JOURNAL OF PALLIATIVE MEDICINE
Volume 25, Number 7, 2022
Mary Ann Liebert, Inc.
DOI: 10.1089/jpm.2021.0423

Adjusting for Patient Characteristics to Compare Quality of Care Provided by Serious Illness Programs

Maria DeYoreo, PhD,[1] Rebecca Anhang Price, PhD,[2] Cheryl K. Montemayor, BS,[1] Anagha Tolpadi, MS,[1] Melissa Bradley, BA,[2] Danielle Schlang, MA,[3] Joan M. Teno, MD, MS,[4] Paul D. Cleary, PhD,[5] and Marc N. Elliott, PhD[1]

Abstract

Background: To compare serious illness programs (SIPs) using recently developed patient experience measures, adjustment must be made for patient characteristics not under control of the programs.
Objectives: To develop a case-mix adjustment model to enable fair comparison of patient experience between SIPs by investigating the roles of patient characteristics, proxy response, and mode of survey administration (mail-only vs. mail with telephone follow-up) in survey responses.
Methods: Using survey data from 2263 patients from 32 home-based SIPs across the United States, we fit regression models to assess the association between patient-level variables and scores for seven quality measures (Communication, Care Coordination, Help for Symptoms, Planning for Care, Support for Family and Friends, and two global assessments of care). Characteristics that are not consequences of the care the program delivered were considered as adjustors.
Results: Final recommended case-mix adjustors are age, education, primary diagnosis, self-reported functional status, self-rated physical health, self-rated mental health, proxy respondent use, and response percentile (a measure of how soon a person responded compared with others in the same program and mode). Age, primary diagnosis, self-rated mental health, and proxy respondent use had the most impact on program-level scores. We also recommend adjusting for mode of survey administration. We find that up to 12 percent of pairs of programs would have their rankings reversed by adjustment.
Conclusions: To ensure fair comparison of programs, scores should be case-mix adjusted for variables that influence patients' reports about care quality, but are not under the control of the program administering care.

Keywords: case-mix; patient experience; patient surveys; serious illness care

Introduction

There is a growing number of high-need, high-cost seriously ill individuals in the United States. These individuals require well-coordinated care that is tailored to their changing health and social needs and is aligned with their preferences and goals.[1] To address these needs, a growing number of programs offer care for seriously ill individuals in their homes. Serious illness programs (SIPs) vary greatly in their structure and staffing, with some focusing on specific populations, such as patients with advanced cancer, and others providing care to those with multiple chronic conditions or frailty.[2–4]

Patient and family centeredness of care are central to the provision of high-quality serious illness care. Initiatives that compare quality of serious illness care and determine payment based on the value and quality of care provided must include measures of patient and family care experiences.[5]

[1]RAND Corporation, Santa Monica, California, USA.
[2]RAND Corporation, Arlington, Virginia, USA.
[3]RAND Corporation, Boston, Massachusetts, USA.
[4]Oregon Health and Science University, Portland, Oregon, USA.
[5]Yale School of Public Health, New Haven, Connecticut, USA.
Accepted December 8, 2021.

Frail, elderly, seriously ill patients are particularly at risk for poor-quality care. However, until recently, there have been few actionable and fair accountability measures for serious illness care in general, and for SIPs in particular.[6] To address this gap, we developed a set of valid and reliable survey-based quality measures to assess patient care experiences in SIPs.[7,8]

Patient-level variables can be related to how patients respond to survey questions.[9–13] These include patient characteristics, mode of survey administration, and the use of a proxy respondent. Since there is great variability across programs with regard to patient mix, program-level scores can be compared fairly only after adjusting for patient characteristics that influence how patients respond to questions about quality of care, but are not under the control of the SIP. Failure to account for differences in patient mix across SIPs through analytic adjustment can result in important errors identifying the highest performing programs.

Comparing the performance of programs assuming comparable patient populations is important to ensure fairness in comparing programs. In addition, quality comparison should not result in incentives that result in programs not providing care to vulnerable patient subgroups. To incentivize high performance on vulnerable patient subgroups, adjustment can be paired with stratified reporting of performance scores so that the quality of care provided to at-risk groups can be highlighted.[14]

Other surveys of patient and family experiences of care have found that age, education, health status, and language spoken at home are important case-mix variables for which to adjust.[10–13,15] Other variables may also be considered for adjustment. For example, serious illness care experiences are sometimes reported by family members or friends who serve as proxy respondents for patients who are unable to respond for themselves. Thus, adjustment should also be made for the use of a proxy respondent, as proxy responses are known to differ from patient responses.[16] Mode of survey administration can also affect patient evaluations of health care. Previous studies have generally found more positive assessment of care by telephone than mail,[10,17,18] although one found in the hospice setting that caregivers in mail-only mode reported significantly better experiences than those in telephone-only mode.[19]

In this study, we assess patterns of survey responses by available patient characteristics and by randomized mode of survey administration. We develop and recommend a case-mix adjustment model to allow for fair comparison of survey-based quality measure scores between SIPs.

Materials and Methods

Sample

From October 2019 to January 2020, we conducted a field test of a survey of care experiences among patients from 32 geographically diverse SIPs across the United States (Northeast (6), South (7), Midwest (8), West (9), and national (2 programs operated in all four regions)). There is no standardized national definition or directory for SIPs. Therefore, to identify programs for participation in the field test, we compiled a master list of programs that had been previously identified as providing care to the seriously ill.[7] Programs were eligible to be recruited for the field test if they provide medical care in patients' homes; almost all of these programs provide after-hour access to care either by phone or in person and have either a physician or a nurse practitioner on the team who makes home visits.

