) contracts (hereafter “plans”). In developing this Health Equity Summary Score (HESS), we aimed to provide a succinct and easy-to-understand overall score as a potential basis for promoting high-quality care to patients with SRFs.

The HESS was constructed as a composite of quality performance for two SRF groups across many quality measures. We designed the HESS to (a) measure both current (cross-

Prior Presentations This work has not been previously presented.

Received June 3, 2019

Revised September 12, 2019

Accepted October 9, 2019

Published online November 11, 2019

sectional) quality-of-care and quality improvement and (b) incentivize good care to both racial-and-ethnic minorities and those dually eligible for Medicare and Medicaid. While different processes are involved in addressing the needs of these SRF groups (and performance can be disaggregated by SRF group), there is significant overlap in these populations. More importantly, combining race-and-ethnicity and dually eligible scores signals that improved care for both groups is important without presupposing that similar mechanisms would be required to improve care for each.

Incorporating cross-sectional and improvement HESS components recognizes good care currently provided to patients with SRFs while also potentially incentivizing low-performing but improving plans. We measured improvement both as the narrowing or widening of within-plan differences in care and as improvement in quality-of-care for those with SRFs relative to national benchmarks. The within-plan improvement component was designed to help plans identify underserved beneficiary groups and incentivize improved care for these lagging groups relative to the highest performing group at baseline. Since many healthcare disparities can be attributed to clustering of at-risk beneficiaries [20–23], the nationally benchmarked improvement component of the HESS was designed to ensure that there is an incentive for absolute improvement in care for beneficiaries with SRFs. This combined approach to measuring improvement targets health equity broadly, even if each individual component does not necessarily measure equity according to some definitions. In any case, our approach allows the disaggregation of these components so plans may understand their improvement according to each of these metrics.

Here, we demonstrate a proof-of-concept application for clinical quality and patient experience, focused on care provided to Medicare beneficiaries by MA plans, which provide care for almost twenty million older and disabled Americans and collect extensive quality measure and SRF data. We relied on two sources of person-level performance data: (a) clinical quality measures from the Healthcare Effectiveness Data and Information Set (HEDIS®) and (b) patient experience measures from the MA Consumer Assessment of Healthcare Providers and Systems (CAHPS®) survey. These measures are among those used to compute MA Contract Quality Bonus Payments, which pay plans more for higher quality care [24] and which are used by the Centers for Medicare & Medicaid Services (CMS) for stratified reporting by race-and-ethnicity within MA plans [25].

METHODS

Data

We analyzed five HEDIS measures that are used both for stratified reporting by race-and-ethnicity and for MA Star Ratings [25]: breast cancer screening, colorectal cancer screening, diabetes care (both nephropathy and retinal exam),

and adult BMI assessment. We also analyzed seven CAHPS composite measures currently used for stratified reporting by race-and-ethnicity: doctor communication, ease of getting needed care, getting care quickly, ease of getting needed prescription drugs, customer service, care coordination, and flu immunization. Measures used by CMS for race-and-ethnicity-stratified reporting meet established criteria for statistical reliability and informativeness [26]. CAHPS scores are adjusted for dual eligibility through case-mix adjustment; Star Ratings for HEDIS measures are adjusted for dual eligibility through the Categorical Adjustment Index [27]. Neither CAHPS nor HEDIS is adjusted for race-and-ethnicity [15, 17]. See Appendix 1 for information about how HEDIS and CAHPS data are collected and detailed descriptions of all measures.

We developed the HESS based on two SRFs that have been prioritized by the National Academy of Medicine: race-and-ethnicity and dual eligibility for Medicare and Medicaid [14]. Race-and-ethnicity was self-reported on the CAHPS survey. We used that information to classify patients into Hispanic and (non-Hispanic) black, Asian and Pacific Islander (API), and white groups, following a variant of the Office of Management and Budget approach (see Appendix 1). American Indian/Alaska Natives and multiracial categories were not common enough to be reliably measured and were excluded from this analysis [26]. Race-and-ethnicity in HEDIS was estimated using CMS' Medicare Bayesian Improved Surname Geocoding (MBISG) 2.0 values [26, 28], yielding a probability of belonging to each of the aforementioned racial-and-ethnic groups based on CMS administrative data, as is done in stratified reporting of these measures. These probabilities were used in all analytic models in the same way that racial-and-ethnic group indicators are typically used [29]. Medicare-Medicaid dual eligibility is determined by income (up to 150% federal poverty level) and assets [30]. It is a proxy for income and wealth and is associated with marked disparities in health and healthcare among Medicare beneficiaries [15]. Dual eligibility status was obtained via Medicare administrative data (see Appendix 1 for details).

Analytic Approach

The HESS was constructed as a composite of cross-sectional and improvement scores constructed from patient experience (CAHPS) and clinical quality (HEDIS) information (each linkable to SRF data at the patient level). Cross-sectional scores attempted to identify plans providing excellent care to those with SRFs in current data. Improvement scores identified plans that particularly improved the care of their members with SRFs.

Our approach to constructing the HESS is illustrated in Figure 1. Cross-sectional and improvement scores were computed both for dual eligibility and race-and-ethnicity as described below. After blending cross-sectional and improvement scores for each SRF according to Figure 2, overall race-

