2a.01) Select the data sources for which the measure is listed.

Electronic health data, electronic health records, chart review

2a.02) If an existing dataset was used, identify the specific dataset.

For reliability and validity testing, we used data from the Michigan Hospital Medicine Safety Consortium (HMS). HMS is a collaborative quality initiative sponsored by Blue Cross Blue Shield of Michigan (https://mi-hms.org/). HMS includes 62 non-governmental hospitals throughout the state of Michigan. In July 2017, HMS hospitals joined in the "Antimicrobial Use Initiative" to collect patient-level data related to hospitalized, medical patients treated for pneumonia (https://mi-hms.org/quality-initiatives/antimicrobial-use-initiative).[1,2,3,4]

For all analyses included in this measure submission, data from HMS are censored as of March 31, 2020, at which time 49 hospitals had contributed data to the dataset.

The dataset includes chart abstracted data, such as:

- Patient demographics (e.g., age, admission, and discharge dates)
- Radiographic imaging
    - The radiologist report from all chest imaging (chest x-ray or chest computed tomography scans [CTs]) and abdominal CTs from two days prior to the hospital encounter and including the first four days of the hospitalization (using the first date of the hospital encounter as day 1)
- Signs and symptoms of pneumonia in the first two days of hospitalization or two days prior to hospital encounter
    - Physical exam findings (e.g., rales)
    - Vital signs (e.g., hypoxia)
    - Documented symptoms (e.g., worsening cough)
    - Laboratory findings (e.g., leukocytosis)
- Antibiotic use during hospitalization and on discharge
- Patient comorbid conditions including dementia, chronic obstructive pulmonary disease, pulmonary fibrosis, interstitial lung disease, asthma, mild immune suppression, heart failure
- Use of home oxygen
- Blood and respiratory cultures
- 30-day adverse events (emergency department visit, mortality, *Clostridioides difficile* infection, antibiotic associated side effects) documented in the medical record
- 30-day adverse events collected via telephone interview (conducted 30-days post discharge)

References:

1. Vaughn VM, Flanders SA, Snyder A, et al. Excess Antibiotic Treatment Duration and Adverse Events in Patients Hospitalized With Pneumonia: A Multihospital Cohort Study. Ann Intern Med. 2019 Aug 6;171(3):153-163.

2. Vaughn VM, Gandhi T, Conlon A, et al. The Association of Antibiotic Stewardship With Fluoroquinolone Prescribing in Michigan Hospitals: A Multi-hospital Cohort Study. Clin Infect Dis. 2019 Sep 27;69(8):1269-1277.
3. Vaughn VM, Gandhi TN, Chopra V, Petty LA, Giesler DL, Malani AN, Bernstein SJ, Hsaiky LM, Pogue JM, Dumkow L, Ratz D, McLaughlin ES, Flanders SA. Antibiotic Overuse After Hospital Discharge: A Multi-hospital Cohort Study. Clin Infect Dis. 2021 Dec 6;73(11):e4499-e4506.
4. Vaughn VM, Gandhi TN, Hofer TP, Petty LA, Malani AN, Osterholzer D, Dumkow LE, Ratz D, Horowitz JK, McLaughlin ES, Czilok T, Flanders SA. A Statewide Collaborative Quality Initiative To Improve Antibiotic Duration And Outcomes Of Patients Hospitalized With Uncomplicated Community-Acquired Pneumonia. Clin Infect Dis. 2021 Nov 13:ciab950.

2a.03) Provide the dates of the data used in testing.

07-01-2017 to 03-31-2020

2a.04) Select the levels of analysis for which the measure is tested.

Facility

2a.05) List the measured entities included in the testing and analysis (by level of analysis and data source).

Table 1. Characteristics of Participating Hospitals

Hospital Characteristic	HMS Hospitals n=49; n (%)	All Michigan Hospitals[1] n=127; n (%)
Academic Hospital[1]	40 (82%)	74 (58%)
Location[2,3]	*	*
Metropolitan	40 (82%)	71 (56%)
Micropolitan	8 (16%)	24 (19%)
Rural	1 (2%)	32 (25%)
Profit Type[2]	*	*
Non-Profit	45 (92%)	116 (59%)
For profit	4 (8%)	9 (33%)
Government	0 (0%)	2 (2%)
Bed Size (Staffed beds)[4]	*	*
≤50	2 (4%)	46 (36%)
51-100	4 (8%)	21 (17%)
101-200	9 (18%)	16 (13%)
>200	34 (69%)	44 (35%)

*Cells intentionally left empty

Data compiled from the following sources:
[1] List of Michigan Hospitals compiled from the Michigan Health & Hospital Association[§] mha.org/about/our-hospitals Accessed January 3, 2022