Adults (18 years of age or older at the time the sample was selected) within each program were eligible to be sampled if they received care at a private home or assisted living facility and had been receiving care from the program for 3 to 24 months as of the date the sample was selected. The survey was designed to be completed by the patient, or if needed, a proxy respondent (i.e., family caregiver). Patients were randomly assigned to mail-only or mixed (mail with telephone follow-up) survey mode. The survey was available in both English and Spanish. The study was approved by the RAND Corporation's Human Subjects Protection Committee, which serves as RAND's IRB.

Survey instrument and evaluative measures

The field test survey instrument contained 56 items that assessed several aspects of serious illness care, with seven quality measures derived from the survey, including five composite measures (Communication, Care Coordination, Help for Symptoms, Planning for Care, and Support for Family and Friends) and two single-item global measures (Overall Rating of the Program and Willingness to Recommend the Program; see Appendix Table A1). Description of the development and validation of these measures is available elsewhere,[7] and final, more concise versions of the survey instrument are available free online.[8]

Responses for all evaluative items were transformed from the original response scale to 0/100 values using top-box scoring. Top-box coding is widely used in public reporting initiatives to promote ease of comprehension by consumers;[20] it classifies the response corresponding to the best quality as 100 (e.g., ''always'') and all others as 0, with the exception that for the overall rating item, both 9 and 10 are classified as top box.[21] Tailored nonapplicable responses (e.g., ''I do not take any medicines'') are removed from the denominator.

Case-mix adjustment variables

We identified patient characteristics that are exogenous to the care provided by the program and both strongly predict survey measure scores and impact program scores for at least one outcome measure. Table 1 describes characteristics of respondents.

Administrative data provided by the SIPs contained information about the patient's sex, age, and primary diagnosis. Survey response data contained information about the patient's education, language spoken at home, and self-reported ratings of mental health, physical health, and functional status. We developed a functional status measure that combined two survey items assessing the respondent's self-reported ability to get out of bed and ability to leave home. In keeping with adjustments for other patient experience surveys, we also created a variable for whether a proxy assisted with completion of the survey on behalf of the patient, using a categorical variable indicating whether and how the proxy helped (''proxy answered questions,'' ''proxy helped in another way,'' and ''survey was completed by patient and not a proxy'').[16,22] The proxy respondent variable captures severity of illness and/or cognitive impairments, and thus may be correlated with other adjustors, such as primary diagnosis and self-reported functional status.

Table 1. Characteristics of Respondents

Characteristic	*Respondents (N = 2263) (%)*
Sex	
Female	58.4
Male	41.6
Age	
18–54	4.4
55–64	9.4
65–69	7.2
70–74	10.6
75–79	12.4
80–84	17.2
85–89	19.5
90 or older	19.4
Primary diagnosis	
Cancer	13.6
Alzheimer's or dementia	9.4
Other	77.1

Percentages calculated excluding missing values as all variables had negligible missingness.

We also considered ''response percentile,'' defined as the rank-ordered time between initiation of survey administration and response for each respondent relative to all eligible patients within program and mode, scaled from 0 to 1. This quantity captures both the program response rate (RR) and how soon a person responded compared with others in the same program and mode.[23–25]

Statistical methods

We used linear regression models to estimate the effect of each potential case-mix adjustor on the survey measure scores and assessed whether the regression coefficient associated with the adjustor was statistically significantly different from zero. We also evaluated the impact of each adjustor on program-level scores by comparing the program-level scores with and without the adjustor of interest.

We first fit regression models where the outcomes were the composite scores or global rating items and the predictors included all candidate case-mix adjustor variables, mode of survey administration (mail-only vs. mixed mode), and program fixed effects. Mode of survey administration has been shown to affect responses to patient experience surveys.[10,19] The regression coefficients can be interpreted as the average effect of each patient characteristic on outcomes within a program. We identified the adjustors that were statistically significantly predictive of at least one outcome ($p < 0.01$) and interpreted the effect of each adjustor on assessments.

To evaluate the impact of each case-mix adjustor on adjustments, we fit a series of models that removed one candidate adjustor at a time. For each composite score and global rating item, we calculated the correlation between the adjusted program-level scores from the full models and the adjusted program-level scores from the models that left out the case-mix adjustor variable of interest, assuming each program had population-average case-mix and mode of survey administration. Adjusted program scores were generated for each item using the estimated regression coefficients and the characteristics of respondents in the program. The quantity $1-r^2$ represents the proportion of variance in the adjusted scores marginally associated with adjustment for that variable and indexes each adjustor's marginal impact on program-level scores. Only characteristics that vary among programs and predict patient responses affect program-level scores.

We additionally reported Kendall's tau, *t,* a rank-based measure of correlation (−1 to +1) that can be used to calculate $p = (1-t)/2$, which can be interpreted as the chance of mistakenly ranking one program as better than the other without adjustment, or the proportion of pairs of programs whose relative rankings are reversed by adjustment. Thus, $t = 1$ indicates that adjustment has no effect on relative rankings, and $t = 0.8$ indicates that there is a 10% chance that one program will be mistakenly ranked as higher than another (without adjustment).

Our recommended set of case-mix adjustors included all variables that were statistically significantly associated with respondent evaluations ($p < 0.01$) and which have an impact ($1-r^2$) of at least 1% for one or more outcome measures. To determine the final case-mix adjustment (CMA) model, we also incorporated feedback from our team's expert advisors regarding variables that did not meet these empirical criteria, but which are important for face validity. For example, if people think that patients with poor function tend to give lower ratings and/or tend to report negative experiences, facilities with many such patients may think they are being unfairly evaluated unless there is an adjustment for the proportion of such patients in each facility, irrespective of empirical evidence of impact. Once we identified an appropriate CMA model, we determined the overall impact of adjustment on program scores by comparing case-mix adjusted program-level scores to those not adjusted for case-mix.