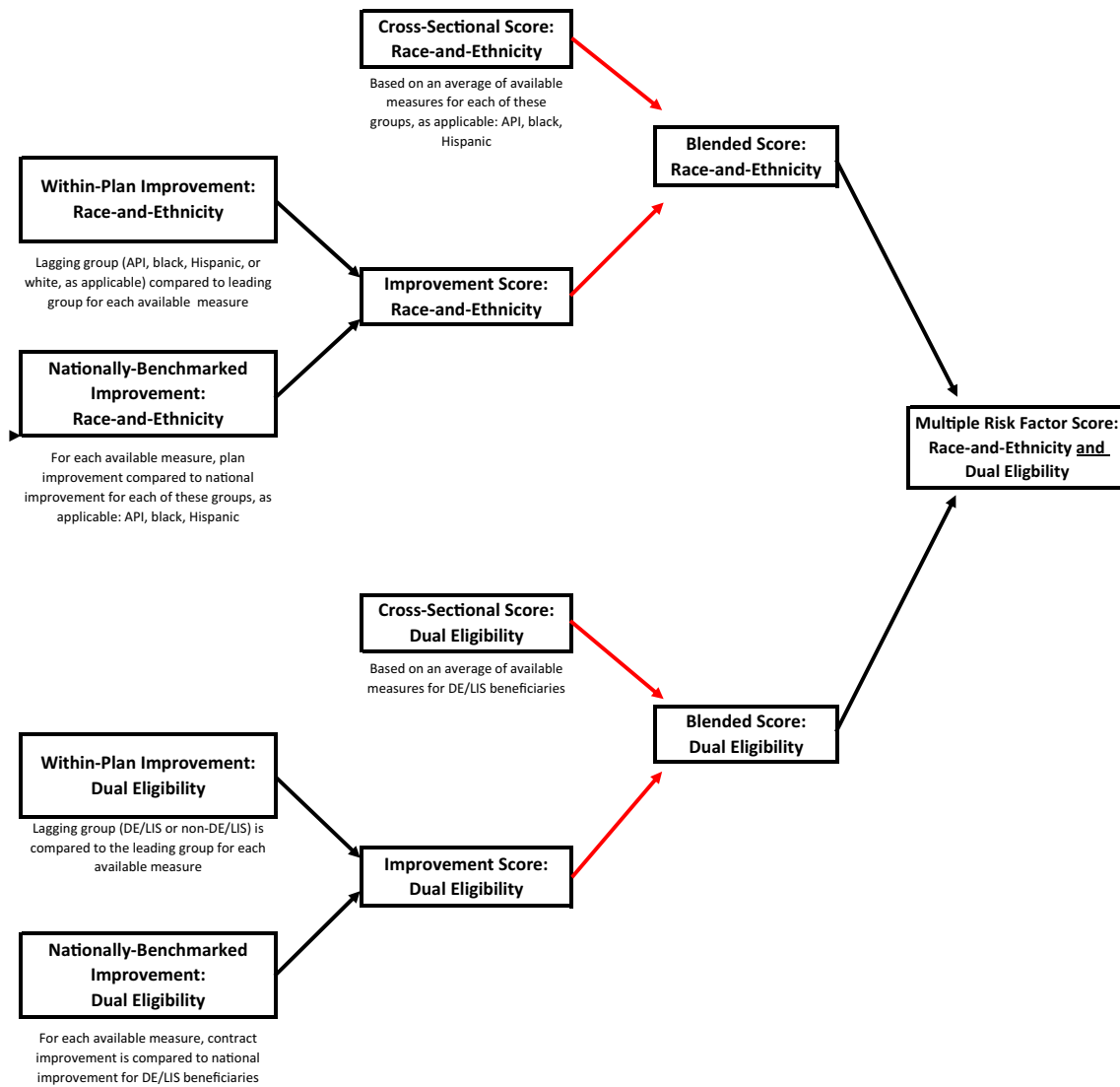


Figure 1 Overview of HESS components and construction. Black arrows refer to averaging non-missing groups. Red arrows refer to using the lookup table in Figure 2.

and-ethnicity and dual-eligibility scores were combined into a single HESS. If a score was only available for either race-and-ethnicity or dual-eligibility, this single score was used as the HESS. Because race-and-ethnicity scores and dual-eligibility scores were standardized to the same scale and averaged, there is no inherent advantage to a plan’s score if its HESS was based on one SRF rather than two. Full computational details appear in [Appendix 2](#).

The HESS was computed separately for the CAHPS and HEDIS domains. As is standard in CMS’ stratified public reporting of quality measures [25], we combined 2 years of data to increase sample size and ensure efficient estimation of cross-sectional performance. Performance period (2015–2016) and baseline (2013–2014) data were used to gauge improvement. We began with the 398 plans that were measurable for overall quality (see [Appendix 2](#)) in these domains in 2016.

We required HESS inputs to have sufficient sample size ($n = 100$) and reliability (> 0.7) for accurate measurement [31]. Each plan’s HESS was based only on those measures and SRF

groups for which it met these measurability requirements. If a plan met these requirements for any measure in any SRF group, it received a HESS score. Plans that did not meet this requirement for any measure in any SRF were considered unmeasurable.

Cross-sectional performance for each measurable racial-and-ethnic group was estimated using linear models, yielding one score for each measurable racial-and-ethnic group for each measure (see Fig. 1). All measures were rescaled to a 0–100 scale and modeled separately. Models for CAHPS data were case-mix adjusted and survey-weighted(see [Appendix 1](#)). Estimates were standardized to put them on a common scale across measures and groups and then combined to yield a single cross-sectional performance score for each plan. Performance scores were then converted to a five-star scale using the MA Part C clustering algorithm [32]. The resulting score represents how well the plan provides care to all racial-and-ethnic minorities. Similar estimation, standardization, and

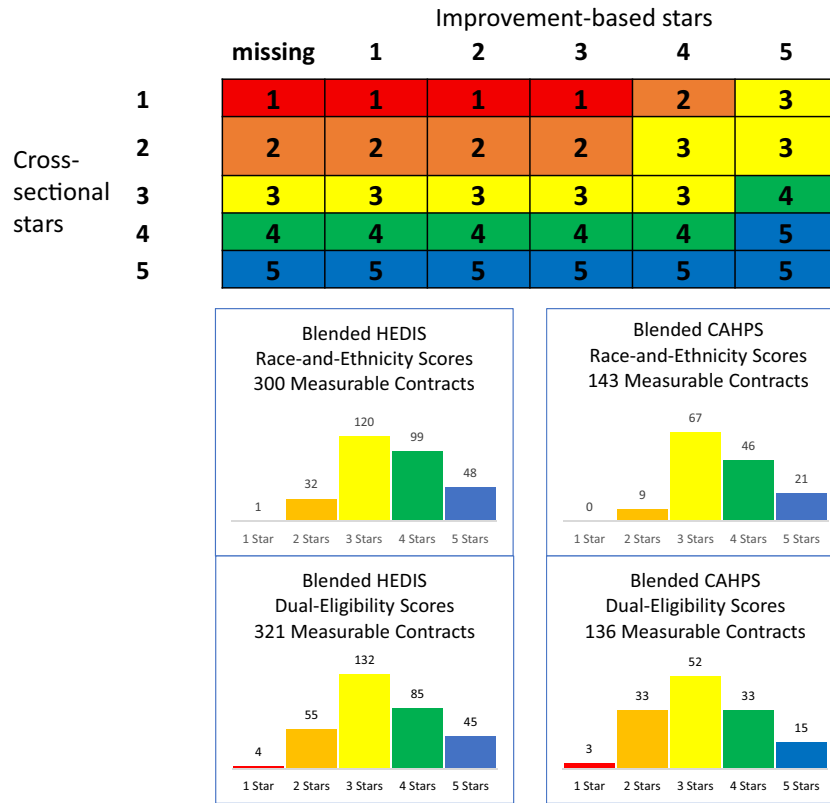


Figure 2 Blended race-and-ethnicity and dual-eligibility scores: relationship to cross-sectional and improvement scores and observed distributions. Cell entries display blended stars, which correspond to the color in the bar charts below. Blended scores are combinations of cross-sectional and improvement scores and are computed separately for each SRF and data source. Cross-sectional scores reflect current care to SRF groups, and improvement scores account for both between-plan and within-plan improvement in care to SRF groups.

