Quality Payment
PROGRAM

# Respiratory Infection Hospitalization

## Measure Testing Form

February 2023

# Contents

# 1.0 Introduction

This Measure Testing Form (MTF) provides a brief summary of the preliminary measure testing results as part of the comprehensive reevaluation process for three episode-based cost measures. Readers may review these results, alongside other documentation, to provide feedback on the measure using the comprehensive reevaluation survey. The testing results reflect both the version of the measure that is currently in-use in MIPS and a revised version of the measure that is undergoing updates potential use in MIPS in future years. Please see the Draft Cost Measure Methodology for a description of the measure specifications and the Draft Measure Codes List for the list of codes used to specify the measure.[1]

## 1.1 Project Title and Overview

The Centers for Medicare & Medicaid Services (CMS) has contracted with Acumen, LLC to develop and maintain episode-based cost measures for potential use in the Merit-Based Incentive Payment System (MIPS) to meet the requirements of the Medicare Access and CHIP Reauthorization Act (MACRA) of 2015. The contract name is "Physician Cost Measures and Patient Relationship Codes (PCMP)." The contract number is 75FCMC18D0015, Task Order 75FCMC19F0004.

## 1.2 Measure Name

Respiratory Infection Hospitalization Episode-based Cost Measure

## 1.3 Type of Measure

Cost/Resource Use

## 1.4 Data

The study period is January 1, 2021 through December 31, 2021. All episodes ending during the study period that meet inclusion and exclusion criteria are included in testing. The measure is calculated with Medicare Parts A and B, administrative claims data, Long-Term Minimum Data Set, Medicare Enrollment Database. For testing purpose, other data sources are used, including the American Community Survey, Common Medicare Environment.

Testing results are presented at a testing volume threshold of 20 episodes for clinician groups and individual practitioners. Clinician groups are identified by a Tax Identification Number (TIN). Individual clinicians are identified using a combination of a Tax Identification Number and National Provider Identifier (TIN-NPI).

[1]These documents will be available on the MACRA Feedback Page once field testing begins. https://www.cms.gov/Medicare/Quality-Payment-Program/Quality-Payment-Program/Give-Feedback

# 2.0 Preliminary Testing Results

This section presents preliminary testing results based on the revised measure as specified for the public comment period. Section 2.1 provides an overview of changes in the measure coverage, clinician population, and reliability between the current measure and the revised version of the measure. Sections 2.2 and 2.3 show additional evidence of scientific acceptability of the measure. Section 2.4 presents empirical results of the risk adjustment and stratification methods used by this measure. Section 2.5 examines the impact of adding social risk factors to the measure's risk adjustment model. Lastly, Section 2.6 examines the impact of exclusion criteria used by the measure through their frequency and resource use patterns.

## 2.1 Impacts of Revisions to the Measure

### 2.1.1 Measure Coverage

Table 1 shows the number of beneficiaries covered by this measure. Table 2 shows the characteristics of TINs and TIN-NPIs who are attributed at least 20 episodes. Compared to the current MIPS version, the revised version captures more beneficiaries.

**Table 1: Measure Coverage**

| Metric | Value | |
|---|---|---|
| | Current MIPS Measure | Revised Measure |
| Number of Beneficiaries | 65,068 | 355,262 |
| Mean Age | 77.08 | 75.43 |
| Female % | 54.04% | 52.78% |

**Table 2: Clinician Characteristics**

| Metric | TIN | | | | TIN-NPI | | | |
|---|---|---|---|---|---|---|---|---|
| | Current MIPS Measure | | Revised Measure | | Current MIPS Measure | | Revised Measure | |
| | Count | % | Count | % | Count | % | Count | % |
| Count | 967 | 100% | 3,401 | 100% | 50 | 100% | 14,190 | 100% |
| **Number of Episodes Attributed** | - | - | - | - | - | - | - | - |
| 20-39 Episodes | 543 | 56.15% | 1,195 | 35.14%% | 48 | 96.00% | 11,930 | 84.07% |
| 40-59 Episodes | 203 | 20.99% | 515 | 15.14% | 1 | 2.00% | 1,732 | 12.21% |
| 60-79 Episodes | 87 | 9.00% | 332 | 9.76% | 1 | 2.00% | 366 | 2.58% |
| 80-99 Episodes | 46 | 0.048% | 249 | 7.32% | 0 | 0.00% | 92 | 0.65% |
| 100-199 Episodes | 71 | 0.073% | 602 | 17.70% | 0 | 0.00% | 70 | 0.49% |

| Metric | TIN | | | | TIN-NPI | | | |
|---|---|---|---|---|---|---|---|---|
| | Current MIPS Measure | | Revised Measure | | Current MIPS Measure | | Revised Measure | |
| | Count | % | Count | % | Count | % | Count | % |
| 200-299 Episodes | 15 | 0.016% | 230 | 6.76% | 0 | 0.00% | 0 | 0.00% |
| 300+ Episodes | 2 | 0.002% | 278 | 8.17% | 0 | 0.00% | 0 | 0.00% |
| **Census Region** | - | - | - | - | - | - | - | - |
| Northeast | 164 | 16.96% | 544 | 16.00% | 4 | 8.00% | 2,842 | 20.03% |
| Midwest | 245 | 25.34% | 785 | 23.08% | 7 | 14.00% | 2,994 | 21.10% |
| South | 428 | 44.26% | 1,510 | 44.40% | 39 | 78.00% | 6,762 | 47.65% |
| West | 129 | 13.34% | 557 | 16.38% | 0 | 0.00% | 1,590 | 11.21% |
| Unknown | 1 | 0.10% | 5 | 0.15% | 0 | 0.00% | 2 | 0.01% |

### 2.1.2 Frequently Attributed Specialties

Table 3 shows the top 10 attributed specialties for this measure, using a 20-episode testing volume threshold. The most frequently attributed specialties reflect the intent of the measure to capture costs of the management of respiratory infections, including internists, pulmonary disease specialists and hospitalists.

