NATIONAL QUALITY FORUM
Driving measurable health improvements together

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3671
Corresponding Measures:
Measure Title: Inappropriate diagnosis of community-acquired pneumonia (CAP) in hospitalized medical patients; Abbreviated form: Inappropriate diagnosis of CAP
Measure Steward: University of Michigan
sp.02. Brief Description of Measure: The inappropriate diagnosis of CAP in hospitalized medical patients (or "Inappropriate Diagnosis of CAP") measure is a process measure that evaluates the annual proportion of hospitalized adult medical patients treated for CAP who do not meet diagnostic criteria for pneumonia (thus are inappropriately diagnosed and treated).
1b.01. Developer Rationale: The goal of this measure is to improve diagnosis and treatment of pneumonia. Literature has demonstrated that while pneumonia is the most common infectious etiology for which patients are hospitalized, it is often inappropriately diagnosed, resulting in unnecessary antibiotic administration and delay in diagnosis of true underlying conditions. The implications of unnecessary antibiotics are well described and include risks of antibiotic-associated adverse events such as *Clostridioides difficile* infection, prolonged length of hospital stay, and antimicrobial resistance, all of which can increase patient morbidity and mortality. Missed or delayed diagnosis of a true underlying condition are equally troubling, as data suggest that diagnostic error results in the highest morbidity, mortality, and malpractice cost of any medical error. Through adoption of this measure, we anticipate a decrease in inappropriate diagnosis of pneumonia, a decrease in unnecessary antibiotic use, and improved patient outcomes.

---

sp.12. Numerator Statement: The measure quantifies adult, hospitalized medical patients inappropriately diagnosed with pneumonia. Here, inappropriate diagnosis is defined as patients treated with antibiotics for CAP who do not meet diagnostic criteria for pneumonia. Patients are considered inappropriately diagnosed if they did not have 2 or more signs or symptoms of pneumonia (documented at some point in the 2 days prior to the hospital encounter through the first 2 days of the hospital encounter) AND meet radiographic criteria for pneumonia.
sp.14. Denominator Statement: The denominator includes all adult, general care, immunocompetent, medical patients hospitalized and treated for CAP who do not have a concomitant infection.
sp.16. Denominator Exclusions:
Patients are excluded from the denominator if they are/have:

- Left against medical advice or refused medical care
- Admitted on hospice
- Pregnant or breastfeeding
- Cystic fibrosis
- Pneumonia-related complication (e.g., empyema)

Measure Type: Process
sp.28. Data Source: Electronic Health Records; Electronic Health Data; Other (specify)
Chart review

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a ***structure, process or intermediate outcome*** measure are that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following description for this measure:

- This is a new process measure at the facility level that evaluates the annual proportion of hospitalized adult medical patients treated for community-acquired pneumonia (CAP) who do not meet diagnostic criteria for pneumonia (thus are inappropriately diagnosed and treated).
- The developer provides a logic model that depicts the connection between patients inappropriately diagnosed with CAP and several negative health outcomes that can result, including a delayed or missed diagnosis of an unrelated underlying condition affecting the patient that might itself result in harm, as well as adverse events from administering the antimicrobial agents, and increasing antimicrobial resistance in the individual patient and in the patient's broader community.

The developer provides the following evidence for this measure:

- Systematic Review of the evidence specific to this measure? ☐ Yes ☒ No
- Quality, Quantity and Consistency of evidence provided? ☐ Yes ☒ No
- Evidence graded? ☐ Yes ☒ No

Summary:

- Although no systematic reviews or guidelines are presented, the developer cited two studies supporting that CAP is inappropriately diagnosed in hospitals:
    - An 2010 study found that 27.3% of patients admitted to the hospital from the emergency department with a diagnosis of CAP had a different, non-pneumonia diagnosis at discharge.
    - A 2019 study found that 29% of patients admitted for CAP had a different diagnosis on discharge.
- The developer provided many studies supporting the harm associated with unnecessary antibiotic use, including:
    - Antibiotic-associated adverse events: a 2017 study found that as many as 20% of patients receiving antibiotics experienced at least 1 antibiotic-associated adverse drug event, and a 2014 study estimated that each day of excess treatment with antibiotic therapies increased the odds of an antibiotic-associated adverse event by 5%, without lowering rates of adverse outcomes.

- Missed or delayed diagnosis: the developers cited studies of missed diagnoses of pulmonary malignancy and heart failure. As well, a 2007 study suggested that a since-revised guideline by the Infection Diseases Society of America recommending the initiation of antibiotic therapy within four hours of hospitalization reduced the rate of final diagnosis of CAP from 75.9% to 58.9%.
- Developing antibiotic resistance: a 2014 systematic review found antibiotic consumption is associated with the development of antibiotic resistance.

Exception to evidence

- N/A

Questions for the Committee:

- *Is the Committee confident that the evidence presented is sufficient to link inappropriate diagnoses of CAP to undesirable health outcomes?*
- *How strong is the evidence for this relationship?*
- *Does the evidence of harms presented by antibiotic therapies in inappropriate diagnoses outweigh potential benefits to administering these even in cases without a clear diagnosis?*

Guidance from the Evidence Algorithm

Process measure not based on a systematic review (Box 3) -> Empirical evidence submitted (Box 7) -> A comprehensive set of studies included (Box 8) -> The submitted evidence indicates a high certainty that benefits clearly outweigh harm (Box 9) -> Rate as MODERATE

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The measure was tested from 07/01/2017-03/31/2020 in 49 Michigan hospitals, finding 18,625 patients treated for a CAP, of whom 12.3% were inappropriately diagnosed.
- The developer reported hospitals by performance decile.
    - In 2017, the median hospital in the best performing decile had 5.6 percent of cases inappropriately diagnosed with CAP. The worst performing decile had 26.8 percent of cases inappropriately diagnosed with a CAP.
    - In 2019, the median hospital in the best performing decile had 4.5 percent of cases inappropriately diagnosed with a CAP. The worst performing decile had 22.4 percent of cases inappropriately diagnosed with a CAP.

Disparities

- In analyzing the demographics of the entire cohort, the developer found no differences in rates of inappropriate diagnosis by gender or race; however, a significant difference was identified by payer.
    - Medicare patients were more likely to be inappropriately diagnosed than those with Medicaid or private insurance. Patients age 65 years or older were also more likely to be inappropriately diagnosed.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

- *Are there any concerns about the presence of disparities in this measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

1a. Evidence

- This measure looks at patients being given antibiotics for CAP who do not meet a modified NHSN HAP criteria. Evidence is presented that inappropriate diagnosis of pneumonia results in excessive antibiotic use and potential harms. While that is certainly true, making a diagnosis of pneumonia (or excluding it) is clinically very challenging even as an experienced ID physician and certainly not captured by an algorithm of various criteria. If only it were that easy. The NHSN criteria, which I have used, is neither sensitive nor specific in defining hospital-onset pneumonia. I have not seen data presented that the criteria adequately identifies CAP. Since this is a metric about diagnosis, it is critical that the criteria being used, does in fact identify community-acquired pneumonia with adequate sensitivity and specificity. This fundamental issue has not been addressed other than having a few physicians review 17 cases thought to be inappropriate diagnosis of CAP. The unintended clinical consequences of delayed diagnosis and sepsis (pneumonia is the most common cause of sepsis Novosad MMWR 2016) have not been evaluated or studied.
- Agree with moderate rating, as evidence provided, while indirect, is compelling about over diagnosis of CAP and potential consequences
- This is a process measure designed to evaluate the annual proportion of hospitalized adult medical patients inappropriately diagnosed and treated for community-acquired pneumonia (CAP). The inappropriate diagnosis of patients with CAP results in overuse of antibiotics, placing patients at increased risk of developing complications such as C.diff. Patients are considered inappropriately diagnosed if they do not meet diagnostic criteria for pneumonia.
- A process measure. Numerator: Patients treated with antibiotics for CAP who do not meet diagnostic criteria. Denominator: All adults hospitalized patient treated for CAP with no accompanying infection. No graded systematic review available. Two studies demonstrate that almost 30% of patients admitted for CAP had a different diagnosis at discharge. Another systematic review demonstrated with abx consumption lead to antimicrobial resistance.
- process measure - moderate evidence
- This is a process measure that evaluates the annual proportion of hospitalized adult medical patients treated for CAP who do not meet diagnostic criteria for pneumonia (thus are inappropriately diagnosed and treated). The developer cited two studies from 2010 and 2019 that showed over a quarter of patients had inappropriate diagnosis of pneumonia in hospitals. Citing a number of studies, the developer supplied a logical model to demonstrate that misdiagnosis of community-acquired pneumonia may lead to antimicrobial-related adverse events and antimicrobial resistance as well as patient harm resulting from delayed or missed true diagnosis. The preliminary rating is moderate.
- Agree
- I don't believe the evidence submitted really supports this measure. The studies show discrepancy between ED/admission diagnosis and discharge diagnosis, that's not the same as having a discharge diagnosis of CAP and not meeting criteria. It's also not clear that having a discharge diagnosis of CAP while not meeting criteria necessarily leads to excessive abx use - possible there's a different source of infection. I don't believe evidence to be sufficient.

- The denominator requires further explanation. The evidence for supporting this measure seems immature to me. I see little harm caused by application of this measure.
- This process measure is supported by a logic model and literature that demonstrates a 27 - 29% rate of misdiagnosis of CAP. Further, evidence from the literature of the impact of poor antibiotic stewardship is offered.

1b. Gap in Care/Opportunity for Improvement and Disparities

- The performance gap seems to center on the difference in admission diagnosis and discharge diagnosis. Otherwise there are data presented showing variability in meeting a set of criteria (which has not yet been shown to be a relevant set of criteria) I am not sure that is an appropriate way to assess performance gap
- Significant spread in the data, suggesting improvement opportunities. Disparities by age appear to be present (Medicare beneficiaries typically older)
- The developer did not provide systematic reviews or guidelines, but cited two studies supporting that CAP is inappropriately diagnosed in hospitals and supporting studies associated with unnecessary antibiotic use.
- Performance variation noted between 2017 and 2019; 2017, best decile at 5.6% of CAP cases inappropriately diagnosed, in 2019 that number decreased to 4.5%. Percentage decreased in worst performing decile as sell (2017 26.8%, 1019 22.4%).
- Moderate gap
- The measure was tested from 07/01/2017-03/31/2020 in 49 Michigan hospitals. Among 18,625 patients treated for a CAP, 12.3% were inappropriately diagnosed, which demonstrates an opportunity for improvement. The developer found no differences in rates of inappropriate diagnosis by gender or race. But a significant difference was identified by the payer: Medicare patients were more likely to be inappropriately diagnosed than those with Medicaid or private insurance. Patients age 65 years or older were also more likely to be inappropriately diagnosed.
- no concerns
- Significant performance gap shown.
- The performance gap is huge, suggesting the need for a national performance measure. No concerns about disparities.
- Yes, data were provided and disparities by payer were noted (Medicare patients more often misdiagnosed).

Criteria 2: Scientific Acceptability of Measure Properties

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluators: Staff

2a. Reliability: Specifications and Testing

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Specifications:

- Measure specifications are complex but clear and precise.

Reliability Testing:

- This dataset represents 18,625 hospitalized patients treated for CAP across 49 hospitals in the Michigan Hospital Medicine Safety Consortium (HMS) from 07/01/2017-03/31/2020.
    - This dataset contained 82 percent academic hospitals, 82 percent metropolitan, 92 percent non-profit, and 69 percent hospitals with greater than 200 staffed beds.

- Reliability testing conducted at the Accountable Entity Level:
    - The developer performed a signal-to-noise analysis using a mixed-effect logistic model run as an empty model to calculate hospital variance (signal), within hospital variance (noise), and total variance, which were used to calculate the intraclass correlation coefficient (ICC).
        - Total variance: 3.4722
        - Hospital variance: 0.18235
        - Within hospital variance: 3.28987
        - Based on signal-to-noise analysis, the developer calculated an ICC of 0.0525
            - An ICC below 0.5 generally indicates poor reliability.
    - The ICC was used in a Spearman Brown formula to calculate reliability for the entire hospital cohort using the median number of case abstracts and to determine the minimum number of cases needed to achieve predetermined reliability thresholds (0.6, 0.7, 0.8, and 0.9).
        - After applying the median number of case extractions, the developer determined that reliability for the entire hospital cohort was 0.911.
        - Within a cohort of 40 hospitals in 2019, 92.5% of hospitals in the cohort were able to abstract 73 or more cases to achieve 0.8 reliability. All but one hospital (with 133 beds) were able to abstract the minimum 43 cases/year needed to reach 0.7 reliability.
- Reliability testing conducted at the Patient/Encounter Level:
    - Encounter-level validity was determined by assessment of effect of abstraction errors and structured implicit case reviews. The developer states that validity testing was conducted on all critical data elements, but since individual data element results were not provided in the submission and only the overall score was provided, NQF does not view this as sufficient to constitute complete patient/encounter level validity testing. It therefore is also insufficient for reliability testing at the patient/encounter level.

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure cannot be consistently implemented (i.e., are measure specifications adequate)?*

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

Specifications are precise and unambiguous (Box 1) -> Reliability was conducted with the measure as specified (Box 2) -> Reliability testing conducted at the accountable entity level (Box 4) -> Signal-to-noise method used to determine reliability but unclear method used for calculating median number of case abstracts and ICC (Box 5) -> Unclear if empirical testing conducted on all critical data elements (Box 8) -> Rate as INSUFFICIENT

In signal-to-noise analysis, the internal variance is greater than the external variance and the intraclass correlation coefficient is well below 0.5, a range generally agreed to show poor reliability. It is not clear from the submission how applying the Spearman Brown prophecy formula leads to an overall reliability of 0.9. Additionally, only an overall score was provided for patient/encounter level reliability testing thus making it difficult to determine if all critical data elements were evaluated.

2b. Validity: Validity testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Validity Testing

- Validity testing conducted at the Patient/Encounter Level:
    - Encounter-level validity was determined by assessment of effect of abstraction errors and structured implicit case reviews. The developer states that validity testing was conducted on all critical data elements, but since individual data element results were not provided in the submission and only the overall score was provided, NQF does not view this as sufficient to constitute complete patient/encounter level validity testing.
    - Using the current measure and data from 33 of hospitals in the cohort using 2021 data, the HMS project manager performed blind audits of 50 consecutive cases of patients counted as inappropriately diagnosed with CAP.
        - Data audit found 93.7% of data elements were abstracted correctly. The developer states that any discrepancies found were minor and resulted in no changes to case classification.
        - The classification of the 50 cases as "CAP" or "inappropriate diagnoses of CAP" by first the abstractor and then the auditor had 100% agreement.
    - Two to three physicians conducted a structured implicit case review to confirm accurate case categorization, using 2020 data. Cases were randomly selected from the "gray areas" identified during measure development (e.g., patients with atelectasis as the only finding on chest imaging). If there was disagreement in classification, the developer prompted a discussion about ways to improve the measure to account for errors in classification.
        - Case review resulted in K=0.72 agreement between physician reviewers on case classification, which the developer considers to be "substantial agreement."
        - Since the case review involved "gray area" cases rather than a random selection, the developer states the true K may be even higher.
        - In 94% of cases (16/17), there was 100% agreement that the cases represented inappropriate diagnosis.

- Validity testing conducted at the Accountable Entity Level:
    - Face Validity:
        - Face validity was assessed in 2021 using the current measure to receive feedback from 38 HMS hospitals, a technical expert panel (TEP), and patients and caregivers.
        - Hospitals were asked "Approximately what percentage of cases called PNA by HMS do you agree are PNA (0-100 percent)? The median response was 90 percent of cases, and the interquartile range 80 percent to 95 percent.
        - Fourteen national experts participated in two weeks of online conversations and responded to survey questions about the measure. The TEP responded to the following:

- a. How much do you agree/disagree with the following statement? "The inappropriate diagnosis of CAP measure as specified can be used to distinguish between better and worse quality hospitals." 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree.
- b. Are there any key data elements you believe are missed or not accurately captured in the inappropriate diagnosis of CAP measure?
- 57% of TEP members (8/14) stated they do not think the measure needs to collect any additional data in order to correctly identify inappropriate diagnosis of CAP.
- The remaining feedback was incorporated into the measure, including adding data on duration of treatment, adding studies of underdiagnosis to the Evidence section, data on outcomes over time, and data on those over age 80.

- Face validity conducted with patients: The developer concluded from panel discussions with patients and caregivers that this group understood the meaning of overdiagnosis and felt that measuring inappropriate diagnosis of infections was both important and meaningful.

- Empirical Validity of Measure Score
  - The developer conducted empirical validity testing by correlating NQF #3671 with NQF# 3690 Inappropriate Diagnosis of Urinary Tract Infection (UTI). The developer states that there were very few measures that assess the same domains of quality as NQF #3671, so after conducting a literature review they found that the inappropriate diagnosis of CAP can represent a signal of hospital quality, which affects patient outcomes.
    - The developer found that NQF #3671 is moderately correlated with NQF# 3690 ($R=0.53$, $p<0.001$). The findings were similar, though slightly less strong, for patients inappropriately diagnosed with either condition in emergency department (ED) settings ($R=0.46$, $p<0.002$).
    - The developer states that this shows that inappropriate diagnosis of the CAP measure may reflect the overall quality of diagnoses made at a hospital.
  - The developer also assessed the association of NQF #3671 with antibiotic-associated adverse events.
    - The developer found that each additional day of antibiotic use in patients inappropriately diagnosed with CAP was associated with an increased odds ratio (1.05) for developing a patient-report antibiotic-associated adverse event.
      - As inappropriate diagnoses of CAP decreased over time, related outcomes improved in HMS hospitals.
      - Death events fell from 3.5% (2017 data, n=6405) to 2.9% (2020 data, n=4961)
      - Adverse-Antibiotic Events fell from 4.8% (2017 data, n=6405), to 3.0% (2020 data, n=4961)

Exclusions

- The developer lists several exclusions to the measure (patients who left against medical advice or refused care, who were pregnant or breastfeeding, who were admitted on hospice or comfort care, who had cystic fibrosis, or who had a pneumonia-related complication) and states the exclusions were determined through careful clinical review and feedback from experts. Exclusions were not common and represented approximately 2.68%-3.35% of the testing population.

Meaningful Differences

- The developer requested all hospitals in the cohort report the distribution of their measure scores then grouped hospitals by quartiles to determine whether the difference in mean measure score was statistically significant.
    - Hospitals in the 10th percentile (better performance) have about 7 fewer patients inappropriately diagnosed with CAP per 100 CAP discharges than the median.
    - Hospitals in the 90th percentile (worse performance) have approximately 10 more patients inappropriately diagnosed with CAP per 100 CAP discharges than the median.
    - The difference in performance between all adjacent quartiles (1st vs. 2nd, 2nd vs. 3rd, 3rd vs. 4th) was statistically significant ($p<0.001$ for all comparisons).

Missing Data

- The developer found missing data to be extremely rare; the percentage of patients in the testing cohort with "unknown" or "not available" values was less than 1.0% (183/18,468 patients).

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

2a. Reliability

- 2a1. Reliability-Specifications
    - The reliability of the component variables are not very good for CXR findings. There is a lot of subjective variation in interpretation and different words used in the reports.
    - Seems very unclear, and the data suggest some very low levels of reliability. This is often challenging to discern in acute care settings
    - The dataset provided represents 18,625 hospitalized patients treated for CAP across 49 hospitals in the Michigan Hospital Medicine Safety Consortium from July 207-March 2020. Reliability testing was completed using a mixed effect logistic model and calculating variance.
    - None
    - not reviewed by SMP - insufficient evidence
    - No concerns.
    - Insufficient
    - I agree with staff that reliability and validity testing is not sufficient.
    - Limited evidence suggests the measure can be consistently applied.
    - at an accountable entity level: Based on signal-to-noise analysis, the developer calculated an ICC of 0.0525 • An ICC below 0.5 generally indicates poor reliability.
- 2a2. Reliability – Testing
    - The developer used intercorrelation coefficients which showed greater variability at the individual level than the facility level, but not sure that should be a concern or threat to reliability on its own.