combination across measures were done for assessing care provided to the dually eligible.

Improvement scores were constructed in a broadly similar fashion but required 4 years of data (2013–2016) and combined two types of improvement. Measurability requirements ($n = 100$, reliability > 0.7) were enforced in both baseline (2013–2014) and performance (2015–2016) periods. We designed the HESS to capture both within-plan and between-plan improvements in the reduction of disparities. If a plan were to focus only on improving its lagging groups to the level of its leading group, this would fail to address the between-plan component of disparities, which is often larger [20, 33, 34]. The between-plan component measured the improvement of each SRF group compared to that group’s national average improvement. By using a national benchmark, we identified plans that provided excellent care to those with SRFs. The within-plan component measured narrowing or widening of within-plan disparities and compared all other groups to the leading group (with highest baseline score) of each plan. These estimates of improvements were standardized and combined to yield a single score for each plan and SRF, measured on a five-star scale.

Cross-sectional and improvement scores were then blended as shown in Figure 2. The blending scheme scored plans highly if they were already providing excellent care to SRF groups (those with high cross-sectional scores) or they were not yet providing excellent care but were demonstrating

improvement (those with high improvement scores). Low improvement scores never resulted in a blended score lower than a plan’s cross-sectional score, which prevented high-performing plans from being penalized for what may be necessarily limited improvement. Since improvement is arguably more important for low-performing plans, high improvement scores could increase a plan’s blended score, with improvement having more influence when cross-sectional performance was lower [35, 36]. Following precedent [37], we prioritized cross-sectional performance and limited the influence of improvement (to a maximum of two additional stars).

Finally, we computed a plan’s HESS by averaging its blended race-and-ethnicity score and blended dual-eligibility score. If the dual-eligibility score was missing, then the HESS was equal to the race-and-ethnicity blended score, and vice versa. For demonstration, we categorized the HESS into high (4–5 stars), medium (2.5–3.5 stars), and low (1–2 stars) groups.

Evaluating the HESS

We examined differences between measurable plans (those that received scores) and unmeasurable plans in plan enrollment, percentage of enrollees in each racial-and-ethnic group, percentage of dually eligible enrollees, and official HEDIS/CAHPS Star Ratings. All characteristics were averaged over 2015–2016. Star Ratings from 2015 to 2016 were averaged across

HEDIS or CAHPS measures individually and then averaged across years.

We similarly compared the characteristics of high- and low-scoring plans to assess whether the HESS only assigned high scores to plans that already received high overall summary scores and whether high-scoring plans served significant percentages of those with SRFs. Finally, we compared HESS scores between the HEDIS and CAHPS domains.

RESULTS

Plan Measurability

Most plans (86% for HEDIS, 49% for CAHPS) were measurable (had sufficient SRF group sample size and reliability) for at least one of the two SRFs and thus received a HESS score. Plans were more likely to be measurable for dual-eligibility than for race-and-ethnicity. HEDIS clinical quality was measurable for 300 plans for race-and-ethnicity (163 with improvement scores) and 321 for dual-eligibility (152 with improvement scores) of the 398 possible plans. CAHPS performance was measurable for 143 plans for race-and-ethnicity (121 with improvement scores) and for 136 plans for dual-eligibility (113 with improvement scores) of the 388 possible plans. In total, 350 (88%) plans received a HESS for either HEDIS (343, 86%) or CAHPS (184, 49%) (Table 1), and 44% were measurable for both.

SRF-Specific Blended Scores

The distributions of the blended stars of SRF-specific performance appear in Figure 2 and demonstrated an expected bell shape, much like overall plan stars based on HEDIS and CAHPS data [32, 38]. More than one-third of plans (35–49%) fell in the high-scoring category (4–5 stars). Of HEDIS-measurable plans, 147 of 300 were high-scoring on the racial-and-ethnic HEDIS blended score, and 130 of 321 on the dual-eligibility blended score. Of CAHPS-measurable plans, 67 of

143 had high (4–5 stars) blended CAHPS race-and-ethnicity scores; 48 of 136 were high-scoring for dual-eligibility.

Characteristics of Measurable Plans

In the HEDIS data, unmeasurable plans had much lower populations of dually eligible (8.3% vs. 28.5% for measurable plans) and black enrollees (4.2% vs. 13.4%; Table 2). Their publicly reported overall stars tended to be a bit lower in unmeasurable plans (3.2 vs. 3.4 stars on average). Unmeasurable plans also tended to be much smaller than measurable ones, with average enrollments of 7917 and 48,230, respectively. Plans not measurable for CAHPS (Table 2) tended to have fewer dually eligible enrollees (7.6% vs. 45.4% in measurable plans), fewer racial-and-ethnic minorities (81.9% white vs. 47.6% in measurable plans), and higher publicly reported overall stars (3.7 vs. 3.0 in measurable plans).

Characteristics of High-Scoring Plans

The HESS was moderately correlated with overall Star Ratings ($r = 0.67$ for HEDIS, 0.66 for CAHPS), suggesting that the HESS identified SRF-specific quality-of-care, and did not just replicate overall performance (if it did, r would be closer to 1). High HEDIS scorers averaged 3.9 publicly reported overall HEDIS stars, while medium and low performers averaged 3.3 and 2.8, respectively (Table 2). Similarly, high CAHPS scorers averaged 3.7 publicly reported overall CAHPS stars, while medium performers averaged 2.6, and low performers 2.1 (Table 2).