For the revised version of the measure, internists and hospitalists together make up over half of all clinicians who meet the testing volume threshold (43.36% and 25.36%, respectively). Pulmonary disease specialists are the third most frequently attributed specialty, comprising 8.39% of all attributed clinicians.

**Table 3: Count of the Top 10 Attributed Specialties**

| Current MIPS Version | | Revised Version | |
|---|---|---|---|
| Specialty | Number of TIN-NPIs Attributed | Specialty | Number of TIN-NPIs Attributed |
| Internal Medicine | 20 | Internal Medicine | 6,153 |
| Pulmonary Disease | 10 | Hospitalist | 3,598 |
| Family Practice | 7 | Pulmonary Disease | 1,190 |
| Hospitalist | 5 | Family Practice | 1,166 |
| Infectious Disease | 2 | Infectious Disease | 752 |
| Critical Care (Intensivists) | 2 | Nurse Practitioner | 617 |
| Cardiology | 1 | Physician Assistant | 249 |
| Nephrology | 1 | Critical Care (Intensivists) | 131 |
| Emergency Medicine | 1 | Emergency Medicine | 95 |
| Physician Assistant | 1 | Nephrology | 84 |

### 2.1.3 Reliability

Reliability evaluates a measure's ability to consistently differentiate the performance of one clinician from another. The signal-to-noise ratio is used to estimate reliability, which indicates how much of the variation in the measure score is explained by differences among clinicians' performance (i.e., signal) instead of differences within each clinician's performance (i.e., noise). Specifically, noise is the variation from one episode to another during the performance period for a particular clinician.

Table 4 shows reliability metrics at various testing volume thresholds. While higher thresholds yield higher reliability results, it is at the cost of further reducing the number of clinicians and clinician groups eligible for the measure, which would reduce the potential impact of the measure. For the purposes of testing, we used a 20-episode volume threshold (bolded in the table below) to align with the current MIPS version. If the measure is implemented in the MIPS in the future, CMS will establish a case minimum through notice-and-comment rulemaking.

**Table 4: Sample Size, Mean Reliability, and Proportion of Clinicians above Moderate Reliability at Various Testing Volume Thresholds**

| Version | Testing Volume Threshold | TIN | | | TIN-NPI | | |
|---|---|---|---|---|---|---|---|
| | | Number of TINs | Mean Reliability | Percent Above 0.4 | Number TIN-NPIs | Mean Reliability | Percent Above 0.4 |
| Current Measure | 10 | 1,854 | 0.43 | 49.73% | 992 | 0.17 | 0.20% |
| Revised Measure | 10 | 5,144 | 0.63 | 81.90% | 36,275 | 0.53 | 90.62% |
| **Current Measure** | **20** | **967** | **0.56** | **95.35%** | **50** | **0.30** | **4.00%** |
| **Revised Measure** | **20** | **3,401** | **0.74** | **100.00%** | **14,190** | **0.64** | **100.00%** |
| Current Measure | 30 | 615 | 0.63 | 100.00% | 11 | 0.38 | 18.18% |
| Revised Measure | 30 | 2,674 | 0.80 | 100.00% | 5,476 | 0.71 | 100.00% |

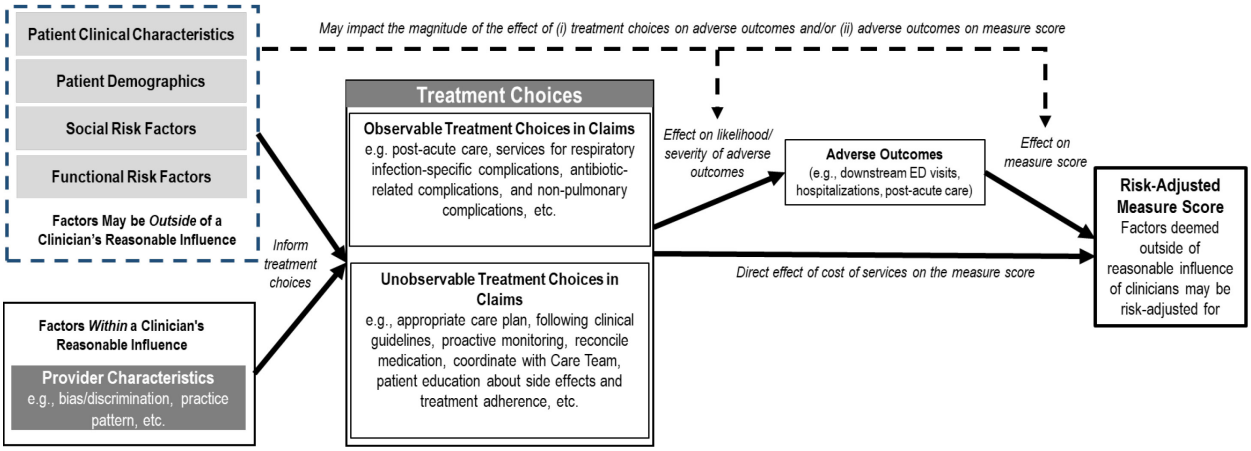
At the testing volume of 20 episodes, the revised version shows moderate to high reliability, specifically 0.74 at the TIN level and 0.64 at the TIN-NPI level (Table 4). CMS generally considers 0.4 as the threshold indicating 'moderate' reliability and 0.7 indicating 'high' reliability, which is supported by previous work into reliability and the threshold was finalized in the 2022 Physician Fee Schedule final rule.[2,3] All TINs and TIN-NPIs meet or exceed the moderate reliability threshold of 0.4 at the 20-episode testing volume threshold. Compared to current measure, the reliability of the revised measure is markedly higher and many more clinicians would be eligible for the revised measure.