- Yes
- The intraclass correlation coefficient (ICC) was 0.0525 where an ICC below 0.5 generally indicated poor reliability. Encounter level validity was determined by assessment of abstraction errors. The SMP concluded that preliminary rating for reliability was "insufficient."
- Poor reliability at the hospital level (ICC of 0.0525); Individual data element results were not provided; insufficient for reliability testing at the patient/encounter level
- insufficient evidence
- Based on signal-to-noise analysis, the developer calculated intraclass correlation coefficient is 0.0525. An ICC below 0.5 generally indicates poor reliability. The preliminary rating on reliability is insufficient.
- yes.
- see above comment
- No.
- It is not clear from the submission how applying the Spearman Brown prophecy formula leads to an overall reliability of 0.9. Validity testing was conducted on all critical data elements, but since individual data element results were not provided in the submission and only the overall score was provided,

2b. Validity

- The clinical validity of using these criteria to appropriately define community-acquired pneumonia has not been established.
- Agree with moderate rating and it appears the TEP recommendations and further testing support the face validity
- The developer presented several different methods of validity testing, including audits and extrapolations.
- At the encounter level audits yielded 93.7% accurate data abstraction; Physician implicit review also performed with K= 0.72 (substantial agreement)--individual data element results were not provided to NQF. At the hospital level: face validity performed with a survey questions to hospitals, TEP members and patients. Correlation between Inappropriate CAP and Inappropriate UTI diagnosis was moderate.
- moderate evidence
- No concerns.
- yes.
- see above comment
- Validity was well demonstrated in the limited situations presented.
- no concerns

2b2-2b6. Potential threats to validity

- 2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment)
    - The fundamental validity has not been established. If the developers were able to do this, the metric would most likely need to be risk adjusted or stratified as it is clear from clinical experience, the diagnosis of CAP is easier or harder in certain groups (NH patients, underlying lung dz, malignancy, heart failure etc
    - Exclusions appropriate and do not seem to be high %
    - Several exclusions are listed including patients who left AMA or refused care, were pregnant or breastfeeding, or admitted to hospice or comfort care or who had CF. These represented 2.68-3.35% of the testing population.

- Process measure.
- ??
- Noy specified.
- not risk adjusted
- N/A
- I see no problems. Exclusions seem appropriate.
- Exclusions appear appropriate. No risk adjustment.

- 2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data)
  - As it stands the metric is too ambiguous and flawed to be able to be meaningfully used.
  - <1%, no concerns
  - The developer found missing data to be extremely rare.
  - No, missing data was rate (less than 1%).
  - no
  - Missing data was found to have minimum impact on validity. Validity is rated as moderate.
  - no concerns
  - no concerns
  - Missing data were rare.
  - Missing data is rare according to the developer. Meaningful differences about quality appear to be captured.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The measure uses chart-abstracted data routinely collected during the normal process of patient care; no additional data are required. In the studied hospital cohort, the developer reported less than 1.0% of encounters had missing data.
- Some data elements needed to calculate the measure must be chart-abstracted, and the developer found the measure was not feasible to transition to an eCQM.
- The data elements that must be abstracted are the symptoms of CAP, which are generally documented in the medical record in free text, with locations that vary based on the medical record and site-specific implementation factors.
- The minimum cases that need to be abstracted in order to meet a reliability threshold of .7 is just 43 cases/year, which all but one of the 49 studied hospitals were able to meet. To meet a reliability threshold of .8, 73 cases must be sampled, which 92.5% of studied hospitals were able to do.
- The developers surveyed studied hospitals, only 20% of whom reported it was “difficult” or “very difficult” to collect the needed data.

Questions for the Committee:

- *Is the Committee confident that the experience of the hospital cohort studied by the developer is broadly representative of hospitals which may report this measure nationwide?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

3. Feasibility

- As i understand the metric, it involves a fair amount of data abstraction on a sample of 73 patients a year
- Abstraction seems very labor intensive
- The developer stated a minimum number of cases that need to abstracted in order to meet a relaibility threshold of .7 is just 43 cases/year.
- Requires some chart abstractions (for symptoms of CAP) and 20% of whom reported it was “difficult” or “very difficult” to collect the needed data.
- moderate
- Among 49 studied hospitals, all but one was able to meet a reliability threshold of o.7 and 92.5% of the studied hospitals were able to meet the liability threshold of 0.8. 80% of survey hospitals report no difficulty in collecting the needed data for the measure. Feasibility is rated as moderate.
- no concerns
- There is added burden of chart abstraction.
- Seems to me that data collection may be a bit tedious given the specific criteria of the measure's numerator.
- While the data needed comes from chart abstraction, it appears that most hospitals were able to complete the abstraction for the number of cases required without difficulty

Criterion 4: Use and Usability

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. Use evaluates the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☐ Yes ☒ No

Current use in an accountability program? ☐ Yes ☒ No ☐ UNCLEAR

Planned use in an accountability program? ☐ Yes ☒ No ☐ NA

Accountability program details

- The measure is currently used in an external benchmarking program. Blue Cross Blue Shield of Michigan sponsors the Michigan Hospital Medicine Safety Consortium (HMS), which benchmarks hospitals against their own prior performance on this measure and offers comparisons to other hospitals in the program.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- Hospitals participating in HMS were given data on their performance relative to other hospitals, and with specific lists of each patients that was considered inappropriately diagnosed with CAP in order to permit hospitals to review those cases. Hospitals then provided case-specific feedback back to the developer.
- The developer also sought open-ended feedback on the measure from participating hospitals, as well as asking about specific barriers to using the measure. In addition, the developer conducted a patient engagement panel, and no concerns were raised about the measure from patients.
- The developer updated the measure to reduce the number of cases needed to abstract to obtain a reliable measure result and to reduce the number of data elements needed.

***Questions for the Committee**:*

- *Has the measure been sufficiently vetted in real-world settings by those being measured or others?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability evaluates the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- Since launching this measure in the HMS accountability program in 2017, the developer observed a 32 percent decrease ($p<0.001$) in the percentage of patients inappropriately diagnosed with CAP, when measured against results in the first quarter of 2020. This improvement is attributable to the external benchmarking program.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

- The developer identified some unexpected benefits from implementation of the measure, including improved awareness of the duration of therapy for other infections, and two hospitals reported changing their default order sets for suspected pneumonia.

Potential harms

- The developer reported that only 22.5% of hospitals in the cohort foresaw possible unintended consequences from the implementation of this measure, including possible delays in administration of

antibiotics for patients who have a CAP and difficulty in obtaining cooperation from prescribing physicians.

***Questions for the Committee**:*

- *Can the performance results be used to continue to improve performance in an accountability program?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for Usability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments:

4a. Use

- The metric is currently not being used outside of HMS
- Used for internal QI data provided back to hospitals by payers
- The data is not publicly reported, used in accountability programs or planned for that.
- No concerns raised a patient engagement panel, and open-ended feedback as also sought from participating hospitals about barriers.
- not currently being used but passed review
- This measure is not currently used in any public reporting and accountability program. And there is no plan to use it in an accountability program. I wonder if this measure can be evolved into an outcome measure. Also, I would like to encourage the developer to think about future plans to make the results of this measure accessible to the public. A hospital with a lower rate of misdiagnosis on CAP, or an overall lower rate on any misdiagnosis, may indicate care quality for patients and healthcare consumers.
- no concerns
- N/A
- Use is OK.
- Not publicly reported. Used as an external benchmarking measure. Feedback has been considered.

4b. Usability

- There are significant concerns for delayed diagnosis, misclassification, and unintended harm. The developers show "improvement" of documentation or meeting the CAP criteria in Michigan hospitals without a concurrent control (and unclear what exactly has improved).
- Some concerns about how data could be used to drive performance in small hospitals
- The developer stated that in the HMS accountability program, there has been a 32 percent decrease in patients inappropriately diagnosed/treated for CAP. Unexpected benefits included improved awareneess of the duration of therapy for other infections. Potential harm included delays in administration of antibiotics for patients who have a CAP and difficulty obtaining cooperation from prescribing physicians.
- 22.5% of hospitals in the cohort foresaw possible unintended consequences from the implementation of this measure, including possible delays in administration of antibiotics for patients who have a CAP--early warning systems may negate this.
- high - moderate

- The developer reported that, since launching this measure in the HMS accountability program in 2017, there has been a 32% reduction in the percentage of patients inappropriately diagnosed with CAP, when measured against results in the first quarter of 2020. This improvement is attributed to this external benchmarking program. Usability is rated as high.
- no concerns
- worry about unintended consequence of underdiagnosis of CAP, which developer comments already occurs.
- The measure appears to be readily usable and is clearly in the best interest of patients.
- Benefits outweighed any potential unintended consequences. There were "side" benefits noted by the developer that affected treatment protocols for other infections.

Criterion 5: Related and Competing Measures

- No related or competing measures identified.

Committee Pre-evaluation Comments:

5: Related and Competing Measures

- n/a
- None
- none
- none identified
- none
- No related and competing measures.
- no
- None given.
- no concerns

Public and NQF Member Comments (Submitted as of June 10, 2022)

Comments

Comment 1 by: Submitted by Valerie Vaughn, Michigan Hospital Medicine Safety Consortium (*#3671 measure developer)*

This public comment is to address concerns about reliability and validity testing at the critical data element level. We did not include data element validity testing in the original submission but rather reported encounter level validity. We also have data element validity available and include it here: SUMMARY: Critical data element validity testing was conducted by a senior project manager who reviewed all critical data elements from 50 abstracted cases (representing 33 hospitals). Overall, the percent agreement for abstractor and auditor for critical data elements for radiographs ranged from 86% to 91% for chest X-rays and 88% to 92% for chest CTs and for signs/symptoms ranged from 86% to 100%. This suggests that data element validity is high and adds to our already submitted information that encounter level validity is high. DETAILS: Critical data elements for chest radiographs (x-ray and CT) and signs/symptoms of pneumonia were examined by the senior project manager in blind audits of 50 consecutive patients with a diagnosis of CAP (appropriate or inappropriate) from 33 hospitals. Data

elements were scored based correctness of data abstraction (1 point received if data element was answered correctly, 0 points if there was disagreement). The proportion of cases in which there was agreement for each data element were tabulated for clinical findings, chest x-ray findings, chest CT data, and overall abstraction accuracy. Audit findings were as follows: Chest X-ray: Percent agreement between abstractor and auditor for critical data elements Air Space Density/Opacity/Disease 86% Aspiration 91% Aspiration Pneumonia 91% Bronchopneumonia 91% Cannot Rule Out Pneumonia 91% Cavitation 91% Consolidation 91% Ground Glass 91% Infection (cannot rule out infection, likely infection) 89% Infiltrate (Single Lobe) 91% Infiltrate (Multiple Lobes) 86% Interstitial lung disease/interstitial disease 91% Interval improvement or resolution 89% Loculations 91% Mass 91% Necrotizing Pneumonia 91% Neoplasm/Metastatic Disease/Malignancy 91% New or Worsening Infiltrates 91% Nodular Airspace Disease 91% Nodules 91% Pleural Effusion 91% Pneumonia 86% Pneumonitis 91% Post Obstructive Pneumonia 91% Pulmonary Edema 88% Pulmonary Vascular Congestion 91% No Evidence of Pneumonia 91% No Change from Previous/No Interval Change 91% Normal/No Abnormalities 91% Chest CT: Percent agreement between abstractor and auditor for critical data elements Air Space Density/Opacity/Disease 92% Aspiration 92% Aspiration Pneumonia 92% Bronchopneumonia 92% Cannot Rule Out Pneumonia 92% Cavitation 92% Consolidation 92% Ground Glass 92% Infection (cannot rule out infection, likely infection) 92% Infiltrate (Single Lobe) 88% Infiltrate (Multiple Lobes) 92% Interstitial lung disease/interstitial disease 92% Interval improvement or resolution 92% Loculations 92% Mass 92% Necrotizing Pneumonia 92% Neoplasm/Metastatic Disease/Malignancy 92% New or Worsening Infiltrates 92% Nodular Airspace Disease 92% Nodules 92% Pleural Effusion 92% Pneumonia 83% Pneumonitis 92% Post Obstructive Pneumonia 92% Pulmonary Edema 92% Pulmonary Vascular Congestion 92% No Evidence of Pneumonia 92% No Change from Previous/No Interval Change 92% Normal/No Abnormalities 92% Signs/Symptoms: Percent agreement between abstractor and auditor for critical data elements New or Increasing Cough 98% New or Increasing Dyspnea/Shortness of Breath 88% Increased/Changed Secretions or Sputum Production 92% Chills 96% Rales 94% Bronchial Breath Sounds 100% Rhonchi 86% Dullness on Percussion 100% Crackles 90% Tachypnea 90% Leukocytosis 100% Abnormal Temperature 91% Hypoxemia 93% Leukopenia 100%

Comment 2 by: Submitted by Valerie Vaughn, Michigan Hospital Medicine Safety Consortium (*#3671 measure developer)*

This public comment is to address concerns about reliability testing at the accountable entitle level. There are concerns that our ICC appears low (0.0525). We would like to clarity that the ICC of 0.0525 applies only if a single case were obtained from each hospital. This indicates that if each hospital performed 1 case abstraction, there would be high variability and poor reliability. However, we do not suggest each hospital only conduct 1 case abstraction. The Spearman Brown Prophecy provides an estimation of reliability after adjusting the number of measurements. When the median number of case counts for the entire cohort (N=184 median cases in measure development hospitals) is applied to the Spearman Brown formula, the overall reliability was 0.911 (well above the 0.5 threshold noted for “poor reliability”). The 0.911 was calculated as follows: Median case abstractions: 184 (IQR 153-201) Reliability or ICC for 184 cases (i.e., ICC/reliability for a typical HMS hospital): (184*0.0525169)/(1+(184-1)*0.0525169)=0.911 Through this same calculation, using the Spearman Brown Prophecy, we calculated the 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 Thus, we attain reliability of 0.8 (standard reliability for a quality metric of this stakes) with 73 cases per hospital which is our suggested target number of cases for the measure.

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? ☒ Yes ☐ No**
2. **Briefly summarize any changes to the measure specifications and/or concerns about the measure specifications.**
   - No concerns.

RELIABILITY: TESTING

3. **Reliability testing level: ☒ Accountable-Entity Level ☒ Patient/Encounter Level ☐ Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure:**

   ☒ Yes ☐ No
5. If accountable-entity level and/or patient/encounter level reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☐ Yes ☐ No
6. **Assess the method(s) used for reliability testing:**
   - Reliability testing was conducted using data from the Michigan Hospital Medicine Safety Consortium (HMS) from 07/01/2017 – 03/31/2020
   - The developer stated they conducted reliability testing at both the patient/encounter level and accountable/entity level
   - Accountable Entity Level Testing
     - The developer performed a signal-to-noise analysis, then calculated hospital variance (signal), within hospital variance (noise) and total variance, which were used to calculate the intraclass correlation coefficient (ICC).
     - The ICC was used in the Spearman Brown formula to determine reliability for the entire hospital cohort using the median number of case abstracts and to determine the minimum case abstracts needed to achieve predetermined reliability thresholds (0.6, 0.7, 0.8, and 0.9)..
   - The developer stated that all critical data elements were tested in validity testing but only provided an overall score. In NQF's assessment this does not show validity of all critical data elements and thus cannot be used to demonstrate reliability.
7. **Assess the results of reliability testing**

   Accountable Entity Level
   - Testing was conducted using data from 49 hospitals in the Michigan Hospital Medicine Safety Consortium (HMS) from July 2017-March 2020. This dataset contained 82% academic hospitals, 82% metropolitan, 92% non-profit, and 69% hospitals with greater than 200 staffed beds.
     - This dataset represents 18,625 hospitalized patients treated for CAP in this time period, all of which were included in reliability and validity testing.
   - The developer performed a signal-to-noise analysis using a mixed-effect logistic model run as an empty model to calculate hospital variance (signal), within hospital variance (noise), and total variance.
     - Hospital variance: 0.18235
     - Within hospital variance: 3.28987
     - Total variance: 3.4722

- The intraclass correlation coefficient (ICC) was calculated from this analysis (ICC = 0.0525) and used in a Spearman Brown formula to calculate reliability for the entire hospital cohort and to determine the minimum number of cases needed to achieve specific reliability thresholds.
    - The developer states the overall reliability for the cohort = 0.911 and is considered to be strong reliability and meets the threshold for reliability for measures considered to be high stakes (>0.9)
    - Number of cases needed to achieve set reliability thresholds:
        - Minimum 28 cases for reliability of 0.6
        - Minimum 43 cases for reliability of 0.7
        - Minimum 73 cases for reliability of 0.8
        - Minimum 163 cases for reliability of 0.9
    - Within the testing cohort, all but one hospital were able to abstract the minimum 43 cases/year needed to reach 0.7 reliability. 92.5% of hospitals in the cohort were able to abstract 73 or more cases to achieve 0.8 reliability.

8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? **NOTE:** If multiple methods used, at least one must be appropriate.

   ☒ **Yes** ☒ **No** ☐ **Not applicable**

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   ☐ **Yes** ☒ **No** ☐ **Not applicable** (patient/encounter level testing was not performed)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and all testing results):

    ☐ **High** (NOTE: Can be HIGH only if accountable-entity level testing has been conducted)

    ☐ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity level testing has not been conducted)

    ☐ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☒ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)

11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

Specifications are precise and unambiguous (Box 1) -> Reliability was conducted with the measure as specified (Box 2) -> Reliability testing conducted at the accountable entity level (Box 4) -> Signal-to-noise method used to determine reliability but unclear method used for calculating median number of case abstracts and ICC (Box 5) -> Unclear if empirical testing conducted on all critical data elements (Box 8) -> Rate as INSUFFICIENT

In signal-to-noise analysis, the internal variance is greater than the external variance and the intraclass correlation coefficient is well below 0.5, a range generally agreed to show poor reliability. It is not clear from the submission how applying the Spearman Brown prophecy formula leads to an overall reliability of 0.9. Additionally, only an overall score was provided for patient/encounter level reliability testing thus making it difficult to determine if all critical data elements were evaluated

VALIDITY: TESTING

12. **Validity testing level:** ☐ **Measure score** ☐ **Data element** ☒ **Both**

13. **If patient/encounter level validity testing was provided, was the method described and appropriate for assessing the accuracy of ALL critical data elements? NOTE:** Data element validation from the literature is acceptable.

☐ **Yes**

☒ **No**

☐ **Not applicable** (patient/encounter level testing was not performed)

14. **Method of establishing validity of the measure score:**

☐ **Face validity**

☒ **Empirical validity testing of the measure score**

☐ **N/A (score-level testing not conducted)**

15. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☐ **No**

☐ **Not applicable** (score-level testing was not performed)

16. **Assess the method(s) for establishing validity**

Patient/Encounter-level Validity

Data Audit

- The developer asked the senior HMS project manager to perform blind audits of 50 consecutive cases of patients counted as inappropriately diagnosed with CAP. This data came from 33 of the hospitals in the cohort using 2021 data.
- The proportion of data elements abstracted correctly was calculated. "Correct data, as abstracted by the HMS project manager, were then reapplied to the measure definition to assess for changes in case classification."

Structured Implicit Case Review

- Using 2020 data, cases were randomly selected from the "gray areas" identified during measure development (e.g., patients with atelectasis as the only finding on chest imaging) and 2-3 physicians reviewed these to confirm accurate case categorization. If there was disagreement in classification, the developer prompted a discussion about ways to improve the measure to account for errors in classification.

Accountable Entity Level Validity

Empirical Validity of Measure Score:

- The developer conducted empirical validity testing by correlating NQF #3671 with NQF# 3690 *Inappropriate Diagnosis of Urinary Tract Infection (UTI)*. The developer states that there were very few measures that assess the same domains of quality as NQF #3671, so after conducting a literature review they found that the inappropriate diagnosis of CAP can represent a signal of hospital quality, which affects patient outcomes.
- They also assessed the association of NQF #3671 with antibiotic-associated adverse events.

Face Validity: Technical Expert Panel Feedback

- 14 national experts, including infectious disease physicians, pharmacists, pulmonologists, radiologists, hospitalists, emergency medicine physicians, regulatory agencies, and individuals with experience in quality improvement, performance measurement, diagnostic error, antibiotic stewardship, and healthcare quality participated in two weeks of online conversations and responded to survey questions about the measure.
    - TEP experts responded to the following:
    - a. How much do you agree/disagree with the following statement? "The inappropriate diagnosis of CAP measure as specified can be used to distinguish between better and

worse quality hospitals." 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree.

- b. Are there any key data elements you believe are missed or not accurately captured in the inappropriate diagnosis of CAP measure?

17. **Assess the results(s) for establishing validity**

Patient/Encounter-level Validity

Data Audit

- Data audit found 93.7% of data elements were abstracted correctly. The developer states that any discrepancies found were minor and resulted in no changes to case classification.
- The classification of the 50 cases as "CAP" or "inappropriate diagnoses of CAP" by first the abstractor and then the auditor had 100% agreement.