High-scoring plans typically enrolled significant percentages of those with SRFs. Plans with high HESS scores for clinical quality averaged 16% black, 16% Hispanic, and 29% dually eligible beneficiaries. Plans with high HESS scores for patient experience had similar percentages of racial-and-ethnic minorities (16% black and 14% Hispanic) and even more dually eligible enrollees (38%).

Table 1 Distribution of Medicare Advantage Health Plans by HEDIS and CAHPS Health Equity Summary Score

		HEDIS					Overall total
		Not measurable	Total measurable	Low	Medium	High	
CAHPS	Not measurable	12%	39%	6%	24%	12%	51%
		(48)	(166)	(24)	(97)	(45)	(214)
	Total measurable	2%	44%	2%	21%	21%	46%
		(7)	(177)	(9)	(85)	(83)	(184)
	Low	< 1%	3% (12)	1%	2%	1%	3%
		(1)		(2)	(7)	(3)	(13)
	Medium	1%	24% (97)	1%	13%	10%	25%
High	(4)		(4)	(52)	(41)	(101)	
	1%	17% (68)	1%	7%	10%	18%	
Overall total	(2)		(3)	(26)	(39)	(70)	
	14%	86%	8%	46%	32%	100%	
	(55)	(343)	(33)	(182)	(128)	(398)	

Percent of all $n = 398$ plans is given in each cell (number of plans given in parenthesis)

*Low-scoring plans have 1–2 Health Equity Summary Score (HESS) stars, medium-scoring plans 2.5–3.5 HESS stars, and high-scoring plans have 4–5 HESS stars. HESS stars computed in either the Healthcare Effectiveness Data and Information Set (HEDIS) or the MA Consumer Assessment of Healthcare Providers and Systems (CAHPS). Overall total is the sum of not measurable and total measurable. Total measurable is the sum of low, medium, and high

Table 2 Comparison of Medicare Advantage Health Plans for HEDIS and CAHPS Measures by Measurability and by HESS

		<i>n</i>	Mean enrolment	Percent Hispanic	Percent Black	Percent Asian/Pacific Islander	Percent dually eligible	Mean overall HEDIS or CAHPS stars
HEDIS	Unmeasurable	55	7917	15.0%	4.2%	2.1%	8.3%	3.2
HESS	Measurable	343	48,230 [†]	13.0% [†]	13.4% [‡]	4.4%	28.5% [‡]	3.4*
	Low (1–2 stars)	33	18,773	6.2%	7.4% [§]	2.6%	14.5%	2.8
	Medium (2.5–3.5 stars, ref)	182	49,837	12.5%	12.8%	4.7%	40.0%	3.3
	High (4–5 stars)	128	53,541	15.5%	15.8%	4.4%	28.6%	3.9
CAHPS	Unmeasurable	204	48,990	3.7%	6.2%	2.0%	7.6%	3.7
HESS	Measurable	184	37,134	21.2% [‡]	16.3% [‡]	6.2% [‡]	45.4% [‡]	3.0 [‡]
	Low (1–2 stars)	13	10,170	21.9%	13.8%	8.4%	61.4%	2.1 [§]
	Medium (2.5–3.5 stars, ref)	101	30,118	24.9%	19.2%	6.3%	48.4%	2.6
	High (4–5 stars)	70	52,264	15.9% [§]	12.7%	5.5%	38.2%	3.7

* $p < 0.05$, [†] $p < 0.01$, [‡] $p < 0.001$ for difference from unmeasurable plans, where measurable plans have at least 100 completed surveys and reliability of at least 0.7 for any SRF group on the MA Consumer Assessment of Healthcare Providers and Systems (CAHPS) or the Healthcare Effectiveness Data and Information Set (HEDIS), respectively. Limited to plans with an overall MA CAHPS or HEDIS Star Rating

[§] $p < 0.05$, ^{||} $p < 0.01$, ^{|||} $p < 0.001$ for difference of high- and low-scoring measurable HESS plans from medium measurable HESS plans

Larger plans were generally more likely to receive high HESS ratings, consistent with overall MA Star Rating patterns [38]. Mean enrollment size was positively associated with HESS score, with high CAHPS performers averaging more than 52,000 enrollees and low performers about 10,000, with similar HEDIS results.

Comparison of HEDIS and CAHPS HESS Scores

There was a small positive correlation between the HEDIS HESS and the CAHPS HESS ($r = 0.23$). Of those 177 plans measurable on both areas of performance, 93 (52%) were rated similarly in both areas (Table 1). About 10% were highly rated for both HEDIS and CAHPS, while an additional 67 were highly rated in one area and medium in the other.

DISCUSSION

We have demonstrated the feasibility of developing a HESS, which aims to promote and incentivize excellent care for SRF groups, with a proof-of-concept application to MA plans. The HESS was constructed using straightforward methods on publicly reported data. It employed a five-star rating system, based on the method CMS uses to convey plan performance on the Medicare Plan Finder and determine MA quality-based bonus payments [27], that here aggregated information on care for racial-and-ethnic minorities and dually eligible enrollees.

The correlation between the HESS and overall plan quality ($r = 0.66$ – 0.67) was not so high that the HESS would be entirely redundant, and not so low as to suggest that it measured something entirely divorced from overall plan quality. There was some correlation between the clinical quality HESS and the patient experience HESS, similar to what has been found for HEDIS and CAHPS more generally [39]. Further, we found that the HESS does not simply reward plans that provide excellent care to those with SRFs but serve very few of them. High-scoring plans had sizeable enrollment of both

racial-and-ethnic minorities (38–42%) and dually eligible beneficiaries (29–38%).

Based on this proof-of-concept application of our methodology, the HESS could be extended to include other SRFs (e.g., disability, educational attainment, rurality), settings (e.g., hospitals), or measures (e.g., plan disenrollment) where data is available.

HESS performance stars could be developed and separately reported for HEDIS, CAHPS, and other measures, with incentives attached to each, according to policy goals. Publicly reported HESS stars could provide visible, comprehensible summaries of how MA plans provide care to those with SRFs to inform patients, quality-improvement staff, payers, and other stakeholders. For example, high-performing safety-net providers might be highlighted by such reporting. Reporting or incentivizing based on the HESS might also incentivize high-performing plans that serve few at-risk beneficiaries to enroll more such beneficiaries. Because unmeasurable plans tend to be smaller, higher performing, and serve fewer beneficiaries with SRFs, a high-performing unmeasurable plan might become measurable and be recognized by enrolling more beneficiaries with SRFs.