[2]Mathematica, Inc., "Memorandum: Reporting Period and Reliability of AHRQ, CMS 30-Day and HAC Quality Measures – Revised," http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/HVBP_Measure_Reliability-.pdf.
[3]CMS, "Medicare Program; CY 2022 Payment Policies Under the Physician Fee Schedule and Other Changes to Part B Payment Policies; Medicare Shared Savings Program Requirements; Provider Enrollment Regulation Updates; and Provider and Supplier Prepayment and Post-Payment Medical Review Requirements," 86 FR 64996-66031.

## 2.2 Validity

Validity is a criterion that evaluates whether the cost measure is able to quantify the construct that it aims to measure, which is the cost directly related to treatment choices and cost of adverse outcomes as a result of care. Validity is evaluated empirically by estimating the effect of relevant treatment choices on the measure score using multiple regression, based on the conceptual model outlined in Figure 1.

**Figure 1: Conceptual Model of the Relationship between Treatment Choices and the Measure Score**




The cost measure is designed to reflect the cost directly related to treatment choices, as well as the cost of adverse outcomes as a result of care. Therefore, treatment choices, either observable in claims or otherwise, by an attributed clinician can directly impact the measure score or indirectly when they're mediated through the cost of adverse outcomes. The cost of adverse outcomes, in turn, contributes to the total costs that are captured by the measure score.

To demonstrate that the measure score is reflective of both the direct and indirect effects of treatment choices, this analysis first estimates the association between treatment choices and the measure score while controlling for the cost of adverse outcomes. Then, the association between treatment choices and the cost of adverse outcomes is estimated to demonstrate the indirect effect.

Generally, adverse outcomes are non-trigger inpatient hospitalizations, non-trigger emergency room visits, and post-acute care. The remaining service categories are generally considered treatment. For each of these categories, the regression models use the mean cost across episodes that were attributed to an individual clinician. The measure score is represented by a clinician's mean observed cost over expected cost ratio across their attributed episodes.

Overall, the results demonstrate that the cost measure is reflective of both the cost directly related to treatment choices, as well as cost of adverse outcomes as a result of care (Table 5). Therefore, there's evidence that the measure is capturing what it purports to measure.

Model 1 shows that the cost of adverse events is associated with a worse measure score, which includes hospitalizations or emergency department visits that are clinically related to respiratory infection. Model 2 shows that costs during the trigger inpatient stay, outpatient physical, occupational, or speech therapies (TIN-NPI reporting level only), and anesthesia (TIN-NPI reporting level only), and dialysis are associated with lower cost of adverse events. However, except for dialysis, these services are associated with a worse score after controlling for the

cost of adverse events, which suggests that they are also prone to over-use. Other services from model 1 all show an association with a worse score.

**Table 5: Estimated Effect of Treatment Choices (Revised Measure)**

| Categories of Service | Coefficient in Thousands [95% Confidence Interval] (p-value) | | | |
|---|---|---|---|---|
| | TIN | | TIN-NPI | |
| | Model 1: Mean O/E = Mean Cost of Treatment Choices + Mean Cost of Adverse Events | Model 2: Mean Cost of Adverse Events = Mean Cost of Treatment Choices | Model 1: Mean O/E = Mean Cost of Treatment Choices + Mean Cost of Adverse Events | Model 2: Mean Cost of Adverse Events = Mean Cost of Treatment Choices |
| Adverse Events | 0.05 [0.05,0.05] (p < 0.01) | - | 0.05 [0.05,0.05] (p < 0.01) | - |
| Outpatient Evaluation & Management Services | 0.09 [0.06,0.12] (p < 0.01) | 9.93 [9.32,10.53] (p < 0.01) | 0.07 [0.05,0.08] (p < 0.01) | 9.79 [9.52,10.06] (p < 0.01) |
| Outpatient Physical, Occupational, or Speech and Language Pathology Therapy | 1.05 [0.41,1.70] (p < 0.01) | -2.43 [-17.43,12.56] (p = 0.75) | 0.49 [0.23,0.75] (p < 0.01) | -8.18 [-14.08,-2.28] (p < 0.01) |
| Imaging Services | 0.38 [0.18,0.59] (p < 0.01) | 4.05 [-0.72,8.81] (p = 0.10) | 0.35 [0.25,0.45] (p < 0.01) | 7.24 [4.93,9.55] (p < 0.01) |
| Inpatient Hospital Trigger | 0.04 [0.04,0.04] (p < 0.01) | -0.11 [-0.14,-0.08] (p < 0.01) | 0.04 [0.04,0.05] (p < 0.01) | -0.05 [-0.06,-0.04] (p < 0.01) |
| Physician Services During Hospitalization Trigger | 0.04 [0.04,0.05] (p < 0.01) | -0.05 [-0.14,0.04] (p = 0.28) | 0.05 [0.05,0.06] (p < 0.01) | -0.16 [-0.20,-0.12] (p < 0.01) |
| Anesthesia Services | 1.03 [0.78,1.28] (p < 0.01) | -2.63 [-8.40,3.15] (p = 0.37) | 0.73 [0.62,0.84] (p < 0.01) | 6.08 [3.60,8.56] (p < 0.01) |
| Dialysis | -0.12 [-0.21,-0.03] (p = 0.01) | -2.33 [-4.47,-0.19] (p = 0.03) | -0.07 [-0.13,-0.01] (p = 0.02) | 2.46 [1.08,3.84] (p < 0.01) |