Structured Implicit Case Review

- Case review resulted in K=0.72 agreement between physician reviewers on case classification, which the developer considers to be "substantial agreement."
  - Since the case review involved "gray area" cases rather than a random selection, the developer states the true K may be even higher.
- In 94% of cases (16/17), there was 100% agreement that the cases represented inappropriate diagnosis.

Accountable Entity Validity

Empirical Validity of the Measure Score

- The developer found that NQF #3671 is moderately correlated with NQF# 3690 (R=0.53, p<0.001). The findings were similar, though slightly less strong, for patients inappropriately diagnosed with either condition in emergency department (ED) settings (R=0.46, p<0.002).
  - The developer states that this shows that inappropriate diagnosis of the CAP measure may reflect the overall quality of diagnoses made at a hospital.
- The developer found that each additional day of antibiotic use in patients inappropriately diagnosed with CAP was associated with an increased odds ratio (1.05) for developing a patient-report antibiotic-associated adverse event.
  - As inappropriate diagnoses of CAP decreased over time, related outcomes improved in HMS hospitals.
    - Death events fell from 3.5% (2017 data, n=6405) to 2.9% (2020 data, n=4961)
    - Adverse-Antibiotic Events fell from 4.8% (2017 data, n=6405), to 3.0% (2020 data, n=4961)

Face Validity

- 57% of TEP members (8/14) stated they do not think the measure needs to collect any additional data in order to correctly identify inappropriate diagnosis of CAP.
- The remaining feedback was incorporated into the measure:
  - Duration of Treatment: data on duration of treatment for patients inappropriately diagnosed with CAP added to submission
  - A Balancing Measure: studies of underdiagnosis added to Evidence section
  - The trend in outcomes of denominator over time as inappropriate diagnosis decreases: data on outcomes over time added to measure
  - Patients 80 years and older with only 1 sign or symptom: data added on those over age 80

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

18. **Please describe any concerns you have with measure exclusions.**
    - The developer lists several exclusions to the measure (patients who left against medical advice or refused care, who were pregnant or breastfeeding, who were admitted on hospice or comfort care, who had cystic fibrosis, or who had a pneumonia-related complication) and states the exclusions were determined through careful clinical review and feedback from experts. Exclusions were not common and represented approximately 2.68%-3.35% of the testing population.
19. **Risk Adjustment**

    19a. **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

    19b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

    ☐ Yes ☐ No ☒ Not applicable

    19c. **Social risk adjustment:**

    19c.1 Are social risk factors included in risk model? ☐ Yes ☐ No ☒ Not applicable

    19c.2 Conceptual rationale for social risk factors included? ☐ Yes ☐ No

    19c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐ Yes ☐ No

    19d.**Risk adjustment summary:**

    19d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

    19d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

    19d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☐ No

    19d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No

    19d.5.Appropriate risk-adjustment strategy included in the measure? ☐ Yes ☐ No

    19e**. Assess the risk-adjustment approach**

    - N/A
20. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

    For cost/resource use measures, does this measure identify meaningful differences about cost and resource use between the measured entities?

    - The developer requested all hospitals in the cohort report the distribution of their measure scores then grouped hospitals by quartiles to determine whether the difference in mean measure score was statistically significant.
        - Hospitals in the 10th percentile (better performance) have about 7 fewer patients inappropriately diagnosed with CAP per 100 CAP discharges than the median.
        - Hospitals in the 90th percentile (worse performance) have approximately 10 more patients inappropriately diagnosed with CAP per 100 CAP discharges than the median.
        - The difference in performance between all adjacent quartiles (1st vs. 2nd, 2nd vs. 3rd, 3rd vs. 4th) was statistically significant ($p<0.001$ for all comparisons).
21. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    - N/A
22. **Please describe any concerns you have regarding missing data.**

- The developer found missing data to be extremely rare; the percentage of patients in the testing cohort with "unknown" or "not available" values was less than 1.0% (183/18,468 patients). For cost/resource use measures ONLY:

If not cost/resource use measure, please skip to question 25.

23. **Are the specifications in alignment with the stated measure intent?**

    ☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**

24. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

25. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

    ☐ **High** (NOTE: Can be HIGH only if accountable-entity level testing has been conducted)

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity level testing has NOT been conducted)

    ☐ **Low** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the accountable-entity level and the patient/encounter level is required; if not conducted, should rate as INSUFFICIENT.)

26. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

    Threats to validity empirically assessed (Box 1) → Testing conducted using the measure as specified (Box 2) → Testing conducted at the accountable entity level (Box 5) → Method was appropriate (Box 6) → Based on the testing results there is moderate certainty that the accountable entity data are a valid indicator of quality (Box 7b) → Rate as MODERATE

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

    - [Summary]

ADDITIONAL RECOMMENDATIONS

29. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

    - [Summary]

Developer Submission

Criteria 1: Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. Measures must be judged to meet all sub criteria to pass this criterion and be evaluated against the remaining criteria

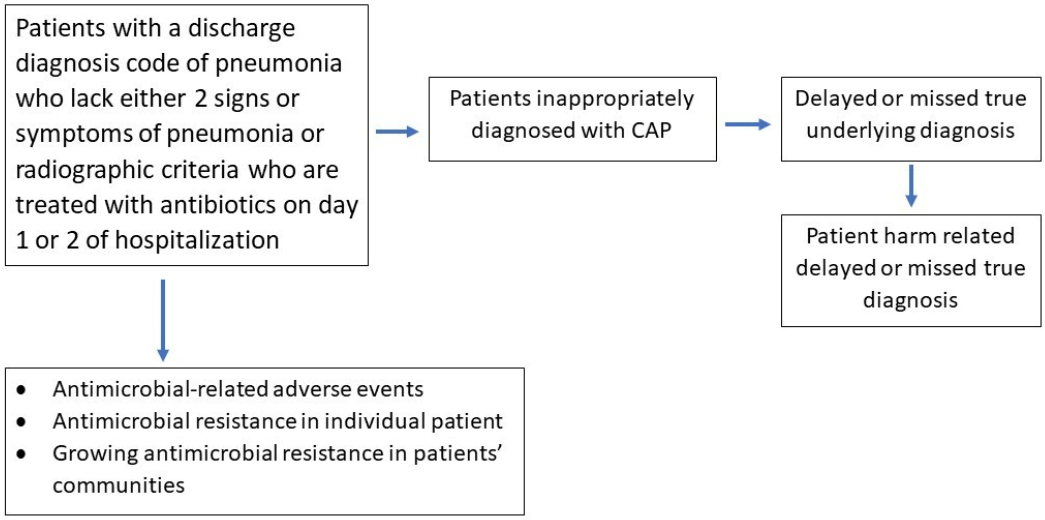
---

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Measure and Report: Evidence section. For example:
2021 Submission:
Updated evidence information here.
2018 Submission:
Evidence from the previous submission here.

1a.01. Provide a logic model.

Briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

[Response Begins]
Figure 1. Logic model for Inappropriate Diagnosis of Community Acquired Pneumonia (CAP) measure



Patients with a discharge diagnosis code of pneumonia who lack either 2 signs or symptoms or radiographic criteria may be inappropriately diagnosed with CAP which in turn may result in delayed or missed true diagnosis which in turn leads to patient harm. Treatment of patients with a discharge diagnosis code of pneumonia who lack either 2 signs or symptoms or radiographic criteria can lead to antimicrobial-related adverse events and antimicrobial resistance in individuals and communities.

[Response Ends]

1a.02. Select the type of source for the systematic review of the body of evidence that supports the performance measure.

A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data.

[Response Begins]

Other (specify)

[Other (specify) Please Explain]

Our definition of inappropriate diagnosis of CAP is based on treatment for CAP in the absence of meeting clinical or radiographic criteria for CAP. Our criteria for CAP are similar to the National Healthcare Safety Network (NHSN) criteria for Clinically Defined Pneumonia.

[Response Ends]

If the evidence is not based on a systematic review, skip to the end of the section and do not complete the repeatable question group below. If you wish to include more than one systematic review, add additional tables by clicking "Add" after the final question in the group.

Evidence - Systematic Reviews Table (Repeatable)

Group 1 - Evidence - Systematic Reviews Table

1a.03. Provide the title, author, date, citation (including page number) and URL for the systematic review.

[Response Begins]

N/A

[Response Ends]

1a.04. Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the systematic review.

[Response Begins]

N/A

[Response Ends]

1a.05. Provide the grade assigned to the evidence associated with the recommendation, and include the definition of the grade.

[Response Begins]

N/A

[Response Ends]

1a.06. Provide all other grades and definitions from the evidence grading system.

[Response Begins]

N/A

[Response Ends]

1a.07. Provide the grade assigned to the recommendation, with definition of the grade.

[Response Begins]

N/A

[Response Ends]

1a.08. Provide all other grades and definitions from the recommendation grading system.

[Response Begins]
N/A
[Response Ends]

1a.09. Detail the quantity (how many studies) and quality (the type of studies) of the evidence.

[Response Begins]
N/A
[Response Ends]

1a.10. Provide the estimates of benefit, and consistency across studies.

[Response Begins]
N/A
[Response Ends]

1a.11. Indicate what, if any, harms were identified in the study.

[Response Begins]
N/A
[Response Ends]

1a.12. Identify any new studies conducted since the systematic review, and indicate whether the new studies change the conclusions from the systematic review.

[Response Begins]
N/A
[Response Ends]

1a.13. If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, describe the evidence on which you are basing the performance measure.

[Response Begins]
Our definition of inappropriate diagnosis of CAP is based on treatment for CAP in the absence of meeting clinical or radiographic criteria for CAP. Our criteria for CAP are similar to the National Healthcare Safety Network (NHSN) criteria for Clinically Defined Pneumonia.[1] While the NHSN algorithms are designed to detect hospital-acquired infections, we have relaxed the requirements for diagnosis of pneumonia such that our measure is less restrictive in its diagnostic criteria. Ultimately, however, the necessity of both radiographic and clinical findings exists in both the NSHN diagnostic criteria and the submitted measure criteria.
Table 1. NHSN criteria for clinically defined pneumonia compared to definition of the inappropriate diagnosis of CAP measure.

NHSN Criteria for Imaging Test Evidence	NHSN Criteria for Signs/Symptoms	Inappropriate Diagnosis of CAP Measure
Two or more serial chest imaging tests results with at least **one** of the following: New and persistent *Or* Progressive and persistent	For ANY PATIENT, at least **one** of the following: • Fever (>38C or >100.4F) • Leukopenia ($\leq$4000 WBC/mm^3) or leukocytosis ($\geq$12,000 WBC/mm^3)	Similar to the NHSN guidelines, appropriate diagnosis of CAP in the submitted measure requires both imaging evidence and presence of signs or symptoms. Whereas the NHSN guidelines require two or more serial abnormal imaging tests, **the submitted**

NHSN Criteria for Imaging Test Evidence	NHSN Criteria for Signs/Symptoms	Inappropriate Diagnosis of CAP Measure
• Infiltrate • Consolidation • Cavitation **Note**: In patients *without* underlying pulmonary or cardiac disease, one definitive imaging test result is acceptable	• For adults ≥ 70 years old, altered mental status with no other recognized cause And at least **two** of the following: • New onset of purulent sputum or change in character of sputum, or increased respiratory secretions, or increased suctioning requirements • New onset or worsening cough, dyspnea, or tachypnea • Rales or bronchial breath sounds • Worsening gas exchange (e.g. O_2 desaturations, increased oxygen requirement)	**measure requires only a single abnormal imaging test** (e.g., chest radiography or computed tomography). Additionally, while the NHSN requires presence of at least one abnormality in temperature, white blood cell count, or mentation and two addition signs or symptoms, **the submitted measure requires only the presence of two *total* abnormal signs or symptoms**.

NHSN criteria for clinically defined pneumonia are compared to the definition of the inappropriate diagnosis of CAP measure. The inappropriate diagnosis of CAP measure requires only a single abnormal imaging test and the presence of two *total* abnormal signs or symptoms.

The necessity for both abnormal radiography and presence of signs and symptoms of pneumonia is highlighted in the American Thoracic Society and Infectious Disease Society of America's clinical practice guideline "Diagnosis and Treatment of Adults with Community-acquired Pneumonia", in which they state that the evidence used to guide recommendations "focused on studies that used radiographic criteria for defining CAP, given the known inaccuracy of clinical signs and symptoms alone for CAP diagnosis."[2] The guideline cites a systematic review evaluating the predictive value of history and physical exam to diagnose pneumonia, in which the authors conclude "There are no combination of history and physical examination findings that confirm the diagnosis of pneumonia. If diagnostic certainty is required in the management of a patient with suspected pneumonia, then chest radiography should be performed."[3]

[Response Ends]

1a.14. Briefly synthesize the evidence that supports the measure.

[Response Begins]

The consequences of inappropriate diagnosis of CAP (which by our measure includes treatment with antibiotics) are considerable and include harm from missing or delayed true diagnoses and harms related to antibiotic therapy. The most recent data from the National Hospital Ambulatory Medical Care Survey by the Centers for Disease Control and Prevention estimated there are nearly 1.5 million annual emergency department (ED) visits for pneumonia each year, among the most of any infectious etiology.[4] While underdiagnosis of CAP occurs, there is a significant body of evidence suggesting that CAP is often inappropriately diagnosed, especially in the ED setting. For example, one study of patients admitted from the ED with a diagnosis of CAP found that 27.3% (95% CI 24-31%) ultimately had a non-pneumonia diagnosis on discharge.[5] A similar study found that within a cohort of consecutively admitted patients, pneumonia was the most common ED admission diagnosis; however, authors noted similar discordance (29%) between the ED admission pneumonia diagnosis and the ultimate ward discharge diagnosis.[6]

Harms related to missed or delayed diagnoses

Inappropriate diagnosis of CAP is not without harm. It often leads to "anchoring" or "premature closure" where further diagnoses are not entertained. Thus, the true underlying diagnosis may be missed, or appropriate care may be delayed. These misses or delays are not benign. For example, in one study of 40,744 patients admitted to Department of Veterans Affairs medical centers between 2002-2007 with a diagnosis of pneumonia, 9.2% were diagnosed with a pulmonary malignancy after their index pneumonia admission. In that study, the median time to diagnosis was 297 days, with only 27% diagnosed with 90-days of admission.[7] Delay of lung cancer diagnosis of this kind often means the cancer has morphed from local, curable disease to metastatic, incurable disease.

Another diagnosis frequently missed in patients inappropriately diagnosed with CAP is acute decompensated heart failure. In one study of patients hospitalized with acute decompensated heart failure, those treated with antibiotics without definitive infection experienced longer lengths of hospital stay (3.0 vs 6.6 days, $P<0.001$) and had over twice the rates of hospital readmission within 30 days.[8]

Additional data suggest that more aggressive diagnosis and treatment of CAP within the ED may result in less optimal patient care. For example, following a since-revised 2003 Infectious Disease Society of America guideline for CAP

recommending initiation of antibiotics within 4 hours, patients were noted to more frequently have a hospital admission diagnosis of CAP without radiographic evidence (28.5% in 2005 after guidelines implementation vs 20.6% in 2003 prior to guideline implementation, p=0.04), more antibiotic utilization per patient, and to less often have a final discharge diagnosis of CAP (58.9% in 2005 vs 75.9% in 2003, P<0.001). There were no noted differences in mortality between the groups.[9]

Prevalence of Inappropriate Antibiotic Use and Antibiotic Associated Adverse Events

Patients inappropriately diagnosed with CAP also experience unnecessary antibiotic use and its associated harm. While inappropriate diagnosis of CAP and inappropriate treatment for CAP often occur in the ED, the use of inappropriate antibiotics generally continues during hospitalization. In a subset of patients from our dataset who were inappropriately diagnosed with CAP within the ED, 76.1% (n=885/1163) were started on antibiotics by an emergency medicine clinician.[10] Of those, 89.9% (n=796/885) remained on antibiotics on day 3 of hospitalization. Antibiotic overuse, and its association with patient harm, is well established.[11-16] One study estimated the incidence of overall antibiotic-associated adverse drug events in hospitalized patients receiving systemic therapy was up to 20%.[12] Similarly, another study showed each day of excess antibiotic treatment for pneumonia was associated with a 5% increase in the odds of patient-reported antibiotic-associated adverse events.[14] These events have implications for both patients and hospital systems, as one study found that antibiotics were implicated in nearly 20% of ED visits for drug-related adverse events.[17] Finally, antibiotic administration has been strongly associated with development of *Clostridioides difficile* infection (CDI),[18,19] while antibiotic stewardship programs have been shown to reduce CDI.[20, 21]

Antibiotic Use and Antimicrobial Resistance

Antibiotic use in patients inappropriately diagnosed with infections continues to be a large driver of antibiotic use and antibiotic resistance. Between 2012 and 2017, overall antibiotic days of therapy in US hospitals were unchanged.[22] While the prevalence of some multi-drug resistant bacteria decreased over that time period (e.g., methicillin-resistant staphylococcus aureus (MRSA)), other highly concerning multi-drug resistant organisms flourished. For example, the incidence of infections resulting from extended-spectrum beta-lactamase (ESBL) producing organisms increased by 53.3% (from 37.55 to 57.12 cases per 10,000 hospitalizations).[23] A systematic review and meta-analysis of the literature found a significant positive relationship between antibiotic consumption and development of antimicrobial resistance, with a pooled odds ratio of 2.3 (95% confidence interval 2.2-2.5).[24] Similarly, a recent study found that recent antibiotic exposure was positively associated with baseline multi-drug resistant colonization (odds ratio 1.70; 95% confidence interval 1.22-2.38).[25] The implications of these antimicrobial resistant organisms are significant. Globally, predictive statistical models estimate 4.95 million (3.62-6.57 million) deaths associated with bacterial antimicrobial resistance in 2019, of which 1.27 million (95% uncertainly interval 0.911-1.71) deaths were attributable to bacterial antimicrobial resistance.[26] In a study of inpatient admissions in the US Department of Veterans Affairs between October 2007 and November 2010, healthcare-associated infections (HAI) with multi-drug resistant gram negative bacteria were associated with a significantly elevated risk of mortality as was HAI or colonization with MRSA.[27]

[Response Ends]

1a.15. Detail the process used to identify the evidence.

[Response Begins]

Evidence was identified through comprehensive Pubmed search of studies as they pertain to diagnosis (as well as over-diagnosis, inappropriate diagnosis, misdiagnosis) of CAP, treatment of CAP, antibiotic side effects, and antimicrobial resistance.

[Response Ends]

1a.16. Provide the citation(s) for the evidence.

[Response Begins]

[1]Pneumonia (Ventilator-associated [VAP] and non-ventilator-associated Pneumonia [PNEU]) Event. National Healthcare Safety Network. Centers for Disease Control. January 2022. < https://www.cdc.gov/nhsn/pdfs/pscmanual/6pscvapcurrent.pdf >

[2] Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K, Cooley LA, Dean NC, Fine MJ, Flanders SA, Griffin MR, Metersky ML, Musher DM, Restrepo MI, Whitney CG. Diagnosis and Treatment of Adults with Community-acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America. Am J Respir Crit Care Med. 2019 Oct 1;200(7):e45-e67.