[2] U.S. Census Bureau, Michigan: 2020 Core Based Statistical Areas and Counties https://www2.census.gov/programs-surveys/metro-micro/reference-maps/2020/state-maps/26_Michigan_2020.pdf

[3] U.S. Census Bureau, Core based statistical areas (CBSAs), metropolitan divisions, and combined statistical areas (CSAs) https://www.census.gov/geographies/reference-files/time-series/demo/metro-micro/delineation-files.html

[4] American Hospital Directory, Individual Hospital Statistics for Michigan https://www.ahd.com/states/hospital_MI.html

[§]The following types of hospitals were excluded:

- Children's hospitals
- Long-term acute care hospitals
- Psychiatric/mental health/substance abuse hospitals
- Rehabilitation hospitals
- Surgical hospitals
- Those providing only specialty services (i.e., cardiac hospital)

Alt-text for Table 1: Participating HMS hospitals (N=49) are compared to all Michigan hospitals (N=127) for proportion classified as academic; location; profit type; and bed size (staffed beds). Relative to all Michigan hospitals, more HMS hospitals were academic (82% vs 58%), located in metropolitan areas (82% vs 56%), were non-profit (92% vs 59%), and had >200 beds (69% vs 35%).

2a.06) Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis), separated by level of analysis and data source; if a sample was used, describe how patients were selected for inclusion in the sample.

Between 7/1/2017 and 3/31/2020 there were 18,625 hospitalized patients treated for CAP across 49 HMS hospitals. All 18,625 patients were used to test validity and reliability of the overdiagnosis of CAP measure. Of the 18,625 patients treated for CAP, 12.3% (2,299) were assessed to be inappropriately diagnosed with CAP. Reliability and validity were assessed at both the hospital and validity was assessed at encounter (i.e., patient) level. Descriptive characteristics of the entire pneumonia cohort are as follows:

Table 2. Descriptive characteristics of the entire pneumonia cohort, patients with appropriate diagnosis, and patients with inappropriate diagnosis

Characteristic	Entire Pneumonia Cohort, n (%)	Appropriate Diagnosis, n (%)	Inappropriate Diagnosis, n (%)
Gender	*	*	*
Male	9,322 (50%)	8,193 (50.2%)	1,129 (49.1%)
Female	9,303 (49.9%)	8,133 (49.8%)	1,170 (50.8%)
Race			
White	14,056 (75.4%)	12,356 (75.7%)	1,700 (73.9%)
Black	3,847 (20.6%)	3,327 (20.4%)	520 (22.6%)
Asian	100 (0.5%)	92 (0.6%)	8 (0.3%)
American Indian	44 (0.2%)	40 (0.2%)	4 (0.2%)

Characteristic	Entire Pneumonia Cohort, n (%)	Appropriate Diagnosis, n (%)	Inappropriate Diagnosis, n (%)
Native Islander	30 (0.2%)	26 (0.2%)	4 (0.2%)
Other	270 (1.4%)	244 (1.5%)	26 (1.1%)
Unknown	220 (1.2%)	186 (1.1%)	34 (1.5%)
Age (years)	*	*	*
18-30	542 (2.9%)	487 (3.0%)	55 (2.4%)
31-40	804 (4.3%)	729 (4.5%)	75 (3.3%)
41-50	1,401 (7.5%)	1,264 (7.7%)	137 (6.0%)
51-60	2,943 (15.8%)	2,601 (15.9%)	342 (14.9%)
61-70	4,216 (22.6%)	3,714 (22.7%)	502 (21.8%)
71-80	4,159 (22.3%)	3,625 (22.2%)	534 (23.2%)
80-90	3,387 (18.2%)	2,911 (17.8%)	476 (20.7%)
91-100	1,127 (6.0%)	958 (5.9%)	169 (7.3%)
100+	52 (0.3%)	41 (0.3%)	11 (0.5%)
Insurance Status	*	*	*
Private	2,568 (13.8%)	2,301 (14.1%)	267 (11.6%)
Medicare	12,024 (64.5%)	10,414 (63.8%)	1,610 (70%)
Medicaid	2,199 (11.8%)	1,962 (12.0%)	237 (10.3%)
Uninsured	267 (1.4%)	242 (1.5%)	25 (1.1%)
Comorbidities	*	*	*
Renal disease	5,300 (28.4%)	4,671 (28.6%)	629 (27.3%)
Liver disease	927 (5.0%)	830 (5.1%)	97 (4.2%)
Congestive heart failure	5,015 (26.9%)	4,413 (27.0%)	602 (26.2%)
Chronic obstructive pulmonary disease	8,888 (47.7%)	7,784 (47.7%)	1,104 (48.0%)
Home oxygen	3,015 (16.2%)	2,664 (16.3%)	351 (15.3%)
Structural lung disease	1,672 (9.0%)	1,484 (9.1%)	188 (8.2%)
Current/Former smoker	12,409 (66.6%)	10,926 (66.9%)	1,483 (64.5%)
Cancer	4,357 (23.4%)	3,864 (23.7%)	493 (21.4%)
Immune compromise	357 (1.9%)	325 (2.0%)	32 (1.4%)
Diabetes mellitus	5,641 (30.3%)	4,896 (30.0%)	745 (32.4%)
Sepsis	6,003 (32.2%)	5,414 (33.1%)	589 (25.6%)
Severe Sepsis	5,679 (30.5%)	5,065 (31.0%)	614 (26.7%)