## 2.3 Performance Gap

Table 6 shows the distribution of the revised measure scores for clinicians and clinician groups. These results align with expectations based on our review of the literature and demonstrate that there is a performance gap in cost measure performance at both the clinician and clinician group levels. The Respiratory Infections Hospitalizations cost measure score at the 90th percentile is much higher than the measure score at the 10th percentile at both the TIN and TIN-NPI levels. The variation in the measure score, indicated by the interquartile range and standard deviation, is in the thousands of dollars. The results suggest that there is opportunity for improvement in performance across providers.

**Table 6: Distribution of the Measure Score (Revised Measure)**

| Metric | TIN | TIN-NPI |
|---|---|---|
| Mean Score | $15,565 | $17,865 |
| Score Interquartile Range (IQR) | $1,677 | $2,612 |
| Standard Deviation | $1,428 | $2,016 |
| Coefficient of Variation | 0.09 | 0.11 |
| **Score Percentile** | | |
| 10th | $13,925 | $15,455 |
| 25th | $14,638 | $16,444 |
| 50th | $15,425 | $17,674 |
| 75th | $16,315 | $19,056 |
| 90th | $17,340 | $20,508 |

## 2.4 Risk Adjustment and Stratification

Figure 1 shows the conceptual model that outlines how patient-level and clinician-level factors can influence the measure score, which is informed by both published external research and our own data analysis.[4,5,6,7,8] The conceptual model includes risk factors that are either known by the literature or informed by the initial and reconvened Clinical Expert Workgroups to be within or outside of the influence of the attributed clinician. Risk factors, including social risk factors (SRFs), can both influence the treatment choices and impact the size of the effect of treatment choices by mitigating the risk of adverse outcomes and the cost of adverse outcomes.

A systematic approach then guides the decision of which factors to include in the risk adjustment model. First, during initial development of the current MIPS measure, we reviewed the literature to gather known risk factors and drivers of resource use. These factors are usually

[4]Centers for Medicare & Medicaid (CMS), Office of Minority Health. "Utilization of Z Codes for Social Determinants of Health among Medicare Fee-for-Service Beneficiaries." (2019) https://www.cms.gov/files/document/z-codes-data-highlight.pdf

[5]Assistant Secretary of Health and Human Services for Planning and Evaluation. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. Washington, D.C. December 2016.

[6]Chen LM, Epstein AM, Orav EJ, Filice CE, Samson LW, Joynt Maddox KE. Association of Practice-Level Social and Medical Risk With Performance in the Medicare Physician Value-Based Payment Modifier Program. JAMA. 2017;318(5):453-461

[7]Medicare Payment Advisory Commission. Beneficiaries Dually Eligible for Medicare and Medicaid. 2018; https://www.macpac.gov/publication/data-book-beneficiaries-dually-eligible-for-medicare-and-medicaid-3/.

[8]Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. https://aspe.hhs.gov/reports/second-report-congress-social-risk-medicares-value-based-purchasing-programs

diagnoses; therefore, the first set of risk adjustors are commonly the Hierarchical Condition Categories. Then, we consulted our clinical expert panels on additional factors that are known to be associated with resource use. Together with our clinical expert panel, we reviewed the stratified results on episode cost across many different patient characteristics. We arrived at the final list of risk adjustors used in the current MIPS measure based on those discussions and consensus among the clinical experts. We also reviewed literature and gathered additional input from the reconvened Clinician Expert Workgroup to determine whether revisions should be made to the risk adjustment model. Additionally, during our testing phases, we also follow a structured and systematic approach to decide whether SRFs should be risk-adjusted for, which is further described in Section 2.5.

### 2.4.1 Discrimination

Discrimination is a statistical criterion that evaluates the measure's ability to distinguish high-cost episodes from low-cost episodes, or the ability to explain the variance in cost of individual episodes. The amount of variance explained is estimated by the R-squared metric with the range between 0 and 1. The R-square value for the measure is 0.134, both before and after adjusting for the model's complexity based on the number of risk adjustors used. In other words, 13.4% of the variation in the actual observed cost of episodes is explained by the risk adjustment model and sub-group stratification.

The remaining unexplained variance is due to variation in factors that are not adjusted for by the measure, such as the clinician's performance. The objective of a cost measure is to evaluate and differentiate the performance of clinicians. Therefore, achieving high explained variance is not essential because not all of the variation in cost of care should be adjusted. In collaboration with the experts from our clinical workgroup, this measure only adjusts for factors that are deemed to be outside of the influence of clinicians. Please see the Draft Cost Measure Methodology for more information on the full list of risk adjustors and sub-groups.

### 2.4.2 Calibration

Calibration evaluates the consistency of the measure in estimating episode cost across the full range of resource use patterns in the population. Calibration is estimated by the average predictive ratios across groups within the population, specifically groups are partitioned by deciles of expected episode cost. The predictive ratio is calculated using the formula of average expected cost / average observed cost for all episodes in each decile. A well-calibrated measure should have predictive ratios close to 1.00 across all deciles. In other words, such results show that the measure is consistent because it does not under- or over-predict cost throughout the range of resource use patterns in the population.

Table 7 shows that the model has consistent predictive ratios across risk score deciles, with each decile having a predictive ratio between 0.99 and 1.01. The average predictive ratio for all risk deciles is 1.00, which demonstrates that the risk adjustment does not under- or over- predict across the full range of resource use patterns in the population.