[3] Metlay JP, Kapoor WN, Fine MJ. Does this patient have communityacquired pneumonia? Diagnosing pneumonia by history and physical examination. JAMA 1997;278:1440–1445.
[4] National Center for Health Statistics. National Hospital Ambulatory Medical Care Survey, 2018. < https://www.cdc.gov/nchs/data/nhamcs/web_tables/2018-ed-web-tables-508.pdf > Accessed 11 March 2022.
[5]Chandra A, Nicks B, Maniago E, Nouh A, Limkakeng A. A multicenter analysis of the ED diagnosis of pneumonia. Am J Emerg Med. 2010 Oct;28(8):862-5.
[6]Atamna A, Shiber S, Yassin M, Drescher MJ, Bishara J. The accuracy of a diagnosis of pneumonia in the emergency department. Int J Infect Dis. 2019 Dec;89:62-65.
[7]Mortensen EM, Copeland LA, Pugh MJ, et al. Diagnosis of pulmonary malignancy after hospitalization for pneumonia. *Am J Med*. 2010;123(1):66-71. doi:10.1016/j.amjmed.2009.08.009
[8]Frisbee J, Heidel RE, Rasnake MS. Adverse Outcomes Associated With Potentially Inappropriate Antibiotic Use in Heart Failure Admissions. Open Forum Infect Dis. 2019 May 8;6(6):ofz220.
[9]Kanwar M, Brar N, Khatib R, Fakih MG. Misdiagnosis of community-acquired pneumonia and inappropriate utilization of antibiotics: side effects of the 4-h antibiotic administration rule. Chest. 2007 Jun;131(6):1865-9.
[10] Gupta A, Petty L, Gandhi T, Flanders S, Hsaiky L, Basu T, Zhang Q, Horowitz J, Masood Z, Chopra V, Vaughn VM. Overdiagnosis of urinary tract infection linked to overdiagnosis of pneumonia: a multihospital cohort study. BMJ Qual Saf. 2022 Jan 5:bmjqs-2021-013565.
[11] Fridkin S, Baggs J, Fagan R, et al. Vital signs: improving antibiotic use among hospitalized patients. MMWR Morbidity and mortality weekly report. 2014;63(9):194-200.
[12] Tamma PD, Avdic E, Li DX, Dzintars K, Cosgrove SE. Association of Adverse Events With Antibiotic Use in Hospitalized Patients. JAMA Intern Med. 2017 Sep 1;177(9):1308-1315
[13] Madaras-Kelly KJ, Burk M, Caplinger C, Bohan JG, Neuhauser MM, Goetz MB, Zhang R, Cunningham FE; Pneumonia Duration of Therapy Medication Utilization Evaluation Group. Total duration of antimicrobial therapy in veterans hospitalized with uncomplicated pneumonia: Results of a national medication utilization evaluation. J Hosp Med. 2016 Dec;11(12):832-839.
[14] Vaughn VM, Flanders SA, Snyder A, Conlon A, Rogers MAM, Malani AN, McLaughlin E, Bloemers S, Srinivasan A, Nagel J, Kaatz S, Osterholzer D, Thyagarajan R, Hsaiky L, Chopra V, Gandhi TN. 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.
[15] Werner NL, Hecker MT, Sethi AK, Donskey CJ. Unnecessary use of fluoroquinolone antibiotics in hospitalized patients. BMC infectious diseases. 2011;11:187.
[16] Hecker MT, Aron DC, Patel NP, Lehmann MK, Donskey CJ. Unnecessary use of antimicrobials in hospitalized patients: Current patterns of misuse with an emphasis on the antianaerobic spectrum of activity. Archives of internal medicine. 2003;163(8):972-978.
[17] Shehab N, Patel PR, Srinivasan A, Budnitz DS. Emergency department visits for antibiotic-associated adverse events. Clin Infect Dis. 2008 Sep 15;47(6):735-43.
[18] Vardakas KZ, Trigkidis KK, Boukouvala E, Falagas ME. Clostridium difficile infection following systemic antibiotic administration in randomised controlled trials: a systematic review and meta-analysis. Int J Antimicrob Agents. 2016 Jul;48(1):1-10.
[19] Slimings C, Riley TV. Antibiotics and healthcare facility-associated Clostridioides difficile infection: systematic review and meta-analysis 2020 update. J Antimicrob Chemother. 2021 Jun 18;76(7):1676-1688.
[20] Dingle KE, Didelot X, Quan TP, Eyre DW, Stoesser N, Golubchik T, Harding RM, Wilson DJ, Griffiths D, Vaughan A, Finney JM, Wyllie DH, Oakley SJ, Fawley WN, Freeman J, Morris K, Martin J, Howard P, Gorbach S, Goldstein EJC, Citron DM, Hopkins S, Hope R, Johnson AP, Wilcox MH, Peto TEA, Walker AS, Crook DW; Modernising Medical Microbiology Informatics Group. Effects of control interventions on Clostridium difficile infection in England: an observational study. Lancet Infect Dis. 2017 Apr;17(4):411-421.
[21] Baur D, Gladstone BP, Burkert F, Carrara E, Foschi F, Döbele S, Tacconelli E. Effect of antibiotic stewardship on the incidence of infection and colonisation with antibiotic-resistant bacteria and Clostridium difficile infection: a systematic review and meta-analysis. Lancet Infect Dis. 2017 Sep;17(9):990-1001.
[22] James Baggs, PhD, Sophia Kazakova, MD, MPH, PhD, Kelly M Hatfield, MSPH, Sujan Reddy, MD, MSc, Arjun Srinivasan, MD, Lauri Hicks, DO, Melinda M Neuhauser, PharmD, MPH, John A Jernigan, MD, MS, 2891. Trends in Inpatient Antibiotic Use in US Hospitals, 2012–2017, *Open Forum Infectious Diseases*, Volume 6, Issue Supplement_2, October 2019, Page S79,
[23] Jernigan JA, Hatfield KM, Wolford H, Nelson RE, Olubajo B, Reddy SC, McCarthy N, Paul P, McDonald LC, Kallen A, Fiore A, Craig M, Baggs J. Multidrug-Resistant Bacterial Infections in U.S. Hospitalized Patients, 2012-2017. N Engl J Med. 2020 Apr 2;382(14):1309-1319.

[24] Bell BG, Schellevis F, Stobberingh E, Goossens H, Pringle M. A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. BMC Infect Dis. 2014 Jan 9;14:13.
[25] Gontjes KJ, Gibson KE, Lansing BJ, Mantey J, Jones KM, Cassone M, Wang J, Mills JP, Mody L, Patel PK. Association of Exposure to High-risk Antibiotics in Acute Care Hospitals With Multidrug-Resistant Organism Burden in Nursing Homes. JAMA Netw Open. 2022 Feb 1;5(2):e2144959.
[26]Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet. 2022 Feb 12;399(10325):629-655.
[27] Nelson RE, Slayton RB, Stevens VW, Jones MM, Khader K, Rubin MA, Jernigan JA, Samore MH. Attributable Mortality of Healthcare-Associated Infections Due to Multidrug-Resistant Gram-Negative Bacteria and Methicillin-Resistant Staphylococcus Aureus. Infect Control Hosp Epidemiol. 2017 Jul;38(7):848-856.

[Response Ends]

1b.01. Briefly explain the rationale for this measure.

Explain how the measure will improve the quality of care, and list the benefits or improvements in quality envisioned by use of this measure.

[Response Begins]
The goal of this measure is to improve diagnosis and treatment of pneumonia. Literature has demonstrated that while pneumonia is the most common infectious etiology for which patients are hospitalized, it is often inappropriately diagnosed, resulting in unnecessary antibiotic administration and delay in diagnosis of true underlying conditions. The implications of unnecessary antibiotics are well described and include risks of antibiotic-associated adverse events such as *Clostridioides difficile* infection, prolonged length of hospital stay, and antimicrobial resistance, all of which can increase patient morbidity and mortality. Missed or delayed diagnosis of a true underlying condition are equally troubling, as data suggest that diagnostic error results in the highest morbidity, mortality, and malpractice cost of any medical error. Through adoption of this measure, we anticipate a decrease in inappropriate diagnosis of pneumonia, a decrease in unnecessary antibiotic use, and improved patient outcomes.
[Response Ends]

1b.02. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.

Include mean, std dev, min, max, interquartile range, and scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]
Data below are from 7/1/2017-3/31/2020 across 49 acute care hospitals in the state of Michigan. This includes 18,625 patients treated for CAP, of whom 12.3% (2,299) were inappropriately diagnosed with CAP. For data of scores over time, we do not report 2020 data as it only includes a single quarter (ending 3/31/2020).
Here, we divided all 49 hospitals each year into performance deciles with decile 1 representing the top performing hospitals. Scores or the percentage of patients treated for pneumonia who were considered inappropriately diagnosed with CAP are then reported by decile, first giving mean (standard deviation [SD]) then providing median (inter-quartile range [IQR]) data.
Table 1. Mean (SD) percent of cases inappropriately diagnosed with CAP (i.e., "score") by Year; N=49 hospitals

Decile	2017; mean (SD)	2018; mean (SD)	2019; mean (SD)
1 (best performing)	5.6 (1.1)	3.9 (1.3)	4.5 (1.0)
2	7.7 (0.6)	5.5 (0.2)	6.3 (0.8)
3	9.6 (1.4)	7.9 (1.0)	8.2 (0.5)
4	12.7 (0.4)	9.2 (0.3)	9.9 (0.8)
5	14.0 (0.5)	10.2 (0.5)	11.4 (0.3)
6	14.9 (0.2)	11.4 (0.3)	12.1 (0.1)
7	15.8 (0.5)	12.7 (0.7)	13.0 (0.3)
8	17.9 (1.6)	14.6 (0.7)	14.5 (0.8)

Decile	2017; mean (SD)	2018; mean (SD)	2019; mean (SD)
9	22.4 (1.0)	17.5 (1.4)	18.1 (1.1)
10 (worst performing)	26.8 (3.4)	21.4 (2.4)	22.4 (2.9)

Mean (SD) percent of cases inappropriately diagnosed with CAP trended downward from 2017 to 2019 in all deciles.

*2020 includes only 1 quarter of data and thus is not reported in the time trend above.

Table 2. Median (IQR) percent of cases inappropriately diagnosed with CAP (i.e., "score") by Year; N=49 hospitals

Decile	2017; median (IQR)	2018; median (IQR)	2019; median (IQR)
1 (best performing)	6.1 (5.0, 6.2)	4.2 (3.0, 4.8)	4.0 (3.6, 5.3)
2	7.3 (7.3, 8.3)	5.4 (5.3, 5.6)	6.0 (5.7, 6.9)
3	8.6 (8.6, 10.9)	8.3 (7.3, 8.5)	8.2 (7.8, 8.5)
4	12.9 (12.6, 13.0)	9.2 (8.9, 9.3)	9.8 (9.4, 10.4)
5	13.8 (13.6, 14.4)	10.0 (9.9, 10.5)	11.6 (11.3, 11.6)
6	14.9 (14.8, 15.0)	11.4 (11.1, 11.6)	12.1 (12.0, 12.2)
7	15.5 (15.5, 16.3)	12.6 (12.4, 13.2)	13.1 (12.8, 13.2)
8	17.6 (16.4, 19.1)	14.3 (14.1, 15.0)	14.4 (13.9, 15.1)
9	22.7 (22.7, 22.8)	17.7 (16.3, 18.7)	17.9 (17.3, 19.0)
10 (worst performing)	27.8 (23.4, 28.0)	21.1 (19.8, 21.1)	21.2 (20.5, 24.3)

Median (IQR) percent of cases inappropriately diagnosed with CAP trended downward from 2017 to 2019 in all deciles.

*2020 includes only 1 quarter of data and thus is not reported in the time trend above.

[Response Ends]

1b.03. If no or limited performance data on the measure as specified is reported above, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement. Include citations.

[Response Begins]
N/A
[Response Ends]

1b.04. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.

Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]
Data below are from 7/1/2017-3/31/2020 across 49 acute care hospitals in the state of Michigan. This includes 18,625 patients treated for CAP, of whom 12.3% (2,299) were inappropriately diagnosed with CAP.
Here, we report the demographics for patients with pneumonia as compared to the demographics of patients inappropriately diagnosed with CAP. We also compare demographics of those inappropriately diagnosed in 2017 to those inappropriately diagnosed in 2020. All comparisons were conducted using chi-squared tests.

Table 3. Demographics of pneumonia cohort and inappropriately diagnosed patients, Year 2017

Variable	Pneumonia, N=2894; % (N)	Inappropriately Diagnosed with CAP, N=489; % (N)	P-value[a]
Medicaid	11.3% (327)	8.8% (43)	**0.02**
Medicare	67.9% (1959)	74.2% (363)	*
Private Insurance	20.7% (598)	17.0% (83)	*
Female	50.1% (1509)	50.2% (251)	0.96

Variable	Pneumonia, N=2894; % (N)	Inappropriately Diagnosed with CAP, N=489; % (N)	P-value[a]
Male	49.9% (1505)	49.8% (249)	*
Race Black	17% (513)	18.2% (91)	0.75
Race Other[b]	3.7% (112)	4.0% (20)	*
Race White	79.3% (2392)	77.8% (389)	*
Age 65 years or older	61.0% (1840)	66.2% (331)	**0.03**
Age < 65 years	39.0% (1177)	33.8% (169)	*

Demographic comparisons of the pneumonia cohort to those inappropriately diagnosed with CAP in 2017 indicate significant differences by payer and by age. Patients inappropriately diagnosed with CAP were more likely to have Medicare insurance (vs. private or Medicaid) compared to patients with CAP. Compared to patients with CAP, patients inappropriately diagnosed with CAP were more likely to be older than 65 years. There were no differences by race or gender.

*cell intentionally left empty

[a] P-value compares demographics of patients with pneumonia to those inappropriately diagnosed with CAP using chi-squared tests. P<0.05 considered significant.

[a]"other" race includes American Indian or Alaskan Native, Arab and Chaldean Ancestries, Asian, Native Hawaiian or Pacific Islander, Other (i.e., if patient demographic information indicates the patient is a race other than what is listed above), and Unknown (i.e., if patient's race is not indicated in the medical record).

Table 4. Demographics of entire pneumonia cohort and inappropriately diagnosed patients, Q1 2020

Variable	Pneumonia, N=1048; % (N)	Inappropriately Diagnosed with CAP, N=150; % (N)	P-value[a]
Medicaid	14.6% (153)	8.7% (13)	**0.02**
Medicare	65.7% (688)	76.7% (115)	*
Private Insurance	19.7% (206)	14.7% (22)	*
Female	50.7% (563)	49.4% (77)	0.76
Male	49.3% (548)	50.6% (79)	*
Race Black	20.7% (230)	20.5% (32)	0.99
Race Other[b]	4.0% (44)	3.9% (6)	*
Race White	75.4% (838)	75.6% (118)	*
Age 65 years or older	57.9% (644)	66.0% (103)	0.05
Age < 65 years	42.1% (468)	34.0% (53)	*

Demographic comparisons of the pneumonia cohort to those inappropriately diagnosed with CAP in quarter 1 of 2020 indicate significant differences by payer and by age. Patients inappropriately diagnosed with CAP were more likely to have Medicare insurance (vs. private or Medicaid) compared to patients with CAP. Compared to patients with CAP, patients inappropriately diagnosed with CAP were more likely to be older than 65 years. There were no differences by race or gender.

*cell intentionally left empty

Abbreviations: Q1: quarter 1

[a] P-value compares demographics of patients with pneumonia to those inappropriately diagnosed with CAP using chi-squared tests. P<0.05 considered significant.

[a]"other" race includes American Indian or Alaskan Native, Arab and Chaldean Ancestries, Asian, Native Hawaiian or Pacific Islander, Other (i.e., if patient demographic information indicates the patient is a race other than what is listed above), and Unknown (i.e., if patient's race is not indicated in the medical record).

Table 5. Trends in demographics of patients inappropriately diagnosed with CAP; 2017 vs. Q1 2020

Variable	2017 Inappropriately Diagnosed with CAP, N=489; % (N)	2020 Inappropriately Diagnosed with CAP, N=150; % (N)	P-value[a]
Medicaid	8.8% (43)	8.7% (13)	0.81
Medicare	74.2% (363)	76.7% (115)	*

Variable	2017 Inappropriately Diagnosed with CAP, N=489; % (N)	2020 Inappropriately Diagnosed with CAP, N=150; % (N)	P-value[a]
Private Insurance	17.0% (83)	14.7% (22)	*
Female	50.2% (251)	49.4% (77)	0.97
Male	49.8% (249)	50.6% (79)	*
Race Black	18.2% (91)	20.5% (32)	0.79
Race Other[b]	4.0% (20)	3.9% (6)	*
Race White	77.8% (389)	75.6% (118)	*
Age 65 years or older	66.2% (331)	66.0% (103)	0.85
Age < 65 years	33.8% (169)	34.0% (53)	*

Comparison of all of 2017 to quarter 1 of 2020 indicated no differences in demographics (payer, gender, race, and age) of patients inappropriately diagnosed with CAP, P=0.79-0.97.

*cell intentionally left empty

Abbreviations: Q1: quarter 1

[a]P-value compares demographics of patients inappropriately diagnosed with CAP in 2017 to those inappropriately diagnosed with CAP in Q1 of 2020 using chi-squared tests. P<0.05 considered significant.

[a]"other" race includes American Indian or Alaskan Native, Arab and Chaldean Ancestries, Asian, Native Hawaiian or Pacific Islander, Other (i.e., if patient demographic information indicates the patient is a race other than what is listed above), and Unknown (i.e., if patient's race is not indicated in the medical record).

[Response Ends]

1b.05. If no or limited data on disparities from the measure as specified is reported above, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in above.

[Response Begins]

N/A

[Response Ends]

Criteria 2: Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.

---

sp.01. Provide the measure title.

Measure titles should be concise yet convey who and what is being measured (see What Good Looks Like).

[Response Begins]
Inappropriate diagnosis of community-acquired pneumonia (CAP) in hospitalized medical patients; Abbreviated form: Inappropriate diagnosis of CAP
[Response Ends]

sp.02. Provide a brief description of the measure.

Including type of score, measure focus, target population, timeframe, (e.g., Percentage of adult patients aged 18-75 years receiving one or more HbA1c tests per year).

[Response Begins]
The inappropriate diagnosis of CAP in hospitalized medical patients (or "Inappropriate Diagnosis of CAP") measure is a process measure that evaluates the annual proportion of hospitalized adult medical patients treated for CAP who do not meet diagnostic criteria for pneumonia (thus are inappropriately diagnosed and treated).
[Response Ends]

sp.04. Check all the clinical condition/topic areas that apply to your measure, below.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Surgery: General*

[Response Begins]
Infectious Diseases (ID): Pneumonia and respiratory infections
Respiratory
Respiratory: Pneumonia
[Response Ends]

sp.05. Check all the non-condition specific measure domain areas that apply to your measure, below.

[Response Begins]
Safety
Safety: Healthcare Associated Infections
Safety: Overuse
[Response Ends]

sp.06. Select one or more target population categories.

Select only those target populations which can be stratified in the reporting of the measure's result.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Populations at Risk: Populations at Risk*

[Response Begins]
Adults (Age >= 18)
Elderly (Age >= 65)
[Response Ends]

sp.07. Select the levels of analysis that apply to your measure.

Check ONLY the levels of analysis for which the measure is SPECIFIED and TESTED.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]
Facility
[Response Ends]

sp.08. Indicate the care settings that apply to your measure.

Check ONLY the settings for which the measure is SPECIFIED and TESTED.
[Response Begins]
Inpatient/Hospital
[Response Ends]

sp.09. Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials.

Do not enter a URL linking to a home page or to general information. If no URL is available, indicate "none available".

[Response Begins]
https://mi-hms.org/inappropriate-diagnosis-community-acquired-pneumonia-cap-hospitalized-medical-patients
[Response Ends]

sp.11. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.

Attach an excel or csv file; if this poses an issue, contact staff . Provide descriptors for any codes. Use one file with multiple worksheets, if needed.
[Response Begins]
Available in attached Excel or csv file
[Response Ends]

Attachment: 3671_Data_Dictionary _CAP_Measure _3.22.22.xlsx

sp.12. State the numerator.

Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome).
DO NOT include the rationale for the measure.

[Response Begins]

The measure quantifies adult, hospitalized medical patients inappropriately diagnosed with pneumonia. Here, inappropriate diagnosis is defined as patients treated with antibiotics for CAP who do not meet diagnostic criteria for pneumonia. Patients are considered inappropriately diagnosed if they did not have 2 or more signs or symptoms of pneumonia (documented at some point in the 2 days prior to the hospital encounter through the first 2 days of the hospital encounter) AND meet radiographic criteria for pneumonia.
[Response Ends]

sp.13. Provide details needed to calculate the numerator.

All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets.
Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]
Patients in the numerator include those that did not have a) ≥2 signs or symptoms of pneumonia (documented at some point in the 2 days prior to the hospital encounter through the first 2 days of the hospital encounter) or did not b) meet radiographic criteria for pneumonia.