Results

Of the 6481 patients sampled, 271 (4.2%) were determined to be ineligible after sampling. There were 2263 eligible respondents, for an overall 36.4% RR (30.4% in mail-only mode and 42.5% in mixed mode). The average number of respondents per program was 71.

An initial multivariate model that adjusted for all candidate CMA variables simultaneously was used for first-stage empirical screening (results not shown). Based on input from the team's expert advisors and given wide variability in the distribution of functional and health status across SIPs, we included in this model but excluded from empirical screening both self-rated physical health and functional status for the sake of face validity. For all other variables, we required that a candidate predictor be statistically significant with at least one of the seven outcomes at the 0.01 significance level. Sex was not significant for any outcome and was excluded on this basis.

The Appendix Table A2 provides the regression coefficient estimates and standard errors from a multivariate model that included all the proposed candidate adjustors, but sex. Directions of association were generally similar to what has been seen previously in the patient experience literature.[11] Response percentile was significantly negatively associated with assessments for two of seven outcomes (Communication and Care Coordination). Education was only significantly

Box 1. Final Recommended Variables for Case-Mix Adjustment
Case-mix adjustors
Age
Education
Response percentile
Primary diagnosis
Proxy respondent
Self-reported functional status
Self-rated physical health
Self-rated mental health

associated with overall rating (negatively); age was significantly associated with two outcomes (negatively, which is atypical).[11] Diagnosis was significant for two outcomes (Help for Symptoms and Planning for Care). For these outcomes, compared to all other diagnoses, cancer diagnosis was associated with significantly more positive assessments and Alzheimer's and other dementia was associated with significantly more positive assessments for Help for Symptoms.

Language was borderline significantly associated with one outcome ($p = 0.01$ for Planning for Care, with Spanish language associated with more negative assessments). Proxy use was significant for four outcomes, with both proxy response and other help from a proxy being associated with more positive responses compared to no proxy. Those with poor function tended to respond more negatively (although functional status was only significant for Care Coordination). Self-rated physical health was not significantly associated with any outcome. Self-rated mental health was significantly associated with five of the outcomes; better ratings of mental health were generally associated with more positive assessments.

Regression coefficients estimate the tendency of respondents to respond more positively or negatively and are used to calculate the adjustments to top-box scores. For example, patients with more than a four-year college degree were 9% less likely to provide the top-box response for overall rating compared to those with a high school degree or graduate equivalency degree. The adjustments exactly counteract the differences in response tendency so that the same level of performance results in the same score, irrespective of patients' characteristics.

To determine which case-mix variables should remain in the model, we calculate the impact of an adjustor as $1 - r^2$. All candidate adjustors subject to empirical screening meet the criteria of statistical significance for at least one outcome at the 0.01 level, in addition to $1 - r^2$ of at least 1% for at least one outcome, except language. We therefore recommend excluding language from the final set of case-mix adjustors.

Retaining self-rated physical health and functional status for face validity (as noted above), our final recommended set of case-mix adjustors consists of age, education, response percentile, primary diagnosis, proxy respondent, and self-reported functional status, physical health and mental health (Box 1). Table 2 contains the impact measure values for these variables. Larger values in Table 2 indicate that an adjustor plays a large role. Age was the most important adjustor overall, followed by diagnosis. Values indicate the proportion of variance in patient-level scores that are uniquely associated with a given predictor, with 3% indicating a moderate impact (impact values exceed 3% for age on two measures, diagnosis on one measure) and 1% indicating a small impact.

Table 3 summarizes the impact of adjustment with all variables proposed for the CMA model (''full adjustment'') on each outcome of interest. Full adjustment has the most impact on Help for Symptoms, with 12% of pairs of programs having their relative rankings reversed by adjustment and 10% of the variance in program scores attributable to extraneous factors controlled for by the case-mix adjustment model.

Mixed mode of survey administration is associated with slightly more positive assessments than mail-only administration for most outcomes (Appendix Table A2). However, when this survey is deployed, survey mode will be a choice that affects a program's entire sample (whereas in this experiment, it is a randomized at the patient level within programs), so survey mode will have a larger impact on scores than is suggested by these patient-level results.

Table 2. Impact of Each Case-Mix Adjustor on Program-Level Scores

Variable removed	*Communication*	*Care coordination*	*Help for symptoms*	*Planning for care*	*Support for family and friends*	*Overall rating of the program*	*Willingness to recommend the program*	*Mean impact*
Age	1.0	3.2	2.3	3.4	0.5	0.4	1.3	1.7
Diagnosis	0.7	0.2	3.9	1.3	0.5	0.6	0.3	1.1
Proxy respondent	0.6	0.6	1.2	0.8	2.3	0.1	0	0.8
Self-rated mental health	0.4	0.6	1.8	0.4	0.6	0.7	0.2	0.7
Response percentile	1.5	0.6	0.2	0	0.7	0.3	0	0.5
Education	0	0.3	0.3	0.4	0.7	1.2	0	0.4
Functional status	0.2	0.5	0.2	0.1	0	0.3	0.1	0.2
Self-rated physical health	0.1	0.1	0.2	0	0	0.2	0.1	0.1

Quantities shown represent the percent of adjustment attributable to each variable.