The approach we describe has potential limitations. First, one cannot accurately measure SRF-specific performance for plans with small sample sizes of beneficiaries with SRFs or, for measures collected at near-constant sample sizes across plans (e.g., CAHPS measures and some HEDIS measures), low proportions of such beneficiaries. Although our approach has limited ability to represent small plans, this is arguably a limitation inherent to all quality measurement [31]. Requiring increased sample sizes or oversampling those with SRFs could improve measurability of plans with low proportions of those with SRFs for CAHPS. Second, only a subset of measures or SRF groups is measurable for some plans due to low sample size or reliability. However, our approach gives all available measures and groups equal influence. Therefore, even when only a subset of SRF groups or measures is used, we obtain unbiased estimates of performance for all plans. Third, not all users will find star ratings easily interpretable and useful for

decision-making. However, higher MA star ratings are associated with both higher enrollment and lower disenrollment [40, 41], suggesting value to consumers. Still, user-testing of the HESS is needed to gauge its comprehensibility and actionability and whether incentives provided by the HESS address barriers to improving care for at-risk groups. Fourth, while race-and-ethnicity and dual-eligibility have been identified as key SRFs [14] and focus on them is likely to have positive spillover effects, incentivizing care for a limited number of SRFs could reduce visibility of other SRFs. If such evidence arose, additional stratified reporting or expansion of HESS SRFs could be considered. Finally, our use of data from 2013 to 2014 and 2015 to 2016 to investigate improvement was driven by data availability. Because this period spanned the implementation of the Affordable Care Act, the improvement reported here may not be typical of other periods.

CONCLUSIONS

We provide a proof-of-concept for a HESS that could serve to promote and incentivize excellent care for racial-and-ethnic minorities and dually eligible enrollees in MA plans, populations for which disparities persist. This methodology could potentially be extended to other SRFs, quality measures, and settings.

Acknowledgments: The authors would like to thank Biayna Darabidian for assistance with manuscript preparation.

Corresponding Author: Marc N. Elliott, PhD; RAND Corporation, Santa Monica, CA, USA (e-mail: Elliott@rand.org).

Funding Information Funding for this study was provided by contract GS-10F-0012Y/HHSM-500-2016-00097G from the Centers for Medicare & Medicaid Services to NCQA.

Compliance with Ethical Standards:

Conflict of Interest: The author declares that he/she does not have a conflict of interest.

Disclaimer: The views expressed in this article are the authors' and do not necessarily represent the views of the U.S. Department of Health and Human Services or Centers for Medicare & Medicaid Services.

APPENDIX 1: DATA SOURCES, RACE-AND-ETHNICITY, AND MEASURES

The Medicare Advantage Consumer Assessment of Healthcare Providers and Systems Surveys

The Medicare Advantage Consumer Assessment of Healthcare Providers and Systems (CAHPS) [42] surveys are mail surveys with telephone follow-ups based on a stratified random sample of Medicare beneficiaries, with plans serving as strata. Surveys represent all Medicare Advantage (MA) beneficiaries from plans that either were required to report (minimum of 600 eligible enrollees) or reported voluntarily

(450–599 enrollees). Data were missing for fewer than 4% of cases for all CAHPS case-mix adjuster variables in 2013–2014 (and at similar rates other years 2015–2016) and were imputed using within-plan means [43]. Data were weighted to represent the Medicare population within each county and plan, followed by a raking procedure [44] to match weighted sample distributions within each plan of 10 beneficiary characteristics available from administrative data.

The full specifications for the CAHPS measures used here are as follows:

- *Doctor communication (four-item composite):* Respondents were asked how often their personal doctor (a) explained things in a way that was easy to understand, (b) listened to them carefully, (c) spent enough time with them, and (d) showed respect for what they had to say.
- *Ease of getting needed care (two-item composite):* Respondents were asked to assess how often it was (a) easy to get appointments with specialists, and (b) to get the care, tests, or treatment they thought they needed through their health plan.
- *Getting care quickly (two-item composite):* Respondents were asked to assess (a) how often they received care promptly if they thought that they needed it right away, and (b) how often they got an appointment for care at a doctor's office or clinic as soon as they thought they needed it.
- *Ease of getting needed prescription drugs (three-item composite):* Respondents enrolled in a prescription drug plan were asked (a) how often it was easy to use their plan to get the drugs their doctor prescribed, (b) how often it was easy to use their plan to fill prescriptions at a local pharmacy, and (c) how often it was easy to use their plan to fill prescriptions by mail.
- *Medicare customer service (three-item composite):* If beneficiaries called their plan's customer service, respondents were asked (a) how often their plan's customer service gave them information or help they needed, and (b) how often their plan's customer service staff treated them with courtesy and respect.
- *Coordination of care (six-item composite):* The frequency with which a respondent's personal care doctor was aware of care he or she received from specialists, the frequency with which the doctor's office provided the respondent with test results, and other aspects that related to the degree to which one's care was coordinated.
- *Flu immunization:* Indicator of getting a vaccine (flu shot) in the past year. Note: flu immunization is not case-mix adjusted.

Information on sample sizes for each year used for analysis are given below:

Year	Plans	Eligible enrollees	Completes	Response rate (%)
2016	441	364,538	153,866	42.2
2015	466	382,857	161,219	42.1
2014	454	448,219	200,469	44.7
2013	463	443,087	203,736	46.0

Information on Race-and-Ethnicity in CAHPS

The CAHPS survey asked beneficiaries, “Are you of Hispanic or Latino origin or descent?” The response options were the following: “Yes, Hispanic or Latino” and “No, not Hispanic or Latino.” The survey then asked, “What is your race? Please mark one or more,” with response options of “White,” “Black or African American,” “Asian,” “Native Hawaiian or other Pacific Islander,” and “American Indian or Alaska Native.” Following a U.S. Office of Management and Budget approach, answers to these two questions were used to classify respondents into one of seven mutually exclusive categories: Hispanic, multiracial, American Indian/Alaska Native (AI/AN), Asian/Pacific Islander (API), black, white, or unknown.