**Table 7: Predictive Ratio by Decile of Predicted Episode Cost (Revised Measure)**

| Decile | Average Predictive Ratio |
|---|---|
| Decile 1 | 1.01 |
| Decile 2 | 1.01 |
| Decile 3 | 0.99 |

| Decile | Average Predictive Ratio |
|---|---|
| Decile 4 | 1.00 |
| Decile 5 | 1.01 |
| Decile 6 | 1.01 |
| Decile 7 | 1.00 |
| Decile 8 | 1.00 |
| Decile 9 | 0.99 |
| Decile 10 | 1.00 |

## 2.5 Social Risk Factor Analysis

Beyond clinical characteristics of patients, the cost of care may be influenced by non-clinical factors related to a patient's social risk factors (SRFs), such as race, income, education, and employment. At the program level, MIPS adjusts for SRFs using the MIPS Complex Patient Bonus to ensure clinicians or groups treating more complex patients are not disadvantaged.[9] At the measure-level, the testing helps to navigate the tension between ensuring fairness for clinicians treating higher shares of vulnerable patients and the possibility of masking poor performance and perpetuating disparity if clinicians are held to different standards.

Table 8 outlines variables that may indicate SRFs and their advantages and disadvantages as indicators of individual-level SRFs. Based on availability of data, this analysis tested all variables except for the ICD-10 Z codes.

**Table 8: Social Risk Factors Available for Analysis (Revised Measure)**

| Variable | Advantages | Disadvantages | Used in Testing |
|---|---|---|---|
| Dual Medicare and Medicaid enrollment status | • Available for all beneficiaries<br>• Most powerful predictor of poor outcomes[10] | • Variation in Medicaid eligibility across states | Yes |
| Race/Ethnicity | • Available for most beneficiaries, except for ambiguous categories of "Unknown" or "Other" | • Social risk driven by someone's race is often correlated with and partially captured by dual status[11]<br>• Only 5 categories available, which may lack granularity to fully capture disparities[12,13] | Yes |

[9] https://qpp-cm-prod-content.s3.amazonaws.com/uploads/966/QPP%20COVID-19%20Response%20Fact%20Sheet.pdf
[10] Refer to footnote 4.
[11] Refer to footnote 4.
[12] Nguyen, Kevin H., Kaitlyn P. Lew, and Amal N. Trivedi. "Trends in Collection of Disaggregated Asian American, Native Hawaiian, and Pacific Islander Data: Opportunities in Federal Health Surveys." *American Journal of Public Health* (2022).
[13] Kader, Farah, Lan N. Doan, Matthew Lee, Matthew K. Chin, Simona C. Kwon, and Stella S. Yi. "Disaggregating Race/Ethnicity Data Categories: Criticisms, Dangers, And Opposing Viewpoints", *Health Affairs Forefront* (2022).

| Variable | Advantages | Disadvantages | Used in Testing |
|---|---|---|---|
| ICD-10 Z codes for social determinants of health | • Reflects individual-level factors that influence health status and contact with health services | • Not routinely and consistently coded on claims, only available for 0.1% of all fee-for-service claims in 2019[14] | No |
| American Community Survey | • Can link beneficiary's ZIP code to socioeconomic (SES) measurement of their neighborhood<br>• Many SES indices can be derived from the survey data (e.g., Agency for Healthcare Research and Quality SES index, deprivation index) | • Only a proxy measure, not always accurate at individual-level | Yes |

First, this analysis evaluated each of the variables for their association with episode cost using step-wise regression. Table 9 shows that dual Medicare and Medicaid enrollment status is a powerful predictor, even in the presence of other variables. This is also consistent with other research that found dual status to be the best proxy of SRFs in predicting health outcomes.[15]

**Table 9: Associations of Available Social Risk Factor Variables and Cost of Care**

| Level | Subgroup Risk Model | Variable | Coefficient (p-value) | | |
|---|---|---|---|---|---|
| | | | Model 1: Base Model + Dual Status | Model 2: Base Model + Dual Status + Race | Model 3: Base Model + Dual Status + Race + AHRQ SES |
| TIN | Respiratory Infections and Inflammations | Dual Status | $972.67 (p: <0.001) | $937.35 (p: <0.001) | $1023.76 (p: <0.001) |
| | | Race – Asian | - | $360.5 (p: <0.001) | $299.94 (p: <0.001) |
| | | Race – Black | - | $280.02 (p: <0.001) | $367.2(p: <0.001) |
| | | Race – Hispanic | - | -$143.11 (p: 0.08) | -$43.72 (p: 0.59) |
| | | Race – North American Native | - | -$218.52 (p: 0.19) | -$39.46 (SD: 166.25, p: 0.81) |
| | | Race – Others | - | $231.56 (p: <0.001) | $192.15 (p: 0.02) |

[14]Centers for Medicare & Medicaid (CMS), Office of Minority Health. "Utilization of Z Codes for Social Determinants of Health among Medicare Fee-for-Service Beneficiaries." (2019) https://www.cms.gov/files/document/z-codes-data-highlight.pdf