Correlations were calculated restricting to only those programs with 10 or more respondents (28 out of 32 programs). Leave-one-out impact measures were generated from adjusted program scores that adjust for all variables in table and include program fixed effects. Adjusted program scores were generated assuming each program had population-average case-mix and mode of survey administration. Table shows results for final set of recommend case-mix adjustors only.

Table 3. Overall Impact of Adjustment

Measure	*Percent of variance in program level scores due to adjustment*	*Kendall's Tau*[a]	*Percent of pairs of programs that switch rankings with adjustment*
Communication	5	0.88	6
Care coordination	6	0.86	7
Help for symptoms	10	0.76	12
Planning for care	8	0.81	10
Support for family and friends	6	0.85	8
Overall rating of the program	5	0.89	6
Willingness to recommend the program	3	0.93	4

[a]Kendall's tau is interpreted as the proportion of pairs of programs whose relative rankings would be reversed by adjustment. The percentage of pairs of programs that would switch rankings because of adjustment is calculated as $(1-t)/2\%$, where t is the value of Kendall's tau. Correlations and Kendall's tau were calculated restricting to only those programs with 10 or more respondents (28 out of 32 programs). Fully adjusted program scores were constructed from models that adjust for the final set of recommend case-mix adjustors and survey mode, and include program fixed effects. Adjusted program scores were generated assuming each program had population-average case-mix and mode of survey administration. Unadjusted program scores were constructed from models that adjust for survey mode and include program fixed effects.

The effect sizes for mixed mode are 0.42 (Help for Symptoms), 0.28 (Care Coordination), 0.25 (Planning for Care), 0.16 (Support for Family and Friends), 0.10 (Communication), –0.04 (Overall Rating), and 0.04 (Willingness to Recommend) program-level standard deviations (Elliott et al.),[10] where 0.2 and 0.5 are small and medium, respectively (Cohen, 1988).[26] Given that mixed mode results in higher scores to an appreciable extent, we recommend adjusting for mode of survey administration when comparing programs that administer the survey in different modes.[10]

Discussion

To ensure that comparison of program scores is fair across SIPs, scores should be adjusted for patient-level variables, which influence patients' reports about care quality, but are not under the control of the program administering care. We identified patient-level variables that are significantly associated with respondent evaluations within programs and have a meaningful impact on program-level scores for at least one outcome measure. The adjustors that had the most impact on program-level scores were age, primary diagnosis, and proxy respondent.

In general, older respondents responded more negatively. This result differs from the trend observed in the Consumer Assessment of Healthcare Providers and Systems Hospital Survey and on Medicare beneficiaries,[27,28] in which older patients responded more positively, as well as the finding that older adults generally report more positive experiences with their medical care and health plan services.[9] However, patients in the SIPs participating in our field test tended to be much older than respondents to other CAHPS surveys (nearly 2 in 5 survey respondents were 85 years of age or older), so the oldest age categories in this analysis are not necessarily well represented in other surveys.

We found that diagnoses of cancer and Alzheimer's or other dementia were associated with more positive assessments than diagnosis of a range of other health conditions, and use of a proxy respondent was associated with more positive responses than when patients completed the survey themselves. In contrast, use of proxy respondents was found to be associated with less positive assessments of care for Medicare.[16] Those with poor function tended to respond more negatively, and those with better ratings of mental health responded more positively, consistent with the findings that better health status and mental health are associated with more positive assessments.[10,29] We found that speaking Spanish was associated with more positive assessments.

While mode of survey administration was not a statistically significant predictor of patient assessments for any outcome variable, use of mixed-mode administration was associated with more positive (but not significant) assessments for five of the seven outcomes. If the survey is implemented in accountability initiatives in which SIPs can select the mode of administration, adjustment for mode should be carefully considered to ensure that program-level decisions regarding mode of administration do not result in unfair comparisons across programs. In addition, since mode is likely to be the same for all patients within a given program when the survey is implemented, rather than randomized at the patient level, the choice of mode will have a larger impact on program-level scores.

To increase awareness around disparities and incentivize high performance on vulnerable patient subgroups, we recommend pairing case-mix adjustment with stratified public reporting of scores by patient subgroups, such as functional status and primary diagnosis.

Our study has several limitations. Although we developed a comprehensive list of SIPs based on numerous sources, there exists no complete list of these programs, and not all programs approached agreed to participate; therefore, programs in our field test may not be representative of all those who provide home-based serious illness care across the United States. Another limitation relates to the variables available for case-mix analyses. Additional information regarding patient diagnosis and primary payer for services may be useful to further refine the CMA model.

As public and private sector initiatives that aim to extend care to seriously ill individuals expand, case-mix adjusted quality measurement is critical for fairly comparing programs that provide these services and assessing quality of care over time, particularly as the characteristics of patients served evolve.

Conclusion

To ensure that comparison of program scores is fair, SIPs' quality measure scores should be adjusted for patient-level variables that influence patients' reports about care quality, but are not under the control of the program administering care, as well as mode of survey administration. Age, primary diagnosis, self-rated mental health, and proxy respondent were the case-mix adjustors with the most impact on program-level scores, but we also recommend adjusting for response percentile (a measure of how soon a person responded compared with others in the same program and mode), age, education, primary diagnosis, proxy respondent, and self-reported functional status and physical health.

Authors' Contributions

All authors contributed to this article in a substantive way. M.D., R.A.P., and M.N.E. conceived the analysis, with input from M.B., D.S., J.M.T., and P.D.C. C.K.M. and A.T. conducted data analysis. M.D., R.A.P., A.T., M.B., D.S., J.M.T., P.D.C., and M.N.E. participated in drafting and refining the article.

Acknowledgment

The authors gratefully acknowledge the serious illness programs, patients, and family caregivers that participated in the field test of the Serious Illness Survey.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

This work was supported by the Gordon and Betty Moore Foundation.