*Cells intentionally left empty

Alt-text for Table 2. Descriptive characteristics of the entire pneumonia cohort, patients with appropriate diagnosis, and patients with inappropriate diagnosis, including gender, race, insurance status, and co-morbidities.

Hospitals within HMS use the following case identification strategy to determine patients to abstract:

- Data collection involves abstraction of eligible cases every two weeks.
- To minimize sampling bias, abstractors are expected to select cases from every day during a two-week time period, including weekends.
- The list of cases eligible for abstraction is created using the below protocol

- For each two-week period, a list of patients admitted to all medical services is created
    - For Inappropriate diagnosis of pneumonia, this list is generally a list of all patients with an ICD-10 code for pneumonia
    - If possible, hospitals apply additional electronic filters to the dataset to screen for inclusion/exclusion criteria. For example, they may exclude patients from the "inappropriate diagnosis of pneumonia" list if they did not received antibiotics on day 1 or 2 of hospitalization or if they were mechanically ventilated during hospitalization.
        - All inclusion/exclusion criteria that are not electronically applied prior to list generation will require manual screening during case review
    - The list of potentially eligible patients is then organized chronologically by date and time of discharge.
    - For each discharge day, the first patient on the chronological list is reviewed for inclusion. If excluded, the next patient is reviewed.
    - This process is repeated, with patients reviewed from the chronological list ensuring that cases are distributed evenly across the two-week timeframe – meaning there are discharge dates across all days of the week – until all cases are identified and abstracted.

We do not report encounter-level reliability as we report encounter-level validity. Please see the validity documents for additional information.

2a.07) If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing.

All data analysis was performed on the same dataset.

Table 3. Description of samples utilized to determine hospital-level and encounter-level reliability and empirical validity

Type of Testing	**Sample Utilized**
Hospital-Level Reliability and Empirical Validity[1]	Entire HMS Pneumonia Dataset (based on case identification protocol outlined in 2a.06)
Encounter-Level Reliability[1]	*Assessment of the Effect of Abstraction Errors*: Review of a random, consecutive subset of 50 encounters within the cohort, representing cases from 33 of 49 participating hospitals. *Structured Implicit Case Review*: Seventeen cases, pseudo-randomly selected, for in-depth review by 2-4 physicians to confirm case classification (appropriate versus inappropriate diagnosis)

[1]Please see validity documents for further information.

Alt-text for Table 3: The entire HMS pneumonia dataset was used to determine hospital-level reliability and empirical validity. Encounter-level reliability was determined by assessment of the effect of abstraction errors and structured implicit case reviews.

2a.08) List the social factors that were available and analyzed.

As this is a process measure, no risk adjustment was performed (including for social factors).

2a.09) Select the level of reliability testing conducted.

Patient or Encounter-Level
Accountable Entity Level (e.g., signal-to-noise)

2a.10) For each level of reliability testing checked above, describe the method of reliability testing and what it tests.

<u>Patient or Encounter Level</u>

Please see validity testing section for encounter-level validity.