- Respondents who endorsed Hispanic ethnicity were classified as Hispanic regardless of races endorsed.
- Non-Hispanic respondents who endorsed two or more races were classified as multiracial, with a single exception: Those who selected both “Asian” and “Native Hawaiian or other Pacific Islander” but no other race were classified as API.
- Non-Hispanic respondents who selected exactly one race were classified as AI/AN, API, black, or white, according to their responses.
- Respondents without data regarding race-and-ethnicity were classified as unknown.
- We do not include the multiracial, unknown, or AI/AN groups because too few plans are measurable for these SRF groups.

The Healthcare Effectiveness Data and Information Set

The Healthcare Effectiveness Data and Information Set (HEDIS) [43] consists of 92 measures across six domains (National Committee for Quality Assurance [NCQA], 2018). These domains are effectiveness of care, access/availability of care, experience of care, utilization and risk-adjusted utilization, health plan descriptive information, and measures collected using electronic clinical data systems. HEDIS measures are developed, tested, and validated under the direction of the National Committee for Quality Assurance. HEDIS measurement data are gathered from a variety of sources including member surveys, insurance claims and medical records for hospitalizations, outpatient visits, procedures, medications, labs, imaging, and other services.

The descriptions of the HEDIS measures used here are as follows:

- *Breast cancer screening*: Indicator of whether appropriate screening for breast cancer took place, limited to women aged 50–74 years. Because of a 2013–2016 specification change, breast cancer screening was not eligible for an improvement score.
- *Colorectal cancer screening*: Indicator of whether appropriate screening for colorectal cancer was received, limited to enrollees aged 50–75.

- *Diabetes care: nephropathy*: Indicator of whether medical attention for nephropathy took place in the past year, limited to enrollees aged 18–75 years with diabetes.
- *Diabetes care: retinal eye exam*: Indicator of whether a retinal eye exam was performed in the past year, limited to enrollees aged 18–75 years with diabetes.
- *Adult BMI assessment*: Indicator of whether the patient’s body mass index was documented in the past two years, limited to enrollees aged 18–74 years who had an outpatient visit.

The total number of enrollees in all plans present in both years of the two time periods eligible for analysis for each measure is given below.

Measure	2013–2014	2015–2016
Breast cancer screening ^a		5,488,654
Colorectal cancer screening	992,247	1,114,647
Diabetes care: nephropathy	612,662	667,405
Diabetes care: retinal eye exam	597,900	624,790
Adult BMI assessment	1,023,408	1,079,274

^aWe did not include breast cancer screening in our 2013–2014 analysis because the measure denominator criteria changed from 2013 to 2014. Prior to 2014, the measure was for women aged 40–69 years; in 2014, it changed to 50–74 years

APPENDIX 2: ADDITIONAL DETAILS

Measurability

Plans were considered measurable if they had a publicly-reported Part C summary star which measures overall plan performance [32, 38], at least one domain-specific star which measures HEDIS- or CAHPS-specific plan performance, and 500 or more enrollees in 2016; 10 plans were not measurable for CAHPS because they had no overall CAHPS stars.

Cross-sectional Score Computation

We outline the steps to compute the cross-sectional score.

1. For every plan that is measurable (minimum sample size of 100 measure completes and reliability of 0.7 or greater) and for a given SRF group and measure, case-mix-adjusted least squares mean estimates were obtained for plan performance on the 0 to 100 scale. Linear models were used for all measures, including binary ones. Linear regression was used for binary measures for ease of interpretation of the results as percentage point changes in probability (the binomial distribution is asymptotically normal, so at large sample sizes, as in this study, linear probability models produce very similar results to logistic regressions [44, 45]). Each plan received a performance score for each SRF group and measure based on its predicted value from the model.
2. Scores were then standardized to put them on a common scale and to ensure equal influence of measures in a set

(HEDIS or CAHPS). For each combination of measure and SRF group, estimates were standardized to put all measures on comparable scales. The standardized estimate was computed as the difference between the performance estimate and the grand mean of the measure across all patients in the dataset, not just reportable groups and plans, all divided by the standard deviation of plan-specific performance. This standard deviation was computed as the square root of the plan-level variance component in a linear mixed-effect model for the measure based on all data, including a plan random effect and any available case-mix adjustment variables.

3. The standardized performance estimates were then entered into the MA Part C clustering algorithm [32], separately for each combination of measure and SRF group, producing one set of cut-points for each SRF group and measure. The input to the clustering algorithm was data from reportable plans only. The output of this clustering algorithm is on a 1–5 star scale. The clustering algorithm assigns plans into the five groups that maximize the ratio of between-plan to within-plan differences in scores and was chosen for comparability to existing methods for scoring these quality measures for MA plans [27].
4. For race-and-ethnicity, stars were averaged across non-whiteracial-and-ethnic groups for each measure and plan to produce “star roll-ups.” For dual-eligibility, where there is only one SRF group, no such averaging was possible or necessary.
5. The star roll-ups were averaged across measures to obtain a cross-sectional score for each plan, which was rounded to the nearest whole star.

Between-Plan Improvement Score Computation

We outline the steps to compute between-plan improvement, which is one of two components of the overall improvement score for a plan.

1. For every plan that is measurable and for a given SRF group and measure, case-mix adjusted estimates were obtained for the difference of the improvement for a given SRF group from the national improvement for that SRF group on that measure, on a -200 to $+200$ scale. These are differences in differences (DIDs). If the DID is positive but the change within the plan is negative, the DID is recoded to 0.
 - To calculate national improvement for each combination of SRF group (e.g., Hispanic, black, API for race-and-ethnicity) and measure, we ran a regression model using all available plans that exist in 2013–2016. Each model included fixed effects for standard case-mix adjusters (for CAHPS, not for HEDIS) and an indicator for “follow-

up,” meaning the survey occurred in 2015–2016. Change from 2013–2014 to 2015–2016 was estimated from these models. A national improvement score for each SRF group and measure was given by the predicted value from the model.

- To calculate plan improvement, we re-ran the improvement models described in the previous step, stratified by plan. Plan-specific improvement scores for each SRF group and measure were given by the predicted value from these models.
 - Linear models on a 0–100 scale were used for all measures, including binary ones.
2. Scores were then standardized to put them on a common scale. The standardized estimate was computed as the DID divided by the standard deviation of plan-specific performance. This standard deviation was computed as the square root of the plan-level variance component in a linear mixed-effect model for the measure based on all 2015–2016 data, including a plan random effect and any available case-mix adjustment variables. This quantity is the same as in the standardization for the cross-sectional score.
 3. For race-and-ethnicity, scores were averaged across non-whiteracial-and-ethnic groups for each measure and plan to produce “score roll-ups.” For dual-eligibility, where there is only one SRF group, no such averaging was possible or necessary.
 4. Score roll-ups were then were averaged across measures to arrive at an average standardized improvement estimate.