References

1. National Consensus Project for Quality Palliative Care: *Clinical Practice Guidelines for Quality Palliative Care, 4th ed*. Richmond, VA: National Coalition for Hospice and Palliative Care, 2018.
2. Long P, Abrams M, Milstein A, et al. (eds): *Effective Care for High-Need Patients: Opportunities for Improving Outcomes, Value, and Health*. Washington, D.C.: National Academy of Medicine, 2017.
3. Cohn J, Corrigan J, Lynn J, et al.: *Community-Based Models of Care Delivery for People with Serious Illness. NAM Perspectives, Discussion Paper*. Washington, DC: National Academy of Medicine, 2017. DOI: 10.31478/201704b.
4. National Academies of Sciences, Engineering, and Medicine: *Models and Strategies to Integrate Palliative Care Principles into Care for People with Serious Illness: Proceedings of a Workshop*. Washington, DC: The National Academies Press, 2018. DOI: 10.17226/24908.
5. Lendon JP, Ahluwalia SC, Walling AM, et al.: Measuring experience with end-of-life care: A systematic literature review. J Pain Symptom Manage 2015;49:904–915.e1–e3.
6. National Quality Forum: Issue Brief: Opportunities for Advancing Quality Measurement in Community-Based Serious Illness Care. 2020: https://store.qualityforum.org/collections/advanced-illness-care/products/issue-brief-opportunities-or-advancing-quality-measurement-in-community-based-serious-illness-care (Last accessed January 3, 2022).
7. Anhang Price R, Bradley M, Cleary P, et al.: Reliable and valid survey-based measures to assess quality of care in home-based serious illness programs. J Palliat Med 2022;25:864–872.
8. RAND Corporation: RAND Serious Illness Survey. www.rand.org/Serious-Illness-Survey. (Last accessed November 4, 2021).
9. Zaslavsky AM, Zaborski LB, Ding L, et al.: Adjusting performance measures to ensure equitable plan comparisons. Health Care Financ Rev 2001;22:109–126.
10. Elliott MN, Zaslavsky AM, Goldstein E, et al.: Effects of survey mode, patient mix, and nonresponse on CAHPS® hospital survey scores. Health Serv Res 2009;44(2 Pt 1):501–518.
11. Cefalu M, Elliott MN, Hays RD: Adjustment of patient experience surveys for how people respond. Med Care 2021;59:202–205.
12. Parast L, Haas A, Tolpadi A, et al.: Effects of caregiver and decedent characteristics on CAHPS Hospice Survey scores. J Pain Symptom Manage 2018;56:519–529.e1.
13. Paddison C, Elliott M, Parker R, et al.: Should measures of patient experience in primary care be adjusted for case mix? Evidence from the UK General Practice Patient Survey. BMJ Qual Saf 2012;21:634–640.
14. National Academies of Sciences, Engineering, and Medicine: *Accounting for Social Risk Factors in Medicare Payment: Criteria, Factors, and Methods*. Washington, DC: The National Academies Press, 2016. DOI: 10.17226/23513.
15. Centers for Medicare & Medicaid Services: *Medicare Advantage and Prescription Drug Plan CAHPS® Survey*. Baltimore, MD. 2018 www.MA-PDPCAHPS.org. https://ma-pdpcahps.org/en/scoring-and-star-ratings/
16. Elliott MN, Beckett MK, Chong K, et al.: How do proxy responses and proxy-assisted responses differ from what Medicare beneficiaries might have reported about their health care? Health Serv Res 2008;43:833–848.
17. Fowler FJ, Jr., Gallagher PM, Nederend S: Comparing telephone and mail responses to the CAHPS survey instrument. Consumer Assessment of Health Plans Study. Med Care 1999;37(3 Suppl):MS41–MS49.
18. de Vries H, Elliott MN, Hepner KA, et al.: Equivalence of mail and telephone responses to the CAHPS Hospital Survey. Health Serv Res 2005;40(6 Pt 2):2120–2139.
19. Parast L, Elliott MN, Hambarsoomian K, et al.: Effects of survey mode on Consumer Assessment of Healthcare Providers and Systems (CAHPS) hospice survey scores. J Am Geriatr Soc 2018;66:546–552.
20. Robert Wood Johnson Foundation: How to Report Results of the CAHPS Clinician & Group Survey. 2010: https://www.ahrq.gov/sites/default/files/wysiwyg/cahps/surveys-guidance/cg/cgkit/HowtoReportResultsofCGCAHPS080610FINAL.pdf (Last accessed January 3, 2022).
21. Damiano PC, Elliott M, Tyler MC, et al.: Differential use of the CAHPS® 0–10 global rating scale by Medicaid and commercial populations. Health Serv Outcomes Res Methodol 2004;5:193–205.
22. Parast L, Mathews M, Tolpadi A, et al.: National testing of the Emergency Department Patient Experience of Care Discharged to Community Survey and implications for adjustment in scoring. Med Care 2019;57:42–48.

23. Rubin HR: Can patients evaluate the quality of hospital care? Med Care Rev 1990;47:267–326.
24. Zaslavsky AM, Zaborski LB, Cleary PD: Factors affecting response rates to the Consumer Assessment of Health Plans Study survey. Med Care 2002;40:485–499.
25. Elliott MN, Edwards C, Angeles J, et al.: Patterns of unit and item nonresponse in the CAHPS® hospital survey. Health Serv Res 2005;40(6 Pt 2):2096–2119.
26. Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
27. O'Malley AJ, Zaslavsky AM, Elliott MN, et al.: Case-mix adjustment of the CAHPS® hospital survey. Health Serv Res 2005;40(6 Pt 2):2162–2181.
28. Elliott MN, Haviland AM, Orr N, et al.: How do the experiences of Medicare beneficiary subgroups differ between managed care and original medicare? Health Serv Res 2011;46:1039–1058.
29. Centers for Medicare & Medicaid Services: https://hcahpsonline.org/globalassets/hcahps/mode-patient-mix-adjustment/october_2021_pma_web_document.pdf. (Last accessed August 16, 2021).