[15]Office of the Assistant Secretary for Planning and Evaluation. "Second report to Congress on social risk and Medicare's value-based purchasing programs." (2020) https://aspe.hhs.gov/pdf-report/second-impact-report-to-congress

| Level | Subgroup Risk Model | Variable | Coefficient (p-value) | | |
|---|---|---|---|---|---|
| | | | Model 1: Base Model + Dual Status | Model 2: Base Model + Dual Status + Race | Model 3: Base Model + Dual Status + Race + AHRQ SES |
| TIN | Respiratory Infections and Inflammations | Race – White | - | ref | ref |
| | | AHRQ SES Index | - | - | $40.14 (p: <0.001) |
| | Simple Pneumonia and Pleurisy | Dual Status | $558.09 (p: <0.001) | $593.64 (p: <0.001) | $633.38 (p: <0.001) |
| | | Race – Asian | - | $62.82 (p: 0.66) | -$43.34 (p: 0.76) |
| | | Race – Black | - | -$96.56 (p: 0.13) | -$59.25 (p: 0.36) |
| | | Race – Hispanic | - | -$547.15 (p: <0.001) | -$514.95 (p: <0.001) |
| | | Race – North American Native | - | $125.30 (p: 0.58) | $182.20 (p: 0.42) |
| | | Race – Others | - | $4.60 (p: 0.97) | -$12.15 (p: 0.92) |
| | | Race – White | - | ref | ref |
| | | AHRQ SES Index | - | - | $14.29 (p: <0.001) |
| TIN_NPI | Respiratory Infections and Inflammations | Dual Status | $836.33 (p: <0.001) | $800.79 (p: <0.001) | $914.49 (p: <0.001) |
| | | Race – Asian | - | $340.44 (p: <0.001) | $261.76 (p: <0.001) |
| | | Race – Black | - | $296.83 (p: <0.001) | $416.31 (p: <0.001) |
| | | Race – Hispanic | - | -$225.77 (p: <0.001) | -$92.21 (p: 0.16) |
| | | Race – North American Native | - | $109.33 (p: 0.37) | $357.11 (p: <0.001) |
| | | Race – Others | - | $375.80 (p: <0.001) | $323.76 (p: <0.001) |
| | | Race – White | - | ref | ref |
| | | AHRQ SES Index | - | - | $55.08 (p: <0.001) |
| | Simple Pneumonia and Pleurisy | Dual Status | $477.31 (p: <0.001) | $504.88 (p: <0.001) | $560.82 (p: <0.001) |
| | | Race – Asian | - | -$192.55 (p: 0.09) | -$314.5 (p: 0.01) |
| | | Race – Black | - | $42.80 (p: 0.39) | $93.16 (p: 0.06) |
| | | Race – Hispanic | - | -$605.27 (p: <0.001) | -$559.79 (p: <0.001) |
| | | Race – North American Native | - | $1349.52 (p: <0.001) | $1441.29 (p: <0.001) |
| | | Race – Others | - | -$60.87 (p: 0.5) | -$96.30 (p: 0.29) |

| Level | Subgroup Risk Model | Variable | Coefficient (p-value) | | |
|---|---|---|---|---|---|
| | | | Model 1: Base Model + Dual Status | Model 2: Base Model + Dual Status + Race | Model 3: Base Model + Dual Status + Race + AHRQ SES |
| TIN-NPI | Simple Pneumonia and Pleurisy | Race – White | - | ref | ref |
| | | AHRQ SES Index | - | - | $22.10 (p: <0.001) |

The subsequent analyses focus on dual status as the main proxy variable for SRFs for risk adjustment. To determine whether it's appropriate to risk adjust for SRFs, the following criteria are considered:

(i) whether there's an association between social risk and performance by examining the coefficient of patient-level dual status when added into the risk model,

(ii) whether the observed association is most influenced by patient-level factors or clinician-level factors by examining the stability of the patient-level dual status coefficient after adding clinician's dual share variable, as well as including the clinician's fixed effects,

(iii) whether the patient's need or complexity (rather than poor quality) is driving the observed performance differences by examining the differences in performance on dual patients versus non-dual patients and if there are many clinicians who are able to perform similarly or better on their dual patients than their non-dual patients, and

(iv) the impact of risk adjusting for SRFs by examining the performance shift of clinicians compared to a risk adjustment model that doesn't risk adjust for SRFs.

There's a statistically significant association between the patient's dual status and episode cost (Table 10). This association decreases after adding variables to account for provider-level factors but remains statistically significant, which suggests that provider-level factor can partially influence episode cost. There is a performance degradation observed with increasing share of dual episodes, and the degradation is present on both dual and non-dual episodes, which suggests that the degradation is not entirely driven by patient-level factors (Table 11). Many providers are still able to perform equally well or significantly better on the dual episodes than their non-dual episodes, which suggest that it is possible to mitigate the effects of SRFs (Table 12). Lastly, risk adjusting for dual status appears to change the performance ranking for the clinicians, but the magnitude of change is marginal (Table 13).

**Table 10: Coefficient of Patient-level Dual Status under Different Models (Revised Measure)**

| Level | Subgroup Risk Model | % of All Episodes | Coefficient of Patient-level Dual Status (P-value) | | |
|---|---|---|---|---|---|
| | | | Base Model + Patient-level Dual Status | Base Model + Patient-level Dual Status + Clinician's Dual Share | Base Model + Patient-level Dual Status + Clinician's Fixed Effect |
| TIN | Respiratory Infections and Inflammations | 75.9% | $972.67 (p: <0.001) | $535.94 (p: <0.001) | $473.47 (p: <0.001) |

| Level | Subgroup Risk Model | % of All Episodes | Coefficient of Patient-level Dual Status (P-value) | | |
|---|---|---|---|---|---|
| | | | Base Model + Patient-level Dual Status | Base Model + Patient-level Dual Status + Clinician's Dual Share | Base Model + Patient-level Dual Status + Clinician's Fixed Effect |
| TIN | Simple Pneumonia and Pleurisy | 24.1% | $558.09 (p: <0.001) | $346.31 (p: <0.001) | $377.38 (p: <0.001) |
| TIN-NPI | Respiratory Infections and Inflammations | 77.68% | $836.33 (p: <0.001) | $484.39 (p: <0.001) | $411.80 (p: <0.001) |
| TIN-NPI | Simple Pneumonia and Pleurisy | 22.32% | $477.31 (p: <0.001) | 318.34 (SD: 40.36, p: <0.001) | $380.12 (p: <0.001) |