Within-Plan Improvement Score Computation

We outline the steps to compute within-plan improvement, which is one of two components of the overall improvement score for a plan.

1. We first identified the *leading group* at baseline for each plan and measure. To identify this highest-performing group at baseline, we ran regression models stratified by plan for just those units that are measurable at both follow-up (2015–2016) and baseline years (2013–2014), using data from all four years. Each model had fixed effects for the follow-up indicator, the indicators for SRF group (black, Hispanic, and API for race-and-ethnicity or a single indicator for dual-eligibility), the interaction of the follow-up indicator with the SRF group indicators and any case-mix adjustment variables. The leading group was identified as the SRF group with the highest positive coefficient for SRF group. If all SRF group coefficients were negative, then the reference group (non-Hispanic white for race-and-ethnicity or non-dually-eligible for dual-eligibility) was considered the leading group.

2. For each measure, within-plan improvement estimates were obtained as the amount by which each *lagging* (non-leading) group gained on the leading group for each plan (“the estimate of gain”). We also measured how much each lagging group improved irrespective of the leading group (“raw improvement”). The estimate of gain corresponds to the coefficient for the interaction between follow-up indicator and the SRF group, and the raw improvement corresponds to the estimate of gain plus the coefficient for follow-up. Within-plan improvement for each SRF group and measure was taken to be the estimate of gain for that group with the following exceptions:
 - a) If a lagging group had positive gain but negative raw improvement, within-plan improvement for that group and measure was set to 0.
 - b) Within-plan improvement was restricted to take a value no larger than the initial disparity between the leading and lagging groups. So, for example, if whites were the leading group, and blacks were initially 5 points behind them, then the highest within-plan improvement estimate for blacks would be 5. Thus, matching and surpassing the leading group were rewarded equally and closing a large difference was given more credit than closing a small one.
- Note that within-plan improvement could be negative. If blacks were initially 5 points behind the leading group, and in the follow-up years the disparity grew to 7 points, then their within-plan improvement for that measure would be -2 . Note also that within-plan improvement includes improvement for non-disadvantaged groups (i.e., whites for race-and-ethnicity and non-dually-eligible for dual-eligibility) when they are not the leading group for a plan and measure.
3. Within-plan improvement scores were then standardized to put them on a common scale. The standardized estimate was computed as the within-plan improvement divided by the standard deviation of plan-specific performance. This standard deviation was computed as the square root of the plan-level variance component in a linear mixed-effect model for the measure based on all 2015–2016 data, including a plan random effect and any available case-mix adjustment variables. This quantity is the same as in the standardization for the cross-sectional score.
 4. Within each plan, standardized within-plan improvement scores were averaged across all lagging groups separately for each measure and then averaged across measures.

Final Improvement Score Computation

A final improvement score was computed by running the MA Part C clustering algorithm on the average of the within-plan and between-plan improvement scores.

REFERENCES

1. **Link BG, Phelan J.** Social conditions as fundamental causes of disease. *J Health Soc Behav.* 1995;Spec No:80–94.
2. **Jiang HJ, Andrews R, Stryer D, Friedman B.** Racial/ethnic disparities in potentially preventable readmissions: the case of diabetes. *Am J Public Health* 2005;95(9):1561–7.
3. **Kim H, Ross JS, Melkus GD, Zhao Z, Boockvar K.** Scheduled and unscheduled hospital readmissions among patients with diabetes. *Am J Manag Care* 2010;16(10):760–7.
4. **Kroch E, Duan M, Martin J, Bankowitz RA.** Patient factors predictive of hospital readmissions within 30 days. *J Health Qual* 2016;38(2):106–15.
5. **Oronce CI, Shao H, Shi L.** Disparities in 30-day readmissions after total hip arthroplasty. *Med Care* 2015;53(11):924–30.
6. **Agarwal S, Garg A, Parashar A, Jaber WA, Menon V.** Outcomes and resource utilization in ST-elevation myocardial infarction in the United States: evidence for socioeconomic disparities. *J Am Heart Assoc* 2014;3(6):e001057.
7. **Bennett KM, Scarborough JE, Pappas TN, Kepler TB.** Patient socioeconomic status is an independent predictor of operative mortality. *Ann Surg* 2010;252(3):552–7; discussion 7–8.
8. **LaPar DJ, Bhamidipati CM, Harris DA, Kozower BD, Jones DR, Kron IL, et al.** Gender, race, and socioeconomic status affects outcomes after lung cancer resections in the United States. *Ann Thorac Surg* 2011;92(2):434–9.
9. Institute of Medicine. Accounting for social risk factors in Medicare payment: identifying social risk factors. Washington, DC: The National Academies Press; 2016. 110 p.
10. **Durfey SNM, Kind AJH, Gutman R, Monteiro K, Buckingham WR, DuGoff EH, et al.** Impact of risk adjustment for socioeconomic status on Medicare Advantage plan quality rankings. *Health Aff (Millwood)* 2018;37(7):1065–72.
11. **Blustein J, Weissman JS, Ryan AM, Doran T, Hasnain-Wynia R.** Analysis raises questions on whether pay-for-performance in Medicaid can efficiently reduce racial and ethnic disparities. *Health Aff (Millwood)*. 2011;30(6):1165–75.
12. **Chin MH.** Creating the business case for achieving health equity. *J Gen Intern Med* 2016;31(7):792–6.
13. **Joynt KE, De Lew N, Sheingold SH, Conway PH, Goodrich K, Epstein AM.** Should Medicare value-based purchasing take social risk into account? *N Engl J Med* 2017;376(6):510–3.
14. National Academies of Sciences E, Medicine. Systems Practices for the Care of Socially At-Risk Populations. Washington, DC: The National Academies Press; 2016. 94 p.
15. National Academies of Sciences E, Medicine. Accounting for Social Risk Factors in Medicare Payment. **Kwan LY, Stratton K, Steinwachs DM,** editors. Washington, DC: The National Academies Press; 2017. 580 p.
16. **Chien AT, Chin MH, Davis AM, Casalino LP.** Pay for performance, public reporting, and racial disparities in health care: how are programs being designed? *Med Care Res Rev.* 2007;64(5 Suppl):283s–304s.
17. **Damberg CL, Elliott MN, Ewing BA.** Pay-for-performance schemes that use patient and provider categories would reduce payment disparities. *Health Aff (Millwood)*. 2015;34(1):134–42.
18. **Casalino LP, Elster A, Eisenberg A, Lewis E, Montgomery J, Ramos D.** Will pay-for-performance and quality reporting affect health care disparities? *Health Aff (Millwood)*. 2007;26(3):w405–14.
19. **Casalino LP.** The unintended consequences of measuring quality on the quality of medical care. *N Engl J Med* 1999;341(15):1147–50.
20. **Goldstein E, Elliott MN, Lehrman WG, Hambarsoomian K, Giordano LA.** Racial/ethnic differences in patients’ perceptions of inpatient care using the HCAHPS survey. *Med Care Res Rev* 2009;67(1):74–92.
21. **Quigley DD, Elliott MN, Hambarsoomian K, Wilson-Frederick SM, Lehrman WG, Agniel D, et al.** Inpatient care experiences differ by preferred language within racial/ethnic groups. *Health Serv Res* 2019;54(S1):263–74.
22. **Trivedi AN, Zaslavsky AM, Schneider EC, Ayanian JZ.** Relationship between quality of care and racial disparities in Medicare health plans. *JAMA.* 2006;296(16):1998–2004.
23. **Lurie N, Zhan C, Sangl J, Bierman AS, Sekscenski ES.** Variation in racial and ethnic differences in consumer assessments of health care. *Am J Manag Care* 2003;9(7):502–9.
24. Centers for Medicare and Medicaid Services. 2018 Part C and D Medicare Star Ratings Data Baltimore, MD: Centers for Medicare and Medicaid Services; 2018 [Available from: <https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovGenIn/Downloads/2018-Part-C-and-D-Medicare-Star-Ratings-Data-v05-10-2018-.zip>].