Address correspondence to:
Maria DeYoreo, PhD
RAND Corporation
1776 Main Street
Santa Monica, CA 90401
USA

E-mail: mdeyoreo@rand.org

Appendix Table A1. Composite and Global Quality Measures for Serious Illness Programs, with Average Unadjusted Person-Level Top-Box Scores Across All Field Test Respondents

Quality measures	*Response options*	*Top-box*	*Average person-level top-box score*
Communication			
In the last three months, how often did people from this program spend enough time with you when they visited?	Never/Sometimes/Usually/Always	Always	72.8%
In the last three months, how often did people from this program explain things to you in a way you could understand?	Never/Sometimes/Usually/Always	Always	77.3%
In the last three months, how often did people from this program listen carefully to you?	Never/Sometimes/Usually/Always	Always	81.7%
In the last three months, how often did you feel that people from this program cared about you as a whole person?	Never/Sometimes/Usually/Always	Always	80.8%
In the last three months, how often did you feel heard and understood by people from this program?	Never/Sometimes/Usually/Always	Always	72.6%
Care coordination			
In the last three months, how often did people from this program seem to know the important information about your medical history?	Never/Sometimes/Usually/Always	Always	67.7%
In the last three months, did someone from this program talk with you about the care or treatment you get from your other doctors or health care providers?	Yes, definitely/Yes, somewhat/No	Yes, definitely	58.5%
In the last three months, did someone from this program talk with you about all the medicines you are taking?	Yes, definitely/Yes, somewhat/No/I do not take any medicines	Yes, definitely	76.6%
In the last three months, did someone from this program talk with you about how to get help with everyday activities?	Yes, definitely/Yes, somewhat/No/I did not want to talk with this program	Yes, definitely	45.5%
In the last three months, when you contacted this program between visits, did you get the help you needed?[a]	Yes, definitely/Yes, somewhat/No	Yes, definitely	79.9%
Help for symptoms			
In the last three months, did you get as much help as you wanted for your pain?[a]	Yes, definitely/Yes, somewhat/No/I did not want help for my pain	Yes, definitely	55.8%

(*Appendix continues*)

Appendix Table A1. (Continued)

Quality measures	*Response options*	*Top-box*	*Average person-level top-box score*
In the last three months, did you get as much help as you wanted for your breathing?[a]	Yes, definitely/Yes, somewhat/No/I did not want help for my breathing	Yes, definitely	63.7%
In the last three months, did you get as much help as you wanted for your feelings of anxiety or sadness?[a]	Yes, definitely/Yes, somewhat/No/I did not want help for my anxiety or sadness	Yes, definitely	47.8%
Planning for care			
Did someone from this program ever talk with you about what you should do during a health emergency?	Yes, definitely/Yes, somewhat/No	Yes, definitely	61.7%
Did someone from this program ever talk with you about what is important in your life?	Yes, definitely/Yes, somewhat/No	Yes, definitely	48.9%
Did someone from this program ever talk with you about what your health care options would be if you got sicker?	Yes, definitely/Yes, somewhat/No	Yes, definitely	45.5%
Support for family and friends			
In the last three months, did the people from the program involve your family members or friends in discussions about your health care as much as you wanted?[a]	Yes, definitely/Yes, somewhat/No	Yes, definitely	70.9%
In the last three months, did your family members or friends get as much emotional support as they wanted from this program?[a]	Yes, definitely/Yes, somewhat/No/My family members or friends did not want emotional support from this program	Yes, definitely	64.8%
Overall rating of the program			
Using any number from 0 to 10, where 0 is the worst care possible and 10 is the best care possible, what number would you use to rate your care from this program?	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	9 and 10	71.2%
Willingness to recommend the program			
Would you recommend this program to your friends and family?	Definitely Yes/Probably Yes/Probably No/ Definitely No	Definitely Yes	70.6%

Scores reflect average unadjusted person-level top box scores.
[a]A screening question(s) determines whether this evaluative survey question is applicable to the respondent.

Appendix Table A2 Effects of Case-Mix Adjustors and Survey Mode on Evaluative Measures