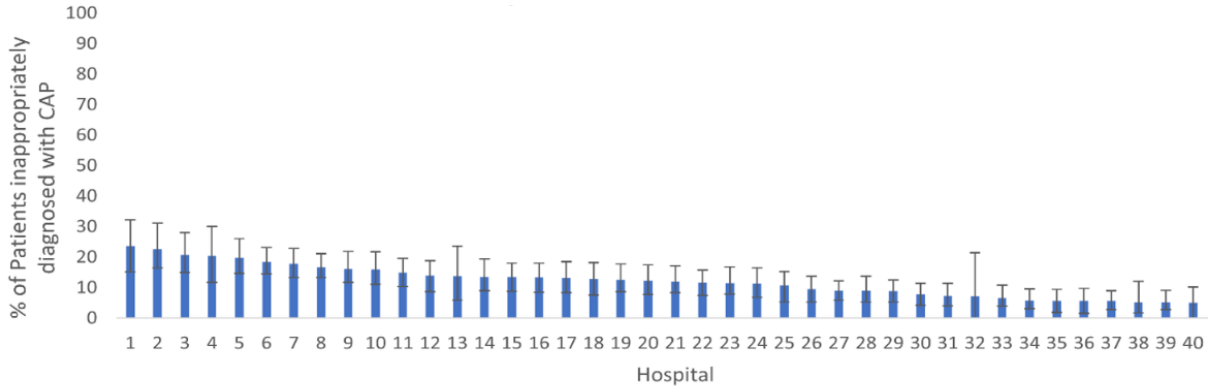
<u>Accountable Entity Level</u>

Signal-to-noise analysis was performed using a mixed-effect logistic model ran as an empty model such that the only effects in the model were the overall intercept and the hospital specific intercepts. This model enabled for the calculation of the hospital variance (signal), the total variance, and the within hospital variance (noise). Based on the hospital variance and the within hospital variance, an intraclass correlation was calculated. The intraclass correlation was utilized within the Spearman Brown formula in two ways: (A) to calculate the reliability for the entire hospital cohort using the median number of case abstractions for the cohort and (B) to understand minimum case abstracts necessary to achieve predetermined reliability thresholds of 0.6, 0.7, 0.8, and 0.9.

2a.11) For each level of reliability testing checked above, what were the statistical results for reliability testing?

Distribution of percentage of patients inappropriately diagnosed with CAP by hospital with 95% confidence intervals is demonstrated below. These data are based on the 4 quarters preceding March 2020 and includes only hospitals that provided data during all four quarters.

Figure 1. Distribution of Inappropriate diagnosis of Community-Acquired Pneumonia by Hospital



Alt-text for Figure: Distribution of percentage of patients inappropriately diagnosed with CAP by hospital with 95% confidence intervals ranges from 4.2% to 23.7%. Data are based on the 4 quarters

From these data, we were able to calculate the following:

From these data, we were able to calculate the following:

Hospital Variance (signal): 0.18235
Total Variance: 3.4722
Within Hospital Variance (noise): 3.28987

Based on this information, an intraclass correlation (ICC) was calculated. This ICC represents the reliability of the cohort if a single measurement (case abstraction) per hospital were included.

ICC=0.18235/(0.18235+3.28987)=0.18235/3.4722=0.0525

A. The Spearman Brown Prophecy allows to an estimation of reliability after adjusting the number of measurements. We can use this formula to estimate the reliability of the measure within the cohort after adjusting the input (in this case the number of case abstractions per site).[1,2] The Spearman Brown Formula states the following:

   $Reliability_{new} = (n*r)/(1+[n-1]*r)$ where n is the number of inputs and r is the prior reliability.

   Adapting to the formula to our variables suggests the following:

   $Reliability_{new}$ = (number of case reviews*ICC)/(1+[number of case reviews-1]*ICC)

   The median case abstraction counts for the entire cohort was applied to the Spearman Brown Formula to obtain the overall reliability for the cohort.

Median case abstractions: 184 (IQR 153-201)
Reliability: (184*0.0525169)/(1+(184-1)*0.0525169)=0.911

1. Spearman, C. (1910), Correlation Calculated From Faulty Data. *British Journal of Psychology*, 1904-1920, 3: 271-295.

2. Warrens MJ. Transforming intraclass correlation coefficients with the Spearman-Brown formula. *J Clin Epidemiol*. 2017 May;85:14-16

B. The ICC was then be applied to the Spearman Brown Formula to calculate the minimum number of cases to achieve pre-specified reliability thresholds based on the outcome distribution of the entire cohort.

Table 1. Number of annual cases needed to achieve each reliability threshold.

Reliability	**Number of annual cases needed**
0.6	28
0.7	43
0.8 (standard)	**73**
0.9	163

Alt-Text for Table 1. In order to achieve a desired reliability of 0.8, each hospital would need to abstract 73 cases annually.

2a.12) Interpret the results, in terms of how they demonstrate reliability.

A. Based on signal-to-noise analysis, we found that reliability of the measure across the entire hospital cohort was strong (0.91), meeting the threshold for reliability for measures considered to be high stakes.

B. Using the current HMS cohort as a representative example, the minimum number of case abstracts per hospital per year to meet pre-specified reliability thresholds of 0.7 and 0.8 are highly attainable. Within a cohort of 40 HMS hospitals participating in 2019, 92.5% of hospitals were able to abstract the minimum of 73 cases to achieve 0.8 reliability. Of those that could not abstract the required number of cases, hospital bed sizes were 68 beds, 133 beds, and 317 beds, the latter two of which had data abstractor hiring challenges. All but one hospital (133 beds) could abstract the 43 cases/year necessary to achieve 0.7 reliability. This cohort of 40 hospitals participating in 2019 was selected as this represented the last year prior to the COVID-19 pandemic.