25. Centers for Medicare and Medicaid Services. Part C and D Performance Data Stratified by Race and Ethnicity Baltimore, MD2018 [Available from: <https://www.cms.gov/About-CMS/Agency-Information/OMH/research-and-data/statistics-and-data/stratified-reporting.html>].
26. **Martino SC, Weinick RM, Kanouse DE, Brown JA, Haviland AM, Goldstein E, et al.** Reporting CAHPS and HEDIS data by race/ethnicity for Medicare beneficiaries. *Health Serv Res* 2013;48(2 Pt 1):417–34.
27. **Sorbero ME, Paddock SM, Damberg CL, Haas A, Kommareddi M, Tolpadi A, et al.** Adjusting Medicare Advantage Star Ratings for socioeconomic status and disability. *Am J Manag Care* 2018;24(9):e285–e91.
28. **Haas A, Elliott MN, Dembosky JW, Adams JL, Wilson-Frederick SM, Mallett JS, et al.** Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. *Health Serv Res*. 2018.
29. **Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N.** Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcome Res Methodol* 2009;9(2):69.
30. Centers for Medicare and Medicaid Services. Seniors & Medicare and Medicaid Enrollees [Available from: <https://www.medicaid.gov/medicaid/eligibility/medicaid-enrollees/index.html>].
31. **Lyratzopoulos G, Elliott MN, Barbieri JM, Staetsky L, Paddison CA, Campbell J, et al.** How can health care organizations be reliably compared?: lessons from a national survey of patient experience. *Med Care* 2011;49(8):724–33.
32. Prescription Drug Coverage - General Information. In: Services CIMA, editor. Baltimore, MD2018.
33. **Hasnain-Wynia R, Baker DW, Nerenz D, Feinglass J, Beal AC, Landrum MB, et al.** Disparities in health care are driven by where minority patients seek care: examination of the hospital quality alliance measures. *Arch Intern Med* 2007;167(12):1233–9.
34. **Jha AK, Orav EJ, Li Z, Epstein AM.** Concentration and quality of hospitals that care for elderly black patients. *Arch Intern Med* 2007;167(11):1177–82.
35. **Rosenthal MB, Frank RG, Li Z, Epstein AM.** Early experience with pay-for-performance: from concept to practice. *JAMA*. 2005;294(14):1788–93.
36. **Rosenthal MB, Dudley RA.** Pay-for-performance: will the latest payment trend improve care? *JAMA*. 2007;297(7):740–4.
37. **Elliott MN, Beckett MK, Lehrman WG, Cleary P, Cohea CW, Giordano LA, et al.** Understanding the role played by Medicare's patient experience points system in hospital reimbursement. *Health Aff (Millwood)*. 2016;35(9):1673–80.
38. Part C and D Performance Data. In: Services CIMA, editor. Baltimore, MD2018.
39. **Weech-Maldonado R, Elliott MN, Adams JL, Haviland AM, Klein DJ, Hambarsoomian K, et al.** Do racial/ethnic disparities in quality and patient experience within Medicare plans generalize across measures and racial/ethnic groups? *Health Serv Res* 2015;50(6):1829–49.
40. **Reid RO, Deb P, Howell BL, Shrank WH.** Association between Medicare Advantage plan star ratings and enrollment. *JAMA*. 2013;309(3):267–74.
41. **Li G, Trivedi AN, Galarraga O, Chernew ME, Weiner DE, Mor V.** Medicare Advantage ratings and voluntary disenrollment among patients with end-stage renal disease. *Health Aff (Millwood)*. 2018;37(1):70–7.
42. Centers for Medicare and Medicaid Services. Quality Assurance Protocols & Technical Specifications. Centers for Medicare and Medicaid Services; 2017.
43. Health Services Advisory Group. MA & PDP CAHPS Variables Used as Case-Mix Adjustors 1998–2006: Health Services Advisory Group; [Available from: https://www.ma-pdpcahps.org/globalassets/ma-pdp/case-mix-adjustment/case-mix_1998_2017_pdf.pdf].
44. **Pedroza C, Thanh Truong VT.** Performance of models for estimating absolute risk difference in multicenter trials with binary outcome. *BMC Med Res Methodol* 2016;16(1):113.
45. **Hosmer DW, Lemeshow S.** Applied Logistic Regression, Second Edition. Hoboken: John Wiley & Sons, Inc.; 2005.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.