**Table 11: Mean Ratio of Observed Cost to Expected Cost (O/E) Stratified by Clinician's Dual Share and Patient's Dual Status (Revised Measure)**

| Dual Share | TIN | | | TIN-NPI | | |
|---|---|---|---|---|---|---|
| | All Episodes | Dual Episodes | Non-Dual Episodes | All Episodes | Dual Episodes | Non-Dual Episodes |
| All | 1.02 | 1.03 | 1.01 | 0.98 | 1.01 | 0.98 |
| 0% | 0.99 | - | 0.99 | 0.97 | - | 0.97 |
| 1-20% | 1.01 | 1.04 | 1.00 | 0.98 | 1.01 | 0.97 |
| 21-40% | 1.01 | 1.04 | 1.00 | 0.98 | 1.00 | 0.97 |
| 41-60% | 1.02 | 1.02 | 1.00 | 1.00 | 1.00 | 0.99 |
| 61-80% | 1.06 | 1.06 | 1.07 | 1.03 | 1.03 | 1.03 |
| 81-99% | 1.14 | 1.14 | 1.08 | 1.13 | 1.13 | 1.10 |
| 100% | 1.16 | 1.16 | - | 1.14 | 1.14 | - |

**Table 12: Proportions of Clinicians Who Perform Significantly Worse, Equally Well, or Significantly Better on Their Dual Episodes than Non-Dual Episodes (Revised Measure)**

| Reporting Level | Significantly Worse | Equally Well | Significantly Better |
|---|---|---|---|
| TIN | 6% | 93% | 1% |
| TIN-NPI | 5% | 94% | 1% |

**Table 13: Clinicians' Performance Shift Measured by the Change in the Average Ratio of Observed Cost to Expected Cost (O/E) (Revised Measure)**

| Reporting Level | Proportions of Clinicians Affected at Various Levels of Performance Shift | |
|---|---|---|
| | Ranking Shift by 1% or more | Ranking Shift by 5% or more |
| TIN | 72.9% | 5.3% |
| TIN-NPI | 63.1% | 2.5% |

## 2.6 Impact of Exclusions

Table 14 displays descriptive statistics of all episodes meeting the revised measure's triggering logic, excluded episodes, and final reportable episodes at both TIN and TIN-NPI levels. These exclusion criteria ensure that the reportable episode populations are more homogenous and comparable than all episodes meeting triggering logic. It is worth noting that only the observed cost is shown, which has not been risk adjusted for using our risk adjustment model. Therefore, the differences in cost may appear much smaller after risk adjustment than as-is.

Overall, exclusion criteria decrease the distribution of observed cost of all episodes meeting trigger logic, from the mean of $16,276 to $15,398 at the TIN-level and $15,744 at the TIN-NPI level (Table 14).

Episodes from non-acute hospitals are excluded because the differences in payment systems may impact the episode cost and may have clinical needs that are different the population captured by the measure. These episodes also have higher mean observed cost than all episodes meeting triggering logic, at $18,984. Similarly, episodes that have concurrent inpatient admissions with the episode are excluded due to potential uncertainty of length of stay and attribution.

Episodes where a beneficiary died before the episode end date are excluded because they do not provide sufficient data in the episode window period. These episodes also have a higher mean observed cost than all episodes meeting triggering logic, at $16,761, likely because the costs are distributed over fewer days than a typical episode.

Episodes classified as outlier cases are excluded because they deviate substantially from the projected cost for a given patient risk profile. Outlier episodes have a mean observed episode cost of $36,248. The wide variability of observed episode costs for outlier cases also supports their exclusion. At the 10th percentile the outlier cases observed cost is $4,682 and at the 90$^{th}$ percentile the observed cost is $74,994.

Episodes where there is not an attributed TIN are excluded because no TIN has met the threshold of 30% evaluation and management claims billed during the inpatient stay. As such, they cannot be used in the measure.

Based on the input from the clinical expert workgroup, several comorbidities are excluded because these episodes can be clinically distinct from the overall population. These exclusions comprise a small percentage of all triggered episodes.