	Communication	*Care coordination*	*Help for symptoms*	*Planning for care*	*Support for family and friends*	*Overall rating*	*Willingness to recommend*
Survey mode							
Mail-mail (ref)							
Mixed (Mail-telephone)	0.01 (0.02), $p=0.58$	0.02 (0.01), $p=0.10$	0.04 (0.02), $p=0.06$	0.03 (0.02), $p=0.14$	0.02 (0.02), $p=0.30$	0.00 (0.02), $p=0.83$	0.00 (0.02), $p=0.84$
Education	$p=1.00$	$p=0.08$	$p=0.30$	$p=0.15$	$p=0.06$	$p<0.01$	$p=0.78$
Eighth grade or less	0.01 (0.03), $p=0.73$	0.02 (0.03), $p=0.42$	−0.04 (0.04), $p=0.38$	0.04 (0.03), $p=0.25$	0.11 (0.04), $p<0.01$	0.01 (0.04), $p=0.84$	0.03 (0.04), $p=0.43$
Some high school, but did not graduate	0.00 (0.03), $p=0.92$	0.04 (0.02), $p=0.15$	0.03 (0.04), $p=0.39$	0.05 (0.03), $p=0.10$	0.03 (0.04), $p=0.41$	0.05 (0.03), $p=0.12$	0.02 (0.03), $p=0.53$
High school graduate or GED (ref)							
Some college or two-year degree	0.00 (0.02), $p=0.90$	0.00 (0.02), $p=0.81$	−0.02 (0.03), $p=0.49$	−0.03 (0.02), $p=0.23$	−0.02 (0.03), $p=0.38$	−0.06 (0.03), $p=0.02$	0.00 (0.03), $p=0.95$
Four-year college graduate	0.00 (0.03), $p=0.93$	0.03 (0.03), $p=0.24$	0.05 (0.04), $p=0.30$	0.01 (0.03), $p=0.81$	0.03 (0.04), $p=0.39$	0.00 (0.04), $p=0.91$	0.05 (0.04), $p=0.20$
More than four-year college degree	0.00 (0.03), $p=1.00$	−0.05 (0.03), $p=0.05$	−0.06 (0.04), $p=0.17$	−0.03 (0.03), $p=0.42$	0.03 (0.04), $p=0.51$	−0.09 (0.04), $p=0.01$	0.00 (0.04), $p=0.90$
Patient age	$p=0.22$	$p<0.01$	$p=0.33$	$p<0.01$	$p=0.69$	$p=0.69$	$p=0.03$
18–54	0.11 (0.04), $p<0.01$	0.15 (0.04), $p<0.01$	0.12 (0.06), $p=0.06$	0.12 (0.05), $p=0.01$	0.07 (0.06), $p=0.26$	0.09 (0.06), $p=0.10$	0.17 (0.06), $p<0.01$
55–64	0.04 (0.03), $p=0.16$	0.10 (0.03), $p<0.01$	0.05 (0.05), $p=0.32$	0.17 (0.04), $p<0.01$	0.01 (0.05), $p=0.83$	0.01 (0.04), $p=0.82$	0.03 (0.04), $p=0.46$
65–69	0.05 (0.03), $p=0.13$	0.11 (0.03), $p<0.01$	0.03 (0.05), $p=0.50$	0.11 (0.04), $p<0.01$	−0.04 (0.05), $p=0.42$	0.03 (0.04), $p=0.45$	0.06 (0.04), $p=0.19$
70–74	0.05 (0.03), $p=0.10$	0.07 (0.03), $p=0.01$	0.04 (0.04), $p=0.30$	0.11 (0.03), $p<0.01$	0.02 (0.04), $p=0.68$	0.04 (0.04), $p=0.25$	0.10 (0.04), $p<0.01$
75–79	0.04 (0.03), $p=0.11$	0.05 (0.03), $p=0.08$	0.00 (0.04), $p=0.96$	0.04 (0.03), $p=0.16$	−0.03 (0.04), $p=0.46$	0.04 (0.04), $p=0.24$	0.02 (0.04), $p=0.61$
80–84	0.02 (0.02), $p=0.48$	0.06 (0.02), $p<0.01$	0.01 (0.04), $p=0.72$	0.07 (0.03), $p=0.02$	−0.02 (0.03), $p=0.65$	0.04 (0.03), $p=0.24$	0.05 (0.03), $p=0.14$
85–89	0.02 (0.02), $p=0.40$	0.01 (0.02), $p=0.55$	−0.03 (0.04), $p=0.37$	0.01 (0.03), $p=0.80$	−0.03 (0.03), $p=0.41$	0.01 (0.03), $p=0.75$	0.01 (0.03), $p=0.77$
≥90 (ref)							
Diagnosis	$p=0.03$	$p=0.22$	$p<0.01$	$p<0.01$	$p=0.21$	$p=0.12$	$p=0.24$
Cancer	0.05 (0.02), $p=0.04$	0.03 (0.02), $p=0.13$	0.18 (0.03), $p<0.01$	0.09 (0.03), $p<0.01$	0.05 (0.03), $p=0.09$	0.06 (0.03), $p=0.05$	0.04 (0.03), $p=0.15$
Alzheimer's & Other Dementia	0.05 (0.03), $p=0.07$	0.03 (0.03), $p=0.31$	0.12 (0.05), $p<0.01$	0.01 (0.03), $p=0.80$	−0.01 (0.04), $p=0.79$	−0.02 (0.04), $p=0.66$	0.04 (0.04), $p=0.32$
Other (ref)							
Language	$p=0.37$	$p=0.34$	$p=0.19$	$p=0.01$	$p=0.92$	$p=0.27$	$p=0.38$
English (ref)							
Spanish	−0.04 (0.04), $p=0.37$	0.06 (0.04), $p=0.14$	0.10 (0.06), $p=0.10$	0.13 (0.05), $p<0.01$	0.02 (0.06), $p=0.73$	0.08 (0.05), $p=0.14$	0.08 (0.05), $p=0.17$
Other	−0.05 (0.04), $p=0.22$	0.01 (0.04), $p=0.81$	0.07 (0.07), $p=0.31$	−0.03 (0.05), $p=0.60$	−0.01 (0.07), $p=0.89$	−0.03 (0.06), $p=0.66$	0.00 (0.06), $p=0.98$
Proxy	$p<0.01$	$p<0.01$	$p=0.02$	$p<0.01$	$p<0.01$	$p=0.46$	$p=0.77$