- Minor numerator exclusions:
    - Those whose only antibiotic treatment was azithromycin (treatment could be related to chronic obstructive pulmonary disease exacerbation): 2.2% (50/2301)
    - Those with a blood culture positive for a pathogenic bacteria: 1.7% (38/2301)
    - Those with a urine antigen positive for streptococcus: (0.9% [20/2301]) or legionella (0.5% [12/2301])

Signs (e.g., tachypnea, leukocytosis) and symptoms (e.g., new cough, shortness of breath) of pneumonia are found in the attached excel file. Any combination of 2 or more signs or symptoms is required to be considered appropriately diagnosed. Any patient who has 0 or 1 eligible signs or symptoms is considered inappropriately diagnosed with CAP and placed in the numerator.
In addition to signs and symptoms, data abstractors are instructed to review the medical record for any chest X-rays, chest computerized tomography (CTs), or abdominal CTs with lung findings to capture language that may be relevant to pneumonia (see excel file for definitions). Chest x-rays, chest CTs, and abdominal CTs that are obtained in the 2 days prior to the hospital encounter through day 4 of the hospital encounter should be included. Imaging results obtained on the day of transfer to the ICU should also be included. Otherwise, imaging results obtained after transfer to the intensive care unit (ICU; e.g., day 2 of transfer) should NOT be included even if it falls within the 4-day window.
Based on descriptions of radiographic criteria identified by abstractors, the following logic is used to determine if the patient met radiographic criteria for CAP for each individual image.

- Highest/first priority radiographic descriptions:
    - If interval improvement/resolution, no change from previous/no interval change, normal/no abnormalities or no evidence of pneumonia is documented, then image considered NOT to meet radiographic criteria
- Second priority radiographic descriptions (overrides other findings except first priority, above):
    - If air space density/opacity/disease, bronchopneumonia, cannot rule out pneumonia, cavitation, infection (cannot rule out infection/likely infection), infiltrate (any lobe specifications), loculations, pneumonia, necrotizing pneumonia, post-obstructive pneumonia, or consolidation is documented, then image considered to meet radiographic criteria
- If none of the above:
    - If ground glass is listed, then image considered to meet radiographic criteria
        - Exception: if ground glass plus interstitial lung disease, pulmonary edema or pulmonary vascular congestion is documented, then image considered NOT to meet radiographic criteria
    - If mass is listed, then image considered to meet radiographic criteria
        - Exception: If neoplasm/metastatic disease/malignancy is documented, then image considered NOT to meet radiographic criteria

- If nodular air space disease, then image considered to meet radiographic criteria
    - Exception: If neoplasm/metastatic disease/malignancy or interstitial lung disease is documented, then image considered NOT to meet radiographic criteria
- If pleural effusion, then image considered to meet radiographic criteria
    - Exception: If pulmonary edema, pulmonary vascular congestion, or ground glass is documented, then image considered NOT to meet radiographic criteria
- If aspiration pneumonia, then image considered to meet radiographic criteria
    - Exception: If pneumonitis is documented, then image considered NOT to meet radiographic criteria

If there were multiple radiographic images, the following prioritization applies:
If available, chest CTs that occur within 1 calendar day (-1,0,+1) of a chest X-ray or abdominal CT are prioritized (even if they conflict with other results)

- If patient has any Chest CT meeting radiographic criteria, then patient considered to meet radiographic criteria
- If the patient's Chest CT does NOT meet radiographic criteria, then the patient is considered NOT to meet radiographic criteria, and then considered inappropriately diagnosed, add to numerator
- Example
    - Chest X-ray and Chest CT on day 1. Chest X-ray says pneumonia. Chest CT says no pneumonia. Patient considered inappropriately diagnosed.
    - Chest X-ray on day 1. Chest CT on day 5. Chest X-ray says pneumonia. Chest CT says no pneumonia. Patient not considered inappropriately diagnosed.

If no chest CT is present, the following will apply

- If Abdominal CT AND/OR Chest X-Ray meet radiographic criteria, then patient considered to meet radiographic criteria
- If NEITHER Abdominal CT or Chest X-Ray meet radiographic criteria, then patient considered NOT to meet radiographic criteria, and considered inappropriately diagnosed, add to numerator

[Response Ends]

sp.14. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]
The denominator includes all adult, general care, immunocompetent, medical patients hospitalized and treated for CAP who do not have a concomitant infection.
[Response Ends]

sp.15. Provide details needed to calculate the denominator.

All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets.
Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]
The denominator includes all sampled patients eligible for abstraction during the measure period (typically annual measurement). Please see excel file (inclusion criteria tab) for detailed operationalized definitions.
Inclusion criteria:

- Adult patient admitted and discharged from the participating hospital with a discharge diagnosis (listed as any discharge diagnosis) of CAP (see excel file for ICD 10 codes)
- Admitted to a general care medicine service
- Received any eligible antibiotic therapy on day 1 or 2 of hospitalization (see excel file for eligible antibiotics)

- Immunocompetent (allowing for mild immune suppression)
- Do not have a concomitant infection (e.g., antibiotic treatment for unrelated infection, COVID-19, fungal pneumonia)

[Response Ends]

sp.16. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]
Patients are excluded from the denominator if they are/have:

- Left against medical advice or refused medical care
- Admitted on hospice
- Pregnant or breastfeeding
- Cystic fibrosis
- Pneumonia-related complication (e.g., empyema)

[Response Ends]

sp.17. Provide details needed to calculate the denominator exclusions.

All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]
Inclusion and exclusion codes and criteria are provided in the attached excel file.
[Response Ends]

sp.18. Provide all information required to stratify the measure results, if necessary.

Include the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate. Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format in the Data Dictionary field.

[Response Begins]
N/A. This measure is not stratified.
[Response Ends]

sp.19. Select the risk adjustment type.

Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.
[Response Begins]
No risk adjustment or risk stratification
[No risk adjustment or risk stratification Please Explain]
Our exclusion criteria are robust and exhaustive in order to negate the need for risk adjustment.

[Response Ends]

sp.20. Select the most relevant type of score.

Attachment: If available, please provide a sample report.
[Response Begins]
Rate/proportion
[Response Ends]

sp.21. Select the appropriate interpretation of the measure score.

Classifies interpretation of score according to whether better quality or resource use is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score
[Response Begins]
Better quality = Lower score
[Response Ends]

sp.22. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

Identify the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period of data, aggregating data; risk adjustment; etc.

[Response Begins]
The measure estimates hospital-level inappropriate diagnosis of CAP. If the hospital has elected to sample patients, they will generate a sample using eligible ICD 10 discharge codes (see excel file for ICD 10 codes). Next, they will apply electronic inclusion criteria (medicine admission, antibiotics on day 1 or 2 of hospitalization) to either their quarterly or monthly patient sample. The resulting list will be randomized, and patients screened in order of randomization. First, patients are screened for inclusion in the denominator. All adult, general care, medical patients hospitalized and treated for CAP are potentially eligible. If the patient meets eligibility criteria and does not have any exclusions, they are placed in the denominator. Patients automatically excluded from the numerator are those treated only with azithromycin, those with blood cultures positive for a pathogenic organism, and those with a positive streptococcal or legionella urinary antigen. Patients are then assessed for whether they meet diagnostic criteria for pneumonia defined as 2 or more symptoms/signs of pneumonia AND meeting radiographic criteria. If a patient does not meet diagnostic criteria they are placed in the numerator. A lower score is considered better diagnostic quality for CAP.
[Response Ends]

sp.25. If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

[Response Begins]
Sampling: Hospitals have the option to sample from their population or submit their entire population. Hospitals also have the option to sample quarterly or monthly. Over the entire year, 73 cases are recommended for the denominator. Thus, hospitals whose Initial Patient Population size is less than or equal to the minimum number of cases per quarter (N=19) or month (N=6) for the measure should not sample. A hospital may choose to use a larger sample size than is required.
Sampling Procedures:
Potentially eligible patient lists should be reviewed monthly or quarterly (as desired). Lists will be determined by the ability of the facility; however, we suggest electronically including the following criteria:

- Initial sample based on ICD-10 discharge diagnostic codes
- Exclude patients who did not receive antibiotics on day 1 or 2 of hospitalization
- Exclude patients admitted to a non-medicine service
- Exclude patients admitted to intensive care

Regardless of the option used, hospital samples must be monitored to ensure that sampling procedures consistently produce statistically valid and useful data. Due to exclusions, hospitals electing to sample cases MUST submit AT LEAST the minimum required sample size.
Eligible lists should then be randomized and reviewed in order until the desired number of cases is included (6-7/month or 19/quarter).
Minimum Sample Size:
Using the Spearman Brown prophecy, we evaluated the number of cases needed to reach each reliability threshold:

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

In order to achieve a desired reliability of 0.8, each hospital would need to abstract 73 cases annually.
Based on these data, for a desired reliability of 0.8, each hospital would need to abstract 73 cases annually or 6-7 cases per month.

[Response Ends]

sp.28. Select only the data sources for which the measure is specified.

[Response Begins]
Electronic Health Data
Electronic Health Records
Other (specify)
[Other (specify) Please Explain]
Chart review

[Response Ends]

sp.29. Identify the specific data source or data collection instrument.

For example, provide the name of the database, clinical registry, collection instrument, etc., and describe how data are collected.

[Response Begins]
Electronic medical record data. The data collection instrument is provided. Those interested in using our online REDCap tool may contact us directly to coordinate.
[Response Ends]

sp.30. Provide the data collection instrument.

[Response Begins]
Available in attached appendix in Question 1 of the Additional Section
[Response Ends]

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate fields in the Scientific Acceptability sections of the Measure Submission Form.

• Measures must be tested for all the data sources and levels of analyses that are specified. If there is more than one set of data specifications or more than one level of analysis, contact NQF staff about how to present all the testing information in one form.
• All required sections must be completed.
• For composites with outcome and resource use measures, Questions 2b.23-2b.37 (Risk Adjustment) also must be completed.
• If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), Questions 2b.11-2b.13 also must be completed.

• An appendix for supplemental materials may be submitted (see Question 1 in the Additional section), but there is no guarantee it will be reviewed.
• Contact NQF staff with any questions. Check for resources at the Submitting Standards webpage .
• For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for the 2021 Measure Evaluation Criteria and Guidance .

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.
2a. Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For instrument-based measures (including PRO-PMs) and composite performance measures, reliability should be demonstrated for the computed performance score.
2b1. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For instrument based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.
2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure;
AND
If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).
2b3. For outcome measures and other measures when indicated (e.g., resource use):
• an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; 14,15 and has demonstrated adequate discrimination and calibration
OR
• rationale/data support no risk adjustment/ stratification.
2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful 16 differences in performance;
OR
there is evidence of overall less-than-optimal performance.
2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.
2b6. Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias.
2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:
2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(if not conducted or results not adequate, justification must be submitted and accepted)

Definitions
Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality

measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.
Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
Risk factors that influence outcomes should not be specified as exclusions.
With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v.$5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Scientific Acceptability sections. For example:
2021 Submission:
Updated testing information here.
2018 Submission:
Testing from the previous submission here.

2a.01. Select only the data sources for which the measure is tested.

[Response Begins]
Electronic Health Data
Electronic Health Records
Other (specify)
[Other (specify) Please Explain]
Chart Review

[Response Ends]

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

The dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).

[Response Begins]
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.

[Response Ends]

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

Use the following format: "MM-DD-YYYY - MM-DD-YYYY"

[Response Begins]
07-01-2017 to 03-31-2020
[Response Ends]

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

Testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]
Facility
[Response Ends]

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

Identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample.

[Response Begins]

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%)

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%).

*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)

[Response Ends]

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.

If there is a minimum case count used for testing, that minimum must be reflected in the specifications.

[Response Begins]

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 inappropriate diagnosis 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 both assessed at the hospital level, 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%)
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%)

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

*Cells intentionally left empty

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 following 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 receive 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.

[Response Ends]

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.

[Response Begins]
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)

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.
[1]Please see validity documents for further information.

[Response Ends]

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

For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

[Response Begins]
As this is a process measure, no risk adjustment was performed (including for social factors).
[Response Ends]

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.07 check patient or encounter-level data; in 2a.08 enter "see validity testing section of data elements"; and enter "N/A" for 2a.09 and 2a.10.

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

Choose one or both levels.
[Response Begins]
Accountable Entity Level (e.g., signal-to-noise analysis)
[Response Ends]

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

Describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used.

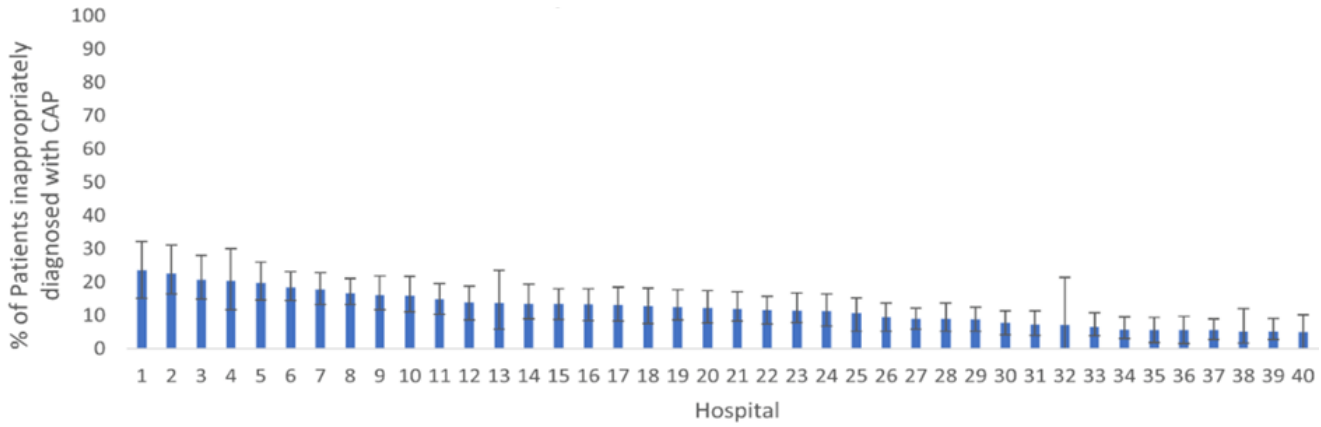
[Response Begins]
Patient or Encounter Level
Please see validity testing section for encounter-level validity.
Accountable Entity Level
Signal-to-noise analysis was performed using a mixed-effect logistic model run as an empty model such that the only effects in the model were the overall intercept and the hospital specific intercepts. This model enabled 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.

[Response Ends]

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

For example, provide the percent agreement and kappa for the critical data elements, or distribution of reliability statistics from a signal-to-noise analysis. For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers). If a particular method yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred (pg. 18, NQF Measure Evaluation Criteria).

[Response Begins]
Distribution of the 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 the 40 hospitals that provided data during all four quarters.
Figure 1. Distribution of Inappropriate diagnosis of Community-Acquired Pneumonia by Hospital



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:
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

In order to achieve a desired reliability of 0.8, each hospital would need to abstract 73 cases annually.

[Response Ends]

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

(In other words, what do the results mean and what are the norms for the test conducted?)

[Response Begins]
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.

[Response Ends]

2b.01. Select the level of validity testing that was conducted.

[Response Begins]
Patient or Encounter-Level (data element validity must address ALL critical data elements)
Empirical validity testing
Systematic assessment of face validity of performance measure score as an indicator of quality or resource use (i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance)
[Response Ends]

2b.02. For each level of testing checked above, describe the method of validity testing and what it tests.

Describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used.

[Response Begins]
We performed validity testing on multiple levels and at multiple stages of measure development. A summary of validity testing is provided in the subsequent table with details provided in the following sections.
Table 1. Summary of Validity Testing

Process	Description (stage of measure)	Results	Interpretation
During Measure Development	*	*	*
A. Face Validity-National Guidelines	Based on National Guidelines and literature review **(Early Measure)**	IDSA/ATS CAP Guidelines[1,2]	Initial basis for definitions
B. Face Validity-Expert Feedback	Data Design and Publications Committee and Michigan Hospital Medicine Safety (HMS) Consortium Hospital Experts **(Early Measure AND Current Measure as Specified)**	Refined inclusion/exclusion criteria and measure specifications to current form	Measure refinement to current measure specifications
During Early Years (2017-2019) of Measure Use	*	*	*
C. Encounter-level Validity: Inappropriate Diagnosis Case Reporting	All inappropriate diagnosis cases reported to participating hospitals **(Early Measure AND Current Measure as Specified)**	Minor adjustments based on feedback from real cases and end-users	Minor measure refinement
During Late Years (2020-2021), Specific Measure Testing	*	*	*
D. Encounter-level Validity: Assessment of Effect of Abstraction Errors	Senior project manager reviewed data elements from 50 cases (representing 33 hospitals) to assess effect of any discrepancies on encounter-level validity **(Current Measure as Specified)**	Overall abstraction accuracy was 93.7%. No changes in inappropriate diagnosis classification due to discrepancies noted in audit	Encounter-level validity is high. Data abstraction is typically accurate; what mistakes are made generally do not affect case classification.
E. Encounter-level Validity: Structured Implicit Case Review	17 cases reviewed by 2-4 physicians to confirm classification **(Late Measure, only minor updates to measure after this assessment)**	The κ for reviewer agreement was 0.72	Indicates substantial agreement

Process	Description (stage of measure)	Results	Interpretation
F. Face Validity: Feedback from HMS hospitals (N=40 hospitals)	"Approximately, what percentage of cases called [inappropriate diagnosis of community acquired pneumonia (CAP)] by HMS do you agree are [inappropriately diagnosed] (0-100%)?" **(Current Measure as Specified)**	Median: 90% IQR: 80%-95%	Most participating hospitals believed the measure was highly accurate
G. Face Validity: National Expert Panel Feedback (N=14 experts)	Individuals form 14 national organizations participated in 2 week online technical expert panel (TEP) which involved discussion of measure. **(Current Measure as Specified)**	Generally, TEP members agreed with face validity. Additional questions/data requests were answered, and responses included below. Survey Question: "The inappropriate diagnosis of CAP measure as specified can be used to distinguish between better and worse quality hospitals." Likert (1=Strongly disagree, 5=Strongly agree) 7/12 respondents (58.3%) reported that they agreed with this statement; 4/12 (33%) were neutral)	Additional feedback to improve utility of measure were provided and incorporated into the measure.
H. Face Validity: Patient Panel Feedback (N=7 patients)	Online focus group including 7 patients who had been hospitalized and treated for an infection **(Current Measure as Specified)**	Patients were asked what [inappropriate] diagnosis of infections meant to them and whether the measure would be valuable. They innately understood inappropriate diagnosis and its consequences.	Patients felt the inappropriate diagnosis of CAP measure was valid and important
I. Empirical Validity: Evaluated association with other measures of diagnostic quality	Evaluated association at hospital level between CAP inappropriate diagnosis and inappropriate diagnosis of UTI. **(Current Measure as Specified)**	Hospitals with higher rates of inappropriate diagnosis of CAP also had higher rates of inappropriate diagnosis of UTI; R=0.53 (i.e., moderate positive correlation)	Hospitals performing better on this measure were also better at appropriately diagnosing UTI
J. Empirical Validity: Evaluated association of inappropriate diagnosis of CAP with outcomes	Characterized antibiotic use in patients inappropriately diagnosed with CAP and the association of antibiotic use with adverse events after hospital discharge **(Current Measure as Specified)**	Median (IQR) 7 (5-9) unnecessary antibiotic days Each day of unnecessary antibiotic use increases odds (aOR: 1.05 [1.01, 1.08]) for developing a patient-reported antibiotic-associated adverse event after discharge.	Inappropriate diagnosis of CAP associated with unnecessary antibiotic use and antibiotic-related harm

Table 1 presents validity testing results and interpretation performed at various stages of measure development. Details are described in the text sections following the table.

*Cells intentionally left empty

A. Face Validity-National Guidelines

The inappropriate diagnosis of CAP measure was based on national guidelines for pneumonia and with additional expert feedback and review.

The 2009 Infectious Diseases Society of America/American Thoracic Society Consensus Guidelines on the Management of Community-Acquired Pneumonia in Adults define pneumonia as the following: "The diagnosis of CAP is based on the

presence of select clinical features (e.g., cough, fever, sputum production, and pleuritic chest pain) and is supported by imaging of the lung, usually by chest radiography."[1] This definition is consistent with our measure which defines inappropriate diagnosis as any patient treated for CAP that is lacking clinical or radiographic criteria. We also evaluated symptom criteria from the Society for Healthcare Epidemiology of America's evaluation of the use of non-specific symptoms in elderly populations.[3]

[1] Mandell LA, Wunderink RG, Anzueto A, et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clin Infect Dis*. 2007;44 Suppl 2:S27-72. doi:10.1086/511159. PCMID: PMC7107997.

[2] Metlay JP, Waterer GW, Long AC, et al. Diagnosis and Treatment of Adults with Community-acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America. *Am J Respir Crit Care Med*. 2019;200(7):e45-e67. doi:10.1164/rccm.201908-1581ST. PCMID: PMC6812437.