**Table 14: Cost Statistics for Measure Exclusions (Revised Measure)**

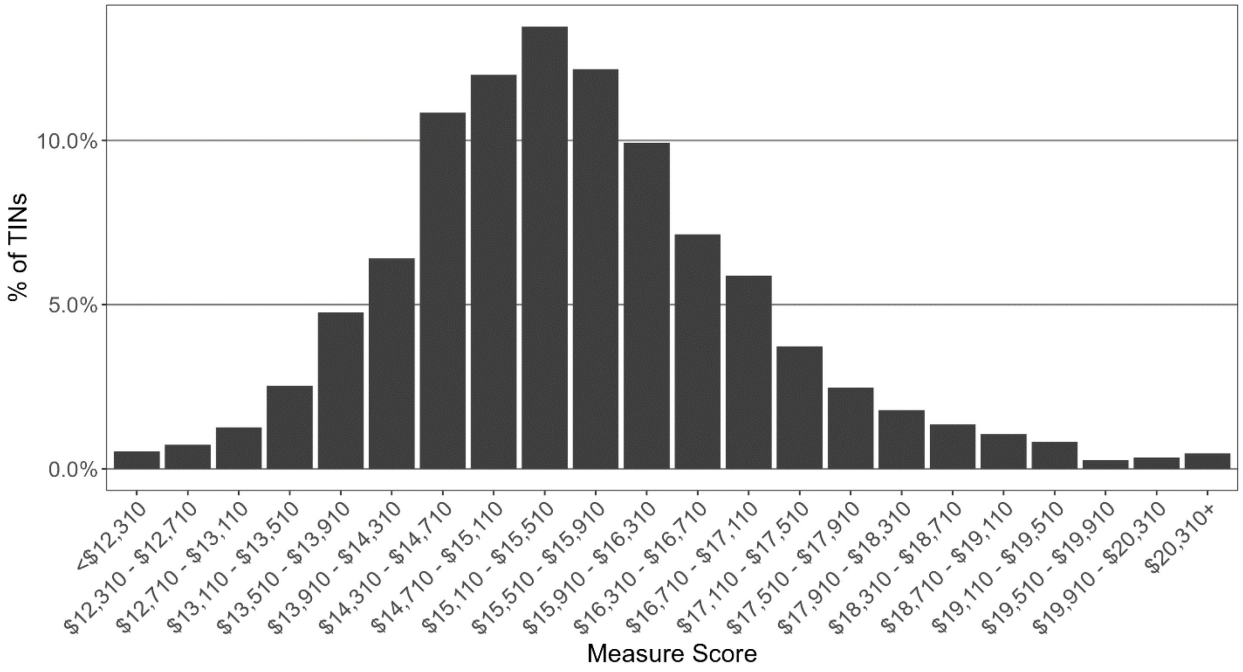
| Exclusion Criteria | Episodes | | Observed Episode Cost | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Percentile | | | | |
| | Count | Percent of All Episodes Meeting Trigger Logic | Mean | 10$^{th}$ | 25$^{th}$ | 50$^{th}$ | 75$^{th}$ | 90$^{th}$ |
| All Episodes Meeting Triggering Logic | 518,332 | 100.00% | $16,276 | $7,948 | $10,917 | $13,220 | $18,661 | $28,293 |
| Not an IPPS Acute Hospital or Psychiatric Facility | 43,920 | 8.47% | $18,984 | $6,819 | $9,837 | $13,234 | $23,784 | $39,049 |

| Exclusion Criteria | Episodes | | Observed Episode Cost | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Percentile | | | | |
| | Count | Percent of All Episodes Meeting Trigger Logic | Mean | 10th | 25th | 50th | 75th | 90th |
| Overlapping IP Admission Days | 1,296 | 0.25% | $16,090 | $5,040 | $5,885 | $13,246 | $19,396 | $30,567 |
| Beneficiary Death in Episode | 87,775 | 16.93% | $16,761 | $9,158 | $12,369 | $13,905 | $18,431 | $26,531 |
| Outlier Cases | 7,634 | 1.47% | $36,248 | $4,862 | $6,922 | $23,595 | $59,068 | $74,994 |
| No Attributed TIN | 12,501 | 2.41% | $21,829 | $10,982 | $13,869 | $16,929 | $27,042 | $37,468 |
| Adverse effects of glucocorticoids and synthetic analogues (T380X5) | 6,939 | 1.34% | $18,540 | $8,653 | $11,018 | $13,951 | $22,758 | $34,253 |
| Pleural Plaque with Presence of Asbestos | 16 | 0.00% | $10,712 | $6,521 | $6,847 | $8,797 | $11,109 | $19,773 |
| Chest Trauma | 4,769 | 0.92% | $16,960 | $7,514 | $10,073 | $13,539 | $21,655 | $30,155 |
| Epidemic Myalgia | * | * | * | * | * | * | * | * |
| Fibrothorax | * | * | * | * | * | * | * | * |
| Influenza due to Avian Flu | 11 | 0.00% | $36,503 | $5,837 | $6,677 | $26,944 | $33,286 | $98,660 |
| Discharged Against Medical Advice | 4,102 | 0.79% | $13,560 | $7,034 | $9,464 | $12,531 | $13,672 | $19,314 |
| Chest Wall Myopathy | 889 | 0.17% | $15,677 | $8,086 | $9,847 | $13,162 | $16,686 | $27,841 |
| Pleural Plaque without Asbestos | * | * | * | * | * | * | * | * |
| Pleurisy | 165 | 0.03% | $8,722 | $4,799 | $5,956 | $6,960 | $9,652 | $12,516 |
| Pleural Condition Unspecified | * | * | * | * | * | * | * | * |
| TIN Does Not Meet Case Minimum | 56,456 | 10.89% | $16,916 | $6,860 | $9,676 | $12,999 | $20,051 | $31,080 |
| TIN-NPI Does Not Meet Case Minimum | 218,110 | 42.08% | $15,901 | $7,222 | $9,939 | $12,944 | $17,785 | $28,016 |
| **Reportable Episodes** (if all clinicians reported as TIN at the Testing Volume Threshold) | 346,391 | 66.83% | $15,398 | $8,107 | $10,739 | $13,053 | $17,624 | $27,196 |
| **Reportable Episodes** (if all clinicians reported as TIN-NPI at the Testing Volume Threshold) | 227,703 | 43.93% | $15,744 | $8,453 | $11,597 | $13,198 | $18,341 | $27,611 |

* This row contains 10 or less episodes. Data is suppressed to protect privacy.

# Appendix A. Distributions of Measure Score (Revised Measure)

### Figure 2: Distribution of Measure Score - TIN




### Figure 3: Distribution of Measure Score - TIN-NPI