Proxy answered questions	0.05 (0.02), $p<0.01$	0.05 (0.02), $p<0.01$	0.09 (0.03), $p<0.01$	0.09 (0.02), $p<0.01$	0.21 (0.03), $p<0.01$	0.03 (0.03), $p=0.22$	0.00 (0.03), $p=0.90$
Proxy helped in some other way	0.08 (0.02), $p<0.01$	0.09 (0.02), $p<0.01$	0.04 (0.04), $p=0.25$	0.09 (0.03), $p<0.01$	0.20 (0.03), $p<0.01$	0.01 (0.03), $p=0.71$	0.02 (0.03), $p=0.57$
No proxy (ref)							
Functional status	$p=0.02$	$p<0.01$	$p=0.41$	$p=0.12$	$p=0.70$	$p=0.08$	$p=0.57$
Not able to get out of bed	−0.08 (0.03), $p=0.01$	−0.09 (0.03), $p<0.01$	0.00 (0.05), $p=0.94$	−0.07 (0.04), $p=0.05$	−0.03 (0.04), $p=0.46$	−0.09 (0.04), $p=0.03$	−0.02 (0.04), $p=0.56$
Can get out of bed, but not leave house	−0.04 (0.03), $p=0.16$	−0.06 (0.03), $p=0.03$	−0.06 (0.05), $p=0.19$	−0.03 (0.03), $p=0.46$	−0.02 (0.04), $p=0.61$	−0.03 (0.04), $p=0.40$	−0.04 (0.04), $p=0.34$
Able to leave house (ref)							
Self-rated physical health	$p=0.79$	$p=0.76$	$p=0.88$	$p=0.69$	$p=0.88$	$p=0.30$	$p=0.36$
Excellent	−0.05 (0.05), $p=0.35$	0.00 (0.05), $p=0.99$	−0.03 (0.09), $p=0.78$	−0.03 (0.06), $p=0.67$	−0.07 (0.08), $p=0.38$	−0.07 (0.07), $p=0.33$	−0.13 (0.07), $p=0.06$
Very good	0.01 (0.03), $p=0.67$	0.04 (0.03), $p=0.19$	0.01 (0.05), $p=0.85$	−0.02 (0.03), $p=0.48$	0.00 (0.04), $p=0.94$	0.05 (0.04), $p=0.19$	0.00 (0.04), $p=1.00$
Good (ref)							
Fair	0.01 (0.02), $p=0.71$	0.01 (0.02), $p=0.60$	−0.03 (0.03), $p=0.37$	0.00 (0.02), $p=0.91$	−0.02 (0.03), $p=0.55$	−0.01 (0.03), $p=0.83$	0.01 (0.03), $p=0.77$
Poor	−0.01 (0.02), $p=0.75$	0.00 (0.02), $p=0.88$	0.00 (0.04), $p=0.89$	0.03 (0.03), $p=0.30$	−0.02 (0.03), $p=0.61$	−0.03 (0.03), $p=0.30$	0.02 (0.03), $p=0.48$
Self-rated mental health	$p<0.01$	$p<0.01$	$p<0.01$	$p<0.01$	$p=0.06$	$p<0.01$	$p=0.04$
Excellent	0.05 (0.03), $p=0.04$	0.09 (0.03), $p<0.01$	0.11 (0.04), $p=0.01$	0.10 (0.03), $p<0.01$	0.09 (0.04), $p=0.02$	0.09 (0.04), $p=0.01$	0.08 (0.04), $p=0.04$
Very good	0.00 (0.02), $p=0.95$	0.00 (0.02), $p=0.91$	0.03 (0.03), $p=0.38$	0.01 (0.02), $p=0.65$	0.01 (0.03), $p=0.68$	0.00 (0.03), $p=1.00$	0.00 (0.03), $p=0.87$
Good (ref)							
Fair	−0.06 (0.02), $p<0.01$	−0.08 (0.02), $p<0.01$	−0.09 (0.03), $p<0.01$	−0.06 (0.02), $p=0.01$	−0.03 (0.03), $p=0.23$	−0.08 (0.03), $p<0.01$	−0.05 (0.03), $p=0.08$
Poor	−0.05 (0.03), $p=0.09$	−0.07 (0.03), $p=0.03$	−0.13 (0.05), $p<0.01$	−0.05 (0.04), $p=0.17$	−0.03 (0.04), $p=0.42$	−0.03 (0.04), $p=0.45$	−0.02 (0.04), $p=0.59$
Response percentile[a]	−0.21 (0.06), $p<0.01$	−0.15 (0.06), $p<0.01$	−0.01 (0.09), $p=0.88$	0.03 (0.07), $p=0.64$	−0.17 (0.08), $p=0.04$	−0.11 (0.08), $p=0.15$	0.00 (0.08), $p=1.00$
Intercept	0.85 (0.06), $p<0.01$	0.67 (0.05), $p<0.01$	0.66 (0.08), $p<0.01$	0.53 (0.06), $p<0.01$	0.70 (0.08), $p<0.01$	0.93 (0.07), $p<0.01$	0.75 (0.08), $p<0.01$

Quantities shown represent regression coefficient point estimates, standard errors, and *p*-values.

Regression coefficient point estimates and standard errors from linear regression models where the outcome was the measure score (e.g., Communication) and the predictors included all characteristics in the table (case-mix adjustors and mode of survey administration) and program fixed effects. *p*-Values from joint tests of significance (*F*-tests) for the adjustor are provided as well as *p*-values that compare each level of the adjustor to the reference level.

We imputed missing patient-level characteristics with the program mean or if still missing, the overall mean across all programs.

[a]Response percentile is defined as the rank-ordered response time for each respondent relative to all eligible patients within program and mode, scaled from 0 to 1.

GED, graduate equivalency degree.