[3] Rowe, T., Jump, R., Andersen, B., et al. (2020). Reliability of nonlocalizing signs and symptoms as indicators of the presence of infection in nursing-home residents. *Infection Control & Hospital Epidemiology*, 1-10. doi:10.1017/ice.2020.1282

B. Face Validity-Expert Feedback

Throughout measure development, we obtained expert and stakeholder input via the following mechanisms:

1. Input from the Data, Design, and Publications (DDP) Committee of the Michigan Hospital Medicine Safety Consortium (HMS) early in measure development
2. Feedback from Experts in Quality, Antibiotic Stewardship, Diagnosis and Patient care from HMS hospitals

The **Data, Design, and Publications (DDP) Workgroup** was an ongoing meeting of champions and experts from HMS hospitals that met to address key issues related to measure methodology, including weighing the pros and cons of measure specifications, modeling, and use (e.g., defining the measure cohort and outcome) to ensure the measure was meaningful, useful, and well-designed. The group met approximately every 2 months during measure development and provided a forum for focused expert review and discussion of technical issues. They also provided final approval of the current submitted measure as specified.

List of DDP Workgroup Members:

- Suhasini Gudipati, MD Ascension Michigan St. Mary's Hospital
- Tina Percha, RN, MSN Beaumont Health
- Rajiv John, MD Beaumont Health
- Lama Hsaiky, PharmD Beaumont Health
- Priscila Bercea, MPH Beaumont Health Dearborn
- Scott Kaatz, DO Henry Ford Health System
- Allison Weinmann, MD Henry Ford Health System
- Emily Nerreter, MBA Henry Ford Health System
- Danielle Osterholzer, MD Hurley Medical Center
- Lisa Dumkow PharmD Mercy Health St. Mary's
- Anurag Malani, MD St. Joseph Mercy Ann Arbor Hospital
- Lakshmi Swaminathan, MD St. Joseph Mercy Ann Arbor Hospital
- Muhammad Nabeel, MD Sparrow Hospital
- Andrea White, PhD University of Utah Health
- Valerie Vaughn, MD, MSc University of Utah Health
- Vineet Chopra, MD, MSc University of Colorado Anschutz Medical Campus

Throughout measure development, we also provided opportunities from experts across the HMS collaborative to provide feedback. This included frontline clinicians, antibiotic stewards, quality improvement experts, c-suite members, and experts in quality measurement.

C. Assessment of Encounter-Level Validity: Inappropriate diagnosis Case Reporting

Once initial measure specifications had been agreed upon, we provided all inappropriate diagnosis cases to participating hospitals for review (N=2,301 cases of inappropriate diagnosis). Hospitals were encouraged to review these "fall-outs" with local experts in antibiotic stewardship, diagnosis, and quality as well as frontline clinicians to perform audit and feedback, identify trends, and assist with overall quality improvement. Occasionally, during this review the local team identified a potential issue with how the fall-out was determined based on the clinical scenario. In some instances, the case was reviewed, and we provided justification for considering the case inappropriately diagnosed. In other instances, modifications to the code and/or additional modifications to the data registry questions were required. Measure adjustments were more common during the initial launch of the measure (2017-2018). Since 2019, there have been no additional modifications to the measure based on this expert review. Since 2021, fall-out reporting has been based on the final submitted measure as currently specified.

D. Assessment of Encounter-Level Validity: Assessment of Effect of Abstraction Errors

To assess encounter-level data validity, the senior HMS project manager performed blind audits of 50 consecutive cases of patients with a diagnosis of CAP (appropriate or inappropriate). These cases included 33 hospitals. Cases were scored based correctness of data abstraction (1 point received if data element was answered correctly, 0 points if there was disagreement). The proportion of data elements abstracted correctly (based on the submitted measure as specified) were tabulated for clinical findings, chest x-ray findings, chest CT data, and overall abstraction accuracy. Correct data, as abstracted by the HMS project manager, were then reapplied to the measure definition to assess for changes in case classification.

E. Assessment of Encounter-Level Validity: Structured Implicit Case Review

In 2020, we conducted structured implicit review of cases of inappropriate diagnosis of CAP by 2-3 physicians to confirm accurate case categorization. Cases were randomly selected from "gray areas" that had been brought up during the initial measure development (e.g., patients with atelectasis as the only finding on chest imaging). During the review process, physician case reviewers had access to copies of medical record information such as diagnostic testing/results, emergency department note, history and physical note, progress notes, vital signs, and documented signs and symptoms. Reviewers were asked to independently assess whether they agreed with the classification of inappropriate diagnosis of CAP and whether they would empirically initiate antibiotics. If there was disagreement in classification, a discussion would commence that included ways to improve the measure to account for any errors in classification. We calculated the inter-rater agreement (prior to discussion) using **κ.** The comments generated through discussion were used as part of the feedback mechanism to improve the measure to the final specifications submitted here (edits in response to this feedback were minor, see details below).

F. Face Validity: Feedback from HMS hospitals (N=38 hospitals)

In October 2021 (after measure specifications had been finalized), we systematically assessed the perceived validity of the inappropriate diagnosis of CAP measure by soliciting feedback from all HMS hospitals. Via online survey, we asked all hospitals to answer the following question: "Approximately, what percentage of cases called [inappropriate diagnosis of CAP] by HMS do you agree are [inappropriately diagnosed] (0-100%)?"

G. Face Validity: National Expert Panel Feedback (N=14 experts)

Throughout measure development, we obtained expert and stakeholder input. In October 2021, we obtained formal expert feedback on the near final measure specifications by holding a two-week national technical expert panel (TEP) where societies and organizations who would potentially be impacted by the measure were asked to send a representative to provide feedback.

In alignment with the CMS Measures Management System guidance on TEP,[4] we convened a TEP to provide input and feedback from a group of recognized experts in relevant fields. To convene the TEP, we reached out to organizations whose members could potentially be impacted by the measure and asked them to nominate individuals for participation. We selected individuals to represent a range of perspectives, including Infectious Diseases physicians, pharmacists, pulmonologists, radiologists, hospitalists, emergency medicine physicians, regulatory agencies, as well as individuals with experience in quality improvement, performance measurement, diagnostic error, antibiotic stewardship, and health care quality. We held two weeks of structured TEP zoom calls consisting of a presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members. We solicited additional input and comments from the TEP via survey after the meeting. A summary of the TEP can be found in the **Appendix**.

Table 2. List of TEP Panelists and their Organizations

Organization/Institution	TEP Member
American College of Emergency Medicine (ACEP)	Larissa May
Centers for Disease Control and Prevention (CDC)	Arjun Srinivasan
Infectious Disease Society of America (IDSA)	Teena Chopra
Pew Research Center	David Hyun
Society for Healthcare Epidemiology of America (SHEA)	Dan Morgan
Society to Improve Diagnosis in Medicine (SIDM)	David Newman-Toker
Association for Professionals in Infection Control and Epidemiology (APIC)	Patty Gray
Society of Infectious Diseases Pharmacists (SIDP)	Jason Pogue
The Joint Commission	David Baker
Emergency Medicine Physician, University of Wisconsin	Michael Pulia
Society of Hospital Medicine (SHM)	Peter Lindenauer
American College of Radiology (ACR)	Ella Kazerooni
American College of Chest Physicians (CHEST)	Marcus Restrepo
American Thoracic Society (ATS)	Mark Metersky

The fourteen TEP panelists and their organizations are listed.

Following the zoom expert panel, all participants completed an online survey that included questions related to validity, reliability, usability, etc. Related to measure validity, we asked TEP members:

a. How much do you agree/disagree with the following statement?
"The inappropriate diagnosis of CAP measure as specified can be used to distinguish between better and worse quality hospitals." 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree.
b. Are there any key data elements you believe are missed or not accurately captured in the inappropriate diagnosis of CAP measure?
[4] "CMS MMS Blueprint Supplemental Material: Technical Expert Panels." September 2021. https://www.cms.gov/files/document/blueprint-technical-expert-panels.pdf

H. Face Validity: Patient Panel Feedback (N=7 patients)

To understand patient perspectives on the inappropriate diagnosis of CAP measure, we solicited patient feedback through a Patient Engagement Panel. This focus group was conducted on December 1, 2021 by the Community Collaboration and Engagement Team (CCET) which is part of the University of Utah Center for Clinical & Translational Science (CCTS). During this focus group, 7 patients and/or the caregivers of patients who had been hospitalized with infections were selected to provide feedback. Topics discussed included: how patients were diagnosed, what treatment they received, their understanding of risks and benefits with antibiotics, their perceptions about their illness and recovery, and how information about how hospitals diagnose and treat infections may inform their medical decisions. The discussion was guided by a Focus Group Discussion Guide (see Engagement Session Report for questions).

I. Empirical Validity: Evaluated association with other measures of diagnostic quality

To assess empirical validity for the inappropriate diagnosis of CAP measure, we identified and assessed the measure's correlation with other measures that target similar domains of quality for similar populations. The goal was to identify if better performance on this measure was related to better performance on other relevant structural or outcome measures. After literature review and consultations with measure experts in the field, there were very few measures identified that assess the same domains of quality.

To better understand whether inappropriate diagnosis is linked across conditions—and thus may reflect the general quality of diagnosis at a hospital—we assessed the association of inappropriate diagnosis of CAP with inappropriate diagnosis of UTI at the hospital level.

J. Empirical Validity: Evaluated association of inappropriate diagnosis of CAP with outcomes

First, we characterized antibiotic use in patients inappropriately diagnosed with CAP using descriptive statistics. Because duration was skewed, we report median (IQR/inter-quartile range) duration of antibiotic therapy.

Next, we evaluated the association of each day of unnecessary antibiotic therapy with patient outcomes at 30-days. Specifically, we were interested in the effect of each day of unnecessary antibiotic use on patient-reported antibiotic-associated adverse events (obtained through 30-day phone calls). We used generalized estimating equation models adjusted for patient characteristics to assess patient outcomes associated with each day of unnecessary antibiotic use.

[Response Ends]

2b.03. Provide the statistical results from validity testing.

Examples may include correlations or t-test results.

[Response Begins]

D. Encounter-level Validity: Assessment of Effect of Abstraction Errors

In 2021, 50 cases were chronologically selected for detailed audit for data accuracy. Audit findings were as follows:

Table 1. Results of detailed audit for data accuracy

Audit Elements	Results
Clinical Findings	95.7% of data elements abstracted correctly
Chest X-ray data	92.3% of data elements abstracted correctly
Chest CT data	94.5% of data elements abstracted correctly
Overall abstraction accuracy	93.7% of data elements abstracted correctly

Results of detailed audit of clinical findings, chest X-ray, and chest CT data. Data accuracy ranged from 92.7% to 95.7%.

When errors found through the data audit were corrected, there were no changes in case classification, as shown in Table 2.

Table 2. Classification of cases in which audited data elements disagreed (n=50)

Abstractor Classification (original)	Auditor Classification (updated)	Number (n=50)
Inappropriate Diagnosis of CAP	Inappropriate Diagnosis of CAP	6

Abstractor Classification (original)	Auditor Classification (updated)	Number (n=50)
CAP	CAP	44
Inappropriate Diagnosis of CAP	CAP	0
CAP	Inappropriate Diagnosis of CAP	0

When errors found through the audit were corrected (n=50 instances), there were no changes in case classification.

E. Encounter-level Validity: Structured Implicit Case Review

In 2020, 17 cases of inappropriate diagnosis of CAP underwent structured implicit case review by 2-4 physicians. **In 94% of cases (16/17) there was 100% agreement by reviewers that the cases represented inappropriate diagnosis**. In the remaining 6% (1/17) 1/3 reviewers agreed it was an inappropriate diagnosis. **The κ for reviewer agreement (prior to reconciliation) was 0.72** indicating substantial agreement. Of note, our case review involved "gray areas" rather than a random selection of cases. Thus, our true κ may be even higher. As a result of this case review process, we made minor refinements to our measure specifications including how chest CTs were assessed (they were given precedence over chest X-rays) and started including abdominal CTs with lung findings in the assessment/classification process.

F. Face Validity: Feedback from HMS hospitals (N=39 hospitals)

We systematically assessed the perceived validity (after finalization of measure specifications) of the inappropriate diagnosis of CAP measure by soliciting feedback from all participating HMS hospitals (N=39 hospitals) via the following question: "Approximately, what percentage of cases called ?PNA by HMS do you agree are ?PNA (0-100%)." Nearly all hospitals (97.4%, 38/39) responded. Respondents were local leaders or quality champions for the measures.

Median: 90% Inter-quartile range: 80%-95%

G. Face Validity: National Expert Panel Feedback

Based on conversations held during our two-week online TEP, the 14 national experts who attended our TEP generally agreed with the face validity and operationalization of the measure. They believed that patients we identified as being inappropriately diagnosed were, in fact, inappropriately diagnosed. The main concern brought up by panelists was a desire for more information on a balancing measure (i.e., under-diagnosis or missed diagnosis of CAP) and patient harm. There were also some concerns about the use of the word "over-diagnosis" in the measure name. As a result, we strengthened our literature review on under-diagnosis/missed diagnosis and added data on antibiotic overuse and patient harm as a result of inappropriate diagnosis of CAP. We also changed the measure name to "inappropriate diagnosis." There were no changes to measure specifications suggested by the TEP.

TEP Survey results:

Table 3. Distribution of TEP responses to ***Question #1***: "*The inappropriate diagnosis of CAP measure as specified can be used to distinguish between better and worse quality hospitals*."

Rating	# of Responses (N=12)	Percent (%)	Cumulative Percent (%)
5 (Strongly agree)	0	0	0
4 (Agree)	7	58.3%	58.3%
3 (Neutral)	4	33.3%	91.7%
2 (Disagree)	0	0	91.7%
1 (Strongly disagree)	1	8.3%	100.0%

The majority (91.6%) of experts on the TEP responded "Agree" or Neutral" (7/12 and 4/23, respectively). There was one response of "Strongly disagree".

Table 4. TEP responses to ***Question #2*.** "*What additional data would you like to see captured related to the inappropriate diagnosis of CAP? (free text)*" N=14 respondents (free text question)

# of Responses N=14	Response	Our Action/Response to Comment
57% (8/14)	None or N/A	None. Confirmed validity of measurement.
14% (2/14)	Duration of Treatment	Added data on duration of treatment for patients inappropriately diagnosed with CAP to measure submission. **Patients inappropriately diagnosed with CAP received a median (IQR) 7 (5-9) antibiotic days, all of which were unnecessary.**
7% (1/14)	Balancing Measure	Added additional resources on studies of underdiagnosis to measure submission (see Evidence section)
7% (1/14)	Trend in Outcomes of Denominator Over Time as Inappropriate Diagnosis Decreases	Added data on outcomes over time to measure submission (see **Table 5**, below)

# of Responses N=14	Response	Our Action/Response to Comment
7% (1/14)	How many patients over 80 years old have only 1 sign or symptom.	Added data on those over 80 (see **Table 6**, below)

The majority (57%) of experts on the TEP indicated that no additional data were needed. Suggestions for additional data included: a) duration of antibiotic treatment (2 panelists), b) balancing measure (1 panelist), c) trend in outcomes of denominator over time as inappropriate diagnosis decreases (1 panelist), and how many patients over 80 years old have only 1 sign or symptom (1 panelist). We addressed each of these in our measure submission.

Table 5. Trend in adverse outcomes over time as inappropriate diagnoses of CAP decreased from 2017 to 2020

Outcome	2017 (N=6405)	2020 (N=4961)
30-day Composite Outcome[a]	26.9% (1723)	25.4% (1260)
Death	3.5% (221)	2.9% (145)
Adverse Antibiotic Event	4.8% (306)	3.0% (147)

From 2017 to 2020, there were decreases in the proportion of adverse outcomes including a 30-day composite outcome (includes readmission, ED visit, death, C. difficile, and physician or patient reported antibiotic-associated adverse events), death, and adverse antibiotic events.

[a] Includes readmission, ED visit, death, *C. difficile*, and physician or patient reported antibiotic-associated adverse events

Table 6. Comparison of inappropriate diagnosis and proportion of patients with only one sign or symptom in patients <80 Years vs patients age 80 or older

Age	All Patients	Inappropriate Dx	1 Symptom Only
<80	13,633	11.8% (1607)	2.3% (311)
≥ 80	4,960	14.0% (694)	3.6% (177)

Table 6 compares the proportion of inappropriately diagnosed patients <80 vs ≥80 years with only 1 sign or symptom. The proportion of inappropriate diagnosis of CAP and having 1 sign or symptom only was greater in the 80 or older group (14.0% vs 11.8% and 3.6% vs 2.3%, respectively for inappropriate diagnosis of CAP and having only 1 sign or symptom).

H. Face Validity: Patient Panel Feedback:

A summary of the findings from the Patient Engagement Panel can be found in the **Appendix.**

Generally, the patients who participated in our panel innately understood the meaning of over-diagnosis or inappropriate diagnosis:

"[over-diagnosis is] taking a somewhat minor issue and overemphasizing it and then maybe overtreating it"

"I was over-diagnosed by the doctor that I went to… I originally went because I had [a cough]… they didn't do any tests; he thought it was pneumonia and never did a test for it; he gave me 3 antibiotics within a 4-week time and so I feel like that is a perfect case of over-diagnosis. [Doctor says] hey, you're sick, I don't want to do a test, so take this." [Note. This participant was later admitted to another hospital with C. diff]

Patients also felt that measuring inappropriate diagnosis of infections was important and meaningful:

"That's [correct diagnosis] step 1… it takes me back to grad school…problem definition – you gotta make sure you're solving the right problem – that's the first step. If you don't, you're going to end up going down all these paths that are not going to lead you to the right answer."

"If you were to have a measure of more correct diagnosis and incorrect diagnosis, and I would do it on the hospital scale, … I feel like if you were to get the correct diagnosis… I would automatically assume that you are getting the correct dose of medicine."

"I would like it if they had a hospital rating… I think it would be beneficial, and I would really appreciate that. I feel that it would affect my decision of where I would go… it would definitely affect where I would guide my family or loved one to go."

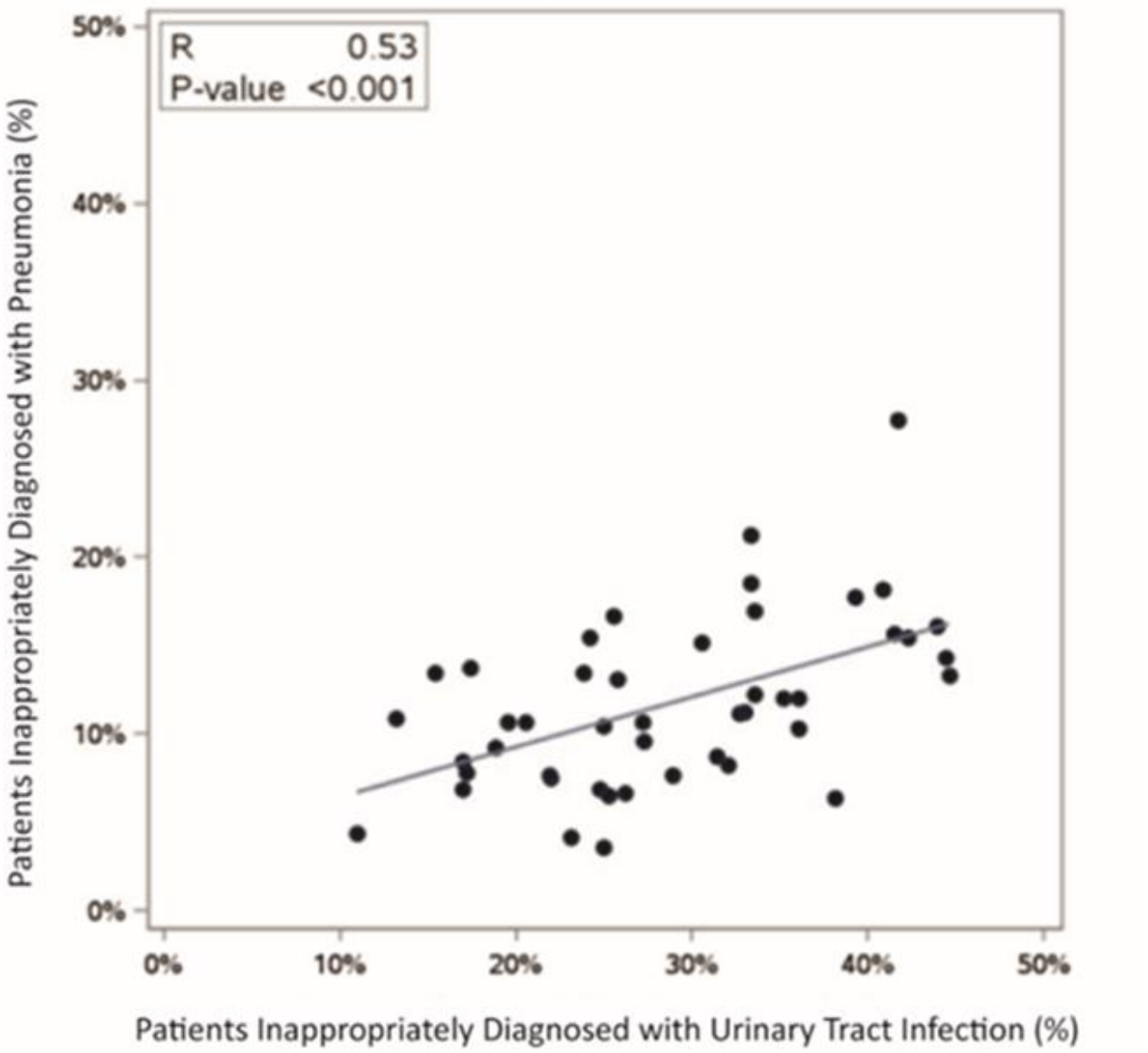
A participant has been looking for a care facility for his 98-year-old mother, utilizing U.S. News & Reports rankings. He said, "So yeah, I've been relying on that and I would definitely use something similar or look for something like that on the internet for a hospital."

I. Empirical Validity: Association with Other Measures of Diagnostic Quality

To address whether inappropriate diagnosis of CAP was correlated with other domains of quality, we assessed whether inappropriate diagnosis of CAP (as currently specified) was related to the inappropriate diagnosis of UTI. This manuscript, Misdiagnosis of Urinary Tract Infection Linked to Misdiagnosis of Pneumonia: A Multi-Hospital Cohort Study, is *in press* at *BMJ Quality & Safety*. In it, we analyzed 10,398 patients treated for UTI and 14,085 patients treated for CAP from HMS

hospitals between July 1, 2017 and March 31, 2020 and found that inappropriate diagnosis of CAP is moderately correlated with inappropriate diagnosis of UTI at the hospital level:

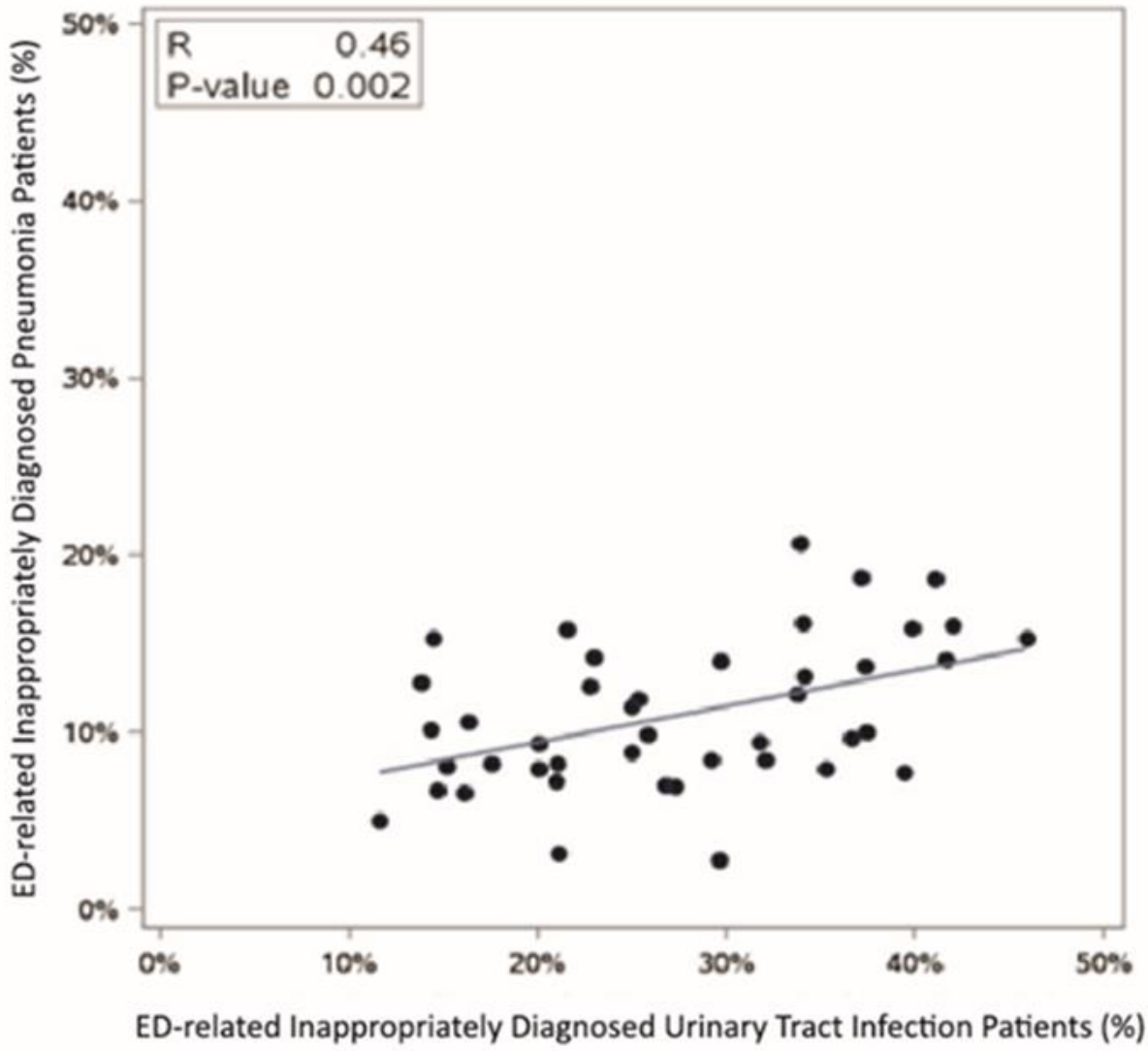
Figure 1. Relationship between inappropriate diagnosis of UTI and inappropriate diagnosis of CAP at the hospital level.



In a sample of 10,398 patients treated for UTI and 14,085 patients treated for CAP from HMS hospitals, the percent of patients with inappropriate diagnosis of UTI is moderately correlated with the percent of patients with inappropriate diagnosis of CAP at the hospital level (R=0.53; P<0.001).

These findings were also true for 2,049 patients initially inappropriately diagnosed in the Emergency Room.

Figure 2. Relationship between inappropriate diagnosis of UTI and inappropriate diagnosis of CAP in Emergency Rooms.



In a sample of 2,049 patients from 46 hospitals and diagnosed in the Emergency Room, the percent of patients with inappropriate diagnosis of UTI is moderately correlated with the percent of patients with inappropriate diagnosis of CAP at the hospital level (R=0.45; P<0.002).

[4] Gupta A, Petty L, Gandhi T, et al. Overdiagnosis of urinary tract infection linked to overdiagnosis of pneumonia: a multihospital cohort study. *BMJ Qual Saf*, 2022. doi:10.1136/bmjqs-2021-013565.

J. Empirical Validity: Association of Inappropriate diagnosis of CAP with Outcomes

There are three main harms associated with inappropriate diagnosis of CAP: delayed time to true diagnosis, antibiotic-associated adverse events, and antibiotic resistance. In our validation cohort of patients inappropriately diagnosed with CAP across HMS hospitals, patients inappropriately diagnosed with CAP received a median (IQR) 7 (5-9) antibiotic days, all of which were unnecessary. Those antibiotics were associated with harm such as antibiotic-associated adverse events, *C. difficile* infection, and antibiotic resistance without benefit (as they did not have bacterial infections). After adjustments, each additional day of antibiotic use in patients inappropriately diagnosed with CAP was associated with an increased odds ratio of 1.05 (1.01, 1.08) for developing a patient-reported antibiotic-associated adverse event.

Furthermore, as noted above in the response to TEP questions, we found that as inappropriate diagnosis of CAP (as currently specified) decreased over time, outcomes improved in HMS hospitals (**Table 5**, above).

[Response Ends]

2b.04. Provide your interpretation of the results in terms of demonstrating validity. (i.e., what do the results mean and what are the norms for the test conducted?)

[Response Begins]
The validity of the inappropriate diagnosis of CAP measure is supported by three types of evidence: (1) strong face validity based on national guidelines and expert opinion and as gauged by feedback from Technical Expert Panel (TEP) members, patients, and end-users (hospitals, patients); (2) strong encounter level validity as demonstrated by implicit review, evaluation of data abstraction errors, and hospital encounter-level feedback; (3) external empiric comparisons with other quality measures; and (4) validity of the outcome.

Face validity
The validity of the measure is supported by strong face validity results, as measured by systematic feedback from the TEP. Perhaps even more important both patients and hospitals—the true end-users of the measure—found the measures to be valid. HMS hospitals who received measure scores found the measures to be highly valid, reporting they believed 90% of cases called inappropriate diagnosis of CAP were in fact inappropriately diagnosed.

Encounter-level Validity
Encounter-level validity is supported by substantial agreement between physician reviewers on case classification (κ=0.72) and by the long-standing general agreement by hospital experts with case classification during data feedback. Furthermore, in an assessment of the effect of abstraction errors on case classification, 93.7% of data elements were abstracted correctly and the minor discrepancies that existed resulted in no changes in case classification.

Empirical Validity Testing
The validity of the measure is further supported by the empiric validation results which demonstrate a correlation (in the expected strength and direction) between the inappropriate diagnosis of CAP measure and measures of inappropriate diagnosis of other infections, namely UTI. As expected, we found hospitals that performed worse on one measure also performed worse on the other. Thus, the inappropriate diagnosis of CAP measure may reflect the overall quality of diagnosis at a hospital.

Validity of the Outcome
The validity of the outcome is supported by the relationship between inappropriate diagnosis of CAP and antibiotic-associated adverse events—including improvement in outcomes over time as measure performance improves.

[Response Ends]

2b.05. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified.

Describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided in Importance to Measure and Report: Gap in Care/Disparities.

[Response Begins]
We used the Spearman Brown prophecy to determine the minimum number of cases that hospitals participating in this measure would need to capture on an annual basis in order to allow us to distinguish performance accurately and reliably. Our analysis suggests that to meet the 0.8 standard for reliability, hospitals would need to abstract 73 cases annually.
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

In order to achieve a desired reliability of 0.8, each hospital would need to abstract 73 cases annually.
Of the 40 hospitals participating in HMS in 2019 (our most recent year), 37/40 (92.5%) were able to meet this minimum standard of 73 annual cases (the 3 that did not were either small hospitals or had abstractor turnover). If we lowered the threshold for reliability to 0.7, 39 of 40 hospitals (97.5%) would have been able to meet this minimum threshold of 43 cases.
To further characterize the degree of variability in the measure score, we analyzed hospitals in the HMS cohort and:

1. Report the distribution of the measure score
2. Calculate the mean; standard deviation; median; and 10th, 25th, 75th, and 90th percentile of the performance scores for each quarter.

3. Group hospitals by quartiles and assess whether the difference in mean measure score between each adjacent quartile was statistically significant.

[Response Ends]

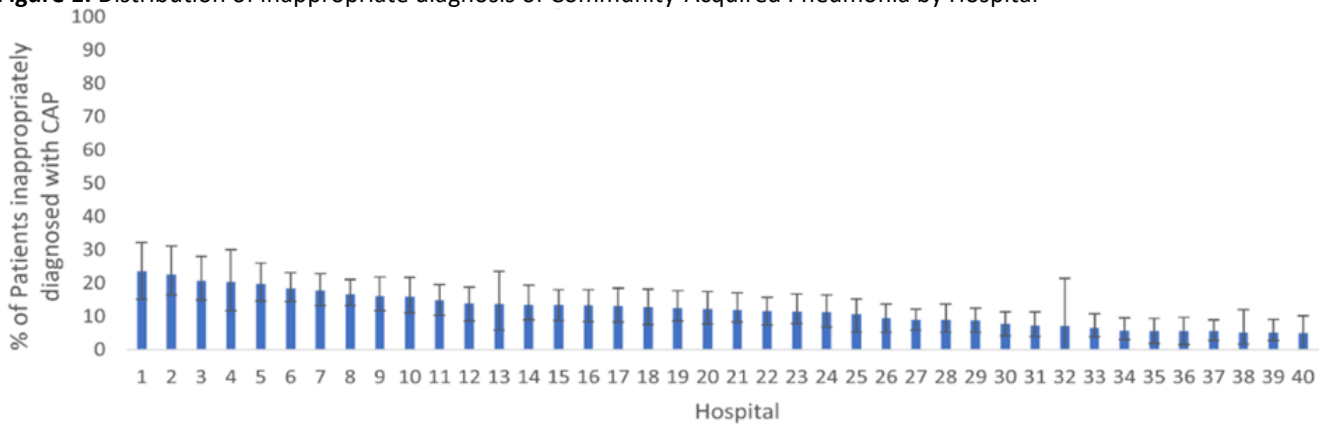
2b.06. Describe the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities.

Examples may include number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined.

[Response Begins]

The distribution of the measure for all 40 hospitals (each hospital=1 blue bar) is shown below in **Figure 1** with error bars representing 95% confidence intervals. **Table 2** shows summary statistics for all years combined, the first 4 quarters, and the final 4 quarters.

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



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 preceding March 2020 and include only hospitals that provided data during all four quarters (N=40 hospitals).

Table 2. Summary Statistics for all years combined, the first 4 quarters, and the final 4 quarters.

Year	Number of Hospitals	Number of Pneumonia Patients	Overall Mean Inappropriate-diagnosis	Hospital Adjusted Mean (SD)	Min-Max	10th Percentile (better performance)	25th Percentile	Median	75th Percentile	90th Percentile (worse performance)
All years	47	18,463	12.4% (2,288/18,463)	12.7% (0.69%)	4.6%-27.8%	6.7%	8.7%	13.1%	15.2%	20.0%
First 4 Quarters	46	6,614	13.2% (881/6614)	13.5% (0.85%)	5.1%-27.8%	6.0%	8.2%	13.8%	18.0%	21.9%
Last 4 quarters	41	7,028	12.2% (857/7028)	12.1% (1.0%)	4.2%-23.7%	5.3%	7.3%	12.0%	15.5%	20.3%

Summary statistics for all years combined, the first 4 quarters, and the final 4 quarters. Percent of patients inappropriately diagnosed with CAP decreased over time from the first 4 quarters to the last 4 quarters: 12.2% to 13.4% overall, 5.3% to 6.7% for the 10th percentile (better performance), and from 20.0% to 21.9% for the 90th percentile (worse performance).

Compared with average-performing hospitals, hospitals in the 10th percentile (better performance) have about 7 fewer patients inappropriately diagnosed with CAP per 100 CAP discharges than the median (~49 fewer unnecessary antibiotic use days/100 CAP discharges), and hospitals in the 90th percentile (worse performing) have approximately 10 more patients inappropriately diagnosed with CAP per 100 CAP discharges than the median (~70 more unnecessary antibiotic use days/100 CAP discharges).
The grouping of hospitals by quartiles for all years, first 4 quarters, and last 4 quarters, is shown in **Table 3**. All quartiles are statistically significantly different from other quartiles.

Table 3. Differences in percent of patients inappropriately diagnosed with CAP between adjacent quartiles of performance

Percentile comparison	Lower Quartile	Higher Quartile	Test statistic	*p*-value
All years: 1st (best) quartile (0-25%) vs. 2nd quartile (25-50%)	6.64%	10.72%	6.99	<0.001
All years: 2nd (25%-50%) vs. 3rd quartile (50%-75%)	10.72%	14.28%	5.27	<0.001
All years: 3rd (50%-75%) vs. 4th (worst) quartile (75%-100%)	14.28%	18.71%	5.52	<0.001
First 4 quarters: 1st (best) quartile (0-25%) vs. 2nd quartile (25-50%)	6.35%	10.66%	4.43	<0.001
First 4 quarters: 2nd (25%-50%) vs. 3rd quartile (50%-75%)	10.66%	15.52%	3.96	<0.001
First 4 quarters: 3rd (50%-75%) vs. 4th (worst) quartile (75%-100%)	15.52%	20.98%	3.74	<0.001
Last 4 quarters: 1st (best) quartile (0-25%) vs. 2nd quartile (25-50%)	5.44%	9.67%	4.43	<0.001
Last 4 quarters: 2nd (25%-50%) vs. 3rd quartile (50%-75%)	9.67%	13.41%	3.51	<0.001
Last 4 quarters: 3rd (50%-75%) vs. 4th (worst) quartile (75%-100%)	13.41%	18.70%	4.27	<0.001

Differences in percent of patients inappropriately diagnosed with CAP between adjacent quartiles of performance for the inappropriate diagnosis of CAP measure were statistically significant (P<0.001) for adjacent quarters overall (all years combined), for the first 4 quarters, and for the last 4 quarters.

[Response Ends]

2b.07. Provide your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities.

In other words, what do the results mean in terms of statistical and meaningful differences?

[Response Begins]
The measure was able to detect facilities with above- and below-average performance. In the first year, facility scores ranged from 5.1% to 27.8% with a mean performance of 13.5%. By the final year, facility scores had improved somewhat and ranged from 4.2% to 23.7% with a mean performance of 12.1%.
Our analysis showed a statistically significant difference in performance between each quartile of hospitals, suggesting consistent performance gaps across facilities and targets for improvement.

[Response Ends]

2b.08. Describe the method of testing conducted to identify the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders). Include how the specified handling of missing data minimizes bias.

Describe the steps—do not just name a method; what statistical analysis was used.

[Response Begins]
This measure is calculated using chart-abstracted data. To limit the effects of missing data, abstractors cannot submit a value of "missing" for individual data elements because the case will be rejected by the abstraction tool. Although abstractors cannot submit missing data, for some data (e.g., white blood cell count) they may submit a value of

"unknown" or "not available." For cases submitted by hospitals from July 2017 through March 2020, we calculated the number of cases that had missing data or had "unknown" values for data elements used in case classification.
[Response Ends]

2b.09. Provide the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data.

For example, provide results of sensitivity analysis of the effect of various rules for missing data/non-response. If no empirical sensitivity analysis was conducted, identify the approaches for handling missing data that were considered and benefits and drawbacks of each).

[Response Begins]
Data that were missing or marked as "unknown/not available" are presented below. Some of these data are accurately missing (e.g., no chest imaging obtained during hospitalization), others are missing due to errors.
As expected, missing data were extremely rare. The percentage of encounters with missing, "unknown," or "not available" values was less than 1.0% (183/18,468) of all included patients.
Table 4. Percent of encounters with data that were missing or marked as "unknown/not available" in N=18,468 hospitalized CAP patients

N=18,468 patients	Missing/Unknown/Not Available
No chest imaging	0.3% (63/18,468)
Hypoxemia	0.2% (36/18,468)
Auscultatory findings	0 (2/18,468)
Temperature	0.1% (14/18,468)
White Blood Cell Count	0.3% (50/18,468)
Cough	0 (2/18,468)
Sputum	0 (2/18,468)
Dyspnea	0.1% (14/18,468)

The percentage of cases with missing or "unknown/not available" clinical symptom data ranged from <0.1% (14/18,468) to 0.3% (63/18,468), suggesting that missing data had little effect on performance results or other findings.

[Response Ends]

2b.10. Provide your interpretation of the results, in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and non-responders), and how the specified handling of missing data minimizes bias.

In other words, what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis was conducted, justify the selected approach for missing data.

[Response Begins]
The percentage of cases that could potentially be affected by missing data is negligible, indicating that missing data did not affect the performance results or other findings.
[Response Ends]

Note: This item is directed to measures that are risk-adjusted (with or without social risk factors) OR to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eCQMs). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

2b.11. Indicate whether there is more than one set of specifications for this measure.

[Response Begins]
No, there is only one set of specifications for this measure
[Response Ends]

2b.12. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications.

Describe the steps—do not just name a method. Indicate what statistical analysis was used.

[Response Begins]
[Response Ends]

2b.13. Provide the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications.

Examples may include correlation, and/or rank order.

[Response Begins]
[Response Ends]

2b.14. Provide your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications.

In other words, what do the results mean and what are the norms for the test conducted.

[Response Begins]
[Response Ends]

2b.15. Indicate whether the measure uses exclusions.

[Response Begins]
Yes, the measure uses exclusions.
[Response Ends]

2b.16. Describe the method of testing exclusions and what was tested.

Describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used?

[Response Begins]
All exclusions were determined by careful clinical review and discussion and feedback from our national expert panel and HMS' Data, Design, and Publications Committee.
Exclusion criteria (and reasoning) include:

- Patients who left against medical advice or refused medical care
    - This exclusion is needed for acceptability of the measure to hospitals, who do not have the opportunity to deliver full care
- Patients who were pregnant or breastfeeding
    - This exclusion is needed for acceptability of the measure to hospitals, as pregnancy/breastfeeding present diagnostic and treatment challenges that may differ from patients who are not pregnant/breastfeeding
- Patients admitted on hospice or comfort care

- This exclusion is needed for acceptability of the measure to hospitals, who may appropriately adjust their treatment and diagnostic procedures to comply with patient desires

- Patients with cystic fibrosis
    - This exclusion is needed for acceptability of the measure to hospitals, as cystic fibrosis presents diagnostic and treatment challenges that may differ from patients without cystic fibrosis
- Patients with a pneumonia-related complication (operationalized by excluding patients discharged on more than 14 days of antibiotic therapy)
    - This exclusion is needed for acceptability of the measure to hospitals. Pneumonia-related complications are not well documented on ICD or other coding but are important reasons to treat patients more aggressively. Generally, patients discharged on more than 14 days of antibiotics do not have typical pneumonia and have an alternative reason or complication for extended therapy.

To assess how common exclusion criteria were, we reviewed the literature—including national databases (Medicaid, Medicare, Premier) to estimate typical numbers of patients excluded for the above reasons. For the final exclusion criterion, we were able to estimate this directly from the HMS database.

[Response Ends]

2b.17. Provide the statistical results from testing exclusions.

Include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores.

[Response Begins]
Our exclusion results are shown below
Table 1. Percent of individuals excluded based on exclusion criteria

Exclusion	Percent of Patients Excluded: Estimates from the Literature/HMS
Patients who left against medical advice	0.37% [1]
Patients who were pregnant or breastfeeding	0.8% [2]
Patients admitted on hospice or comfort care	0.33% (Medicaid) to 0.62% (Medicare) 1% (Premier) [3]
Patients with cystic fibrosis	0.18% (Premier) [4]
Pneumonia-related complication (>14 days of antibiotics at discharge)	0.3% (9/3197)- HMS Estimates
Total	**2.68%-3.35%**

The total percent of patients excluded based on exclusion criteria estimated from the literature and HMS data ranges from 2.68% to 3.35%. Individual exclusion criteria would exclude 0.18% to 1.0% of patients.
In addition, we provided all exclusion criteria to participating hospitals and our technical expert panel to ensure they appeared feasible and reasonable. There was generally agreement across our groups that the exclusions led to a more accurate and fair assessment of patients over-diagnosed with community-acquired pneumonia. For example, one surveyed respondent reported, "I think the exclusion criteria for this initiative protect against vulnerable patients who actually have an infection being untreated because of this measure. Therefore, no issues on my end." There were no additional exclusion criteria requested or suggested criteria to be removed from the TEP.

[1] YNHHSC/CORE. Excess Days in Acute Care (EDAC) Measures Methodology. CMS.gov. Methodology Web site. https://qualitynet.cms.gov/inpatient/measures/edac/methodology. Published 2021. Accessed 11/20/2021.
[2] Dinh A, Ropers J, Duran C, et al. Discontinuing beta-lactam treatment after 3 days for patients with community-acquired pneumonia in non-critical care wards (PTC): a double-blind, randomised, placebo-controlled, non-inferiority trial. *Lancet*. 2021;397(10280):1195-1203. doi:10.1016/S0140-6736(21)00313-5.
[3] Lindenauer PK, Stefan MS, Shieh MS, Pekow PS, Rothberg MB, Hill NS. Outcomes associated with invasive and noninvasive ventilation among patients hospitalized with exacerbations of chronic obstructive pulmonary disease. *JAMA Intern Med*. 2014;174(12):1982-1993. doi:10.1001/jamainternmed.2014.5430. PCMID: PMC4501470.
[4] Rothberg MB, Pekow PS, Priya A, et al. Using highly detailed administrative data to predict pneumonia mortality. *PLoS One*. 2014;9(1):e87382. doi:10.1371/journal.pone.0087382. PCMID: PMC3909106.

[Response Ends]

2b.18. Provide your interpretation of the results, in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results.

In other words, the value outweighs the burden of increased data collection and analysis. Note: If patient preference is an exclusion, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion.

[Response Begins]
Exclusions were uncommon. When present they were needed to improve acceptability by the hospitals. Feedback from our TEP and from end-user hospitals was supportive of the exclusions in their current form.
[Response Ends]

2b.19. Check all methods used to address risk factors.

[Response Begins]
No risk adjustment or stratification
[Response Ends]

2b.20. If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, risk factor data sources, coefficients, equations, codes with descriptors, and definitions.

[Response Begins]
n/a
[Response Ends]

2b.21. If an outcome or resource use measure is not risk-adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (i.e., case mix) is not needed to achieve fair comparisons across measured entities.

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure.
In the context of healthcare performance assessment, the purpose of the risk model is to reduce bias due to case mix characteristics present at the start of care (i.e., to risk adjust), not to totally explain variation in outcomes, which would require also including variables about quality of care. Variables related to quality of care are purposely not included in risk models for performance measures used to assess quality.[5]
Specifically, CMS notes:

- "Process measures are not risk-adjusted; rather the target population of a process measure is defined to include all patients for whom the process measure is appropriate."
- "The variation in measured entity-level (e.g., clinician or facility) performance may be due to variation in quality or variation in factors that are independent of quality (e.g., factors like the age or severity of illness of patients). Independent of quality means that the clinician treats the patients exactly the same way, but patients who have the factor (older or sicker) have worse outcomes than patients who do not (younger or less sick)."

[5] Measures Management System Risk Adjustment. Centers for Medicare & Medicaid. Measure Management & You Web site. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/Downloads/Risk-Adjustment.pdf. Published 2017. Accessed 11/30/2021.

[Response Ends]

2b.22. Select all applicable resources and methods used to develop the conceptual model of how social risk impacts this outcome.

[Response Begins]
Published literature
Internal data analysis
[Response Ends]

2b.23. Describe the conceptual and statistical methods and criteria used to test and select patient-level risk factors (e.g., clinical factors, social risk factors) used in the statistical risk model or for stratification by risk.

Please be sure to address the following: potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10 or other statistical tests; correlation of x or higher. Patient factors should be present at the start of care, if applicable. Also discuss any "ordering" of risk factor inclusion; note whether social risk factors are added after all clinical factors. Discuss any considerations regarding data sources (e.g., availability, specificity).

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure. No risk model/stratification.
[Response Ends]

2b.24. Detail the statistical results of the analyses used to test and select risk factors for inclusion in or exclusion from the risk model/stratification.

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure. No risk model/stratification.
[Response Ends]

2b.25. Describe the analyses and interpretation resulting in the decision to select or not select social risk factors.

Examples may include prevalence of the factor across measured entities, availability of the data source, empirical association with the outcome, contribution of unique variation in the outcome, or assessment of between-unit effects and within-unit effects. Also describe the impact of adjusting for risk (or making no adjustment) on providers at high or low extremes of risk.

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure. No risk model/stratification.
[Response Ends]

2b.26. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (describe the steps—do not just name a method; what statistical analysis was used). Provide the statistical results from testing the approach to control for differences in patient characteristics (i.e., case mix) below. If stratified ONLY, enter "N/A" for questions about the statistical risk model discrimination and calibration statistics.

Validation testing should be conducted in a data set that is separate from the one used to develop the model.

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure. No risk model/stratification.
[Response Ends]

2b.27. Provide risk model discrimination statistics.

For example, provide c-statistics or R-squared values.

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure. No risk model/stratification.
[Response Ends]

2b.28. Provide the statistical risk model calibration statistics (e.g., Hosmer-Lemeshow statistic).

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure.
In the context of healthcare performance assessment, the purpose of the risk model is to reduce bias due to case mix characteristics present at the start of care (i.e., to risk adjust), not to totally explain variation in outcomes, which would require also including variables about quality of care. Variables related to quality of care are purposely not included in risk models for performance measures used to assess quality.[5]
Specifically, CMS notes:

- "Process measures are not risk-adjusted; rather the target population of a process measure is defined to include all patients for whom the process measure is appropriate."
- "The variation in measured entity-level (e.g., clinician or facility) performance may be due to variation in quality or variation in factors that are independent of quality (e.g., factors like the age or severity of illness of patients). Independent of quality means that the clinician treats the patients exactly the same way, but patients who have the factor (older or sicker) have worse outcomes than patients who do not (younger or less sick)."

[5] Measures Management System Risk Adjustment. Centers for Medicare & Medicaid. Measure Management & You Web site. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/Downloads/Risk-Adjustment.pdf. Published 2017. Accessed 11/30/2021.

[Response Ends]

2b.29. Provide the risk decile plots or calibration curves used in calibrating the statistical risk model.

The preferred file format is .png, but most image formats are acceptable.

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure. No risk model/stratification.
[Response Ends]

2b.30. Provide the results of the risk stratification analysis.

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure. No risk model/stratification.
[Response Ends]

2b.31. Provide your interpretation of the results, in terms of demonstrating adequacy of controlling for differences in patient characteristics (i.e., case mix).

In other words, what do the results mean and what are the norms for the test conducted?

[Response Begins]
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure. No risk model/stratification.
[Response Ends]

2b.32. Describe any additional testing conducted to justify the risk adjustment approach used in specifying the measure.

Not required but would provide additional support of adequacy of the risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed.

[Response Begins]
N/A. No risk model/stratification.
[Response Ends]

Criteria 3: Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3.01. Check all methods below that are used to generate the data elements needed to compute the measure score.

[Response Begins]
Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
[Response Ends]

3.02. Detail to what extent the specified data elements are available electronically in defined fields.

In other words, indicate whether data elements that are needed to compute the performance measure score are in defined, computer-readable fields.
[Response Begins]
Some data elements are in defined fields in electronic sources
[Response Ends]

3.03. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using data elements not from electronic sources.

[Response Begins]
Currently, some of the inappropriate diagnosis of CAP data elements can be captured electronically in discrete fields (e.g., vital signs, laboratory values). However, not all documentation required to report the inappropriate diagnosis of CAP measure can be captured electronically in discrete fields. In particular **symptoms of pneumonia** and **radiographic findings** are not in defined, computer-readable fields.
Rationale for Using Data Elements not from Electronic Sources
While efforts are being made to facilitate an electronic measure (see below), gaps remain in the ability to electronically capture all of the required data for measure validity. The inappropriate diagnosis of CAP measure requires data abstractors to review documentation in various formats, including narrative free-text, to identify the specific information necessary to report the measure. Preliminary efforts to convert the inappropriate diagnosis of CAP measure to an eCQM within the current Health Quality Measure Format/Quality Data Model frameworks showed that the transition is not immediately feasible.
First, symptoms are generally located in free-text spaces within the medical record, and their location varies by hospital. Symptoms are critical to measure validity, as pneumonia is a clinical diagnosis and radiographs and other laboratory data are non-specific. Measures of diagnostic accuracy of pneumonia thus require clinical data—namely, symptoms.
Second, radiographic interpretation currently requires use of data elements that are not discrete.
One potential method to reduce manual data collection needs would be to remove symptoms from the measure definition, and to rely on radiographs alone to determine whether a patient did or did not meet criteria for pneumonia. We tested this simplified measure to assess whether evaluation of radiographic data alone would remain sufficiently valid to assess quality of diagnosis of pneumonia in individual patients and across hospitals. The breakdown of inappropriately diagnosed cases classified by number of signs or symptoms and radiographic findings is shown below. If signs and symptoms were not considered part of the inappropriate diagnosis of pneumonia criteria, 23.7% (544/2,299) of cases of inappropriate diagnosis would be missed because they met radiographic criteria but had 0 or only 1 sign or symptom of pneumonia. Thus, diagnosis using radiographs alone substantially reduced patient-level validity.

Table 1. Symptoms and Radiographic Criteria for Inappropriately Diagnosed Cases (N=2299)

Signs/Symptoms	CAP radiographic criteria[1] not met	CAP radiographic[1] criteria met
No Symptoms	0.3% (8/2,299)	4.0% (93/2,299)*

Signs/Symptoms	CAP radiographic criteria[1] not met	CAP radiographic[1] criteria met
1 Symptom	1.6% (37/2,299)	19.6% (451/2,299)*
2 or more Symptoms	74.4% (1,710/2,299)	0
Total	**76.3%** (1,755/2,299)	**23.7%** (544/2299)****

Evaluation of radiographic data alone resulted in failure to detect 23.7% (544/2299) of inappropriately diagnosed pneumonia cases.

[1]From chest CT or chest X-ray

*cases that would be misclassified if symptoms were not included in measure specifications

We then tested whether the simplified measure using radiographs alone (without symptoms) was a valid assessment of diagnostic accuracy across hospitals. When the simplified definition was applied across all 49 HMS hospitals, the absolute percent of patients considered inappropriately diagnosed with pneumonia changed, on average, by 2% per hospital (median 3%). For example, the average hospital whose prior inappropriate diagnosis of pneumonia score was 13% now only had 9% of their CAP patients classified as inappropriately diagnosed. This misclassification did not affect all hospitals equally (range 0-13% change). We divided hospitals into performance quartiles and compared their performance quartile calculated by the full method vs. the simplified radiological assessment only. When comparing hospital performance quartiles, this new definition would change performance quartile for nearly a third 31% [15/49]) of hospitals. Therefore, this attempt to limit the need for unstructured data was insufficient to accurately assess inappropriate diagnosis of CAP at the hospital level when compared to our proposed measure which uses both radiographic and symptom data to determine diagnostic accuracy.

Finally, an alternative method to electronically determine inappropriate diagnosis of pneumonia would be to combine radiographic findings with signs that could potentially be captured electronically, namely: respiratory rate >20, white blood cell count >10,000 or <4,000, temperature <36.1 C or >38.0 C, oxygen saturation <90%, or partial pressure of arterial oxygen <60 mmHg. If we redefined inappropriate diagnosis of CAP as any patient treated for pneumonia that lacked radiographic findings or had zero of the above clinical findings, then, compared to the full definition, 5.7% (1,068/18,625) of patients would change classification. The difference between the full proposed definition vs. this new definition is shown:

Table 2. Case Classification Comparing the Full Definition and a Modified Version only Including Signs that Could be Captured Electronically, N= 18,625 patients

*	Pneumonia by Full Definition, N=16,326	Inappropriate Diagnosis of CAP by Full Definition, N=2299
Pneumonia by Modified Definition^	83.5% (15556/18625)	1.6% (298/18625)*
Inappropriate Diagnosis of CAP by Modified Definition	4.1% (770/18625)*	10.7% (2001/18625)

*Indicates the table cell left intentionally blank

Compared to the full definition, the modified definition (which only includes signs potentially captured electronically) inappropriately classifies 5.7% of patients treated for pneumonia including 1.6% it incorrectly classifies as pneumonia and 4.1% it incorrectly classifies as inappropriate diagnosis of CAP.

^Modified definition considers any patients not meeting radiographic criteria or not having at least one sign that could be captured electronically (i.e., respiratory rate >20, white blood cell count >10,000 or <4,000, temperature <36.1 C or >38.0 C, oxygen saturation <90%, or partial pressure of arterial oxygen <60 mmHg) to be inappropriately diagnosed

*Cases whose classification would change if the criteria were changed

The sensitivity and specificity for this modified measure to detect inappropriate diagnosis of CAP is 87.0% and 95.3%, respectively, compared to the full proposed definition. Unfortunately, when the new definition was applied across all 49 HMS hospitals, and hospitals were re-divided into performance quartiles, this new definition would change performance quartile for 39% (19/49) of hospitals. Thus, this simplified version of the measure is insufficiently accurate to assess inappropriate diagnosis of CAP at the hospital level when compared to our proposed measure which includes non-discrete signs and symptom data.

Consequently, at this time there is a strong rationale to require chart review of symptoms and radiographs in order to ensure a valid, reliable measure of inappropriate diagnosis of CAP. Fortunately, there is a credible, near-term path to electronic measurement development, as described below.

[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]

There are several promising pathways to eCQM development of the Inappropriate Diagnosis of CAP measure. First, there are methods under development to assess diagnostic accuracy of pneumonia using natural language processing.[1] Several tools have been developed to identify evidence of pneumonia within chest radiograph reports [2-4] and directly from images,[5,6] some of which are in clinical use (e.g., in Kaiser Permanente hospitals).[7] These have not been validated in diverse systems, and additional prospective validation is needed. In addition, limitations of computational infrastructure are barriers to natural language processing, particularly in hospitals not using standardized electronic health systems or with limited technologic support. Second, there have been promising steps toward electronic capture of symptoms either through natural language processing or by making symptoms part of discrete data fields (e.g., through templated notes or required indications when antibiotics are ordered).[8,9] Third, there have been efforts to standardize radiology reports using templates to more easily allow electronic interpretation of radiographs.[10,11] These ongoing efforts provide a solid roadmap for how the measure can segue from chart review to electronic based. When this occurs, the measure would need to be retested to ensure validity before being adapted as an eCQM. Multiple research efforts (e.g., through the Centers for Disease Control and Prevention Shepherd projects and through the Gordon and Betty Moore Foundation) continue to make progress on eCQM development for pneumonia diagnosis.

References

1. Hwang EJ, Park S, Jin KN, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA Netw Open. 2019;2(3):e191095.
2. Dublin S, Baldwin E, Walker RL, et al. Natural Language Processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf. 2013;22(8):834-841.
3. Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. J Biomed Inform. 2001;34(1):4-14.
4. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc. 2000;7(6):593-604.
5. Irvin JA, Pareek A, Long J, et al. CheXED: Comparison of a Deep Learning Model to a Clinical Decision Support System for Pneumonia in the Emergency Department. J Thorac Imaging. 2021.
6. Majkowska A, Mittal S, Steiner DF, et al. Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. Radiology. 2020;294(2):421-431.
7. Liu V, Clark MP, Mendoza M, et al. Automated identification of pneumonia in chest radiograph reports in critically ill patients. BMC Med Inform Decis Mak. 2013;13:90.
8. Bastarache L, Brown JS, Cimino JJ, et al. Developing real-world evidence from real-world data: Transforming raw data into analytical datasets. Learn Health Syst. 2022;6(1):e10293.
9. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. 2019;26(4):364-379.
10. Ganeshan D, Duong PT, Probyn L, et al. Structured Reporting in Radiology. Acad Radiol. 2018;25(1):66-73.
11. Chung CY, Makeeva V, Yan J, et al. Improving Billing Accuracy Through Enterprise-Wide Standardized Structured Reporting With Cross-Divisional Shared Templates. J Am Coll Radiol. 2020;17(1 Pt B):157-164.

[Response Ends]

3.06. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

[Response Begins]

Data Collection, Availability, Missing Data

This measure is calculated using chart-abstracted data and using only data typically captured during patient encounters. No additional data are required for the measure that are not captured during the typical process of patient care. Because these data are captured as standard practice, missing data were extremely rare. The percentage of encounters with missing, "unknown," or "not available" values was less than 1.0% (183/18,468) of all included patients.

Timing/Frequency of Data Collection and Patient Sampling

Hospitals have the option to sample from their population or submit their entire population. Hospitals also have the option to sample quarterly or monthly. Over the entire year, 73 cases are recommended for the denominator. Thus, hospitals whose Initial Patient Population size is less than or equal to the minimum number of cases per quarter (N=19) or month (N=6) for the measure should not sample. A hospital may choose to use a larger sample size than is required.

Using the current HMS hospital 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) was able to abstract the 43 cases/year necessary to achieve 0.7 reliability.

Patient Confidentiality

Data are deidentified.

Time and Cost

To improve feasibility and reduce time and cost of data collection, we removed all non-essential data collection elements from the measure during measure testing. We also reviewed exclusion criteria to remove those that were uncommon and would not impact measure outcomes. This pared down data collection form was tested at 4 hospitals in Utah to estimate the time needed for case review. Those results follow:

- Review of eligibility criteria to determine whether a patient would be included vs. excluded took 1-3 minutes.
- Review time could be reduced by adding exclusion criteria (e.g., ICU admission) electronically to lists for review.
- Across the 4 hospitals, 39.2%-60.2% of patients reviewed for inclusion were eligible (see Table for details)
- Once determined to be eligible, case review took 15 to 30 minutes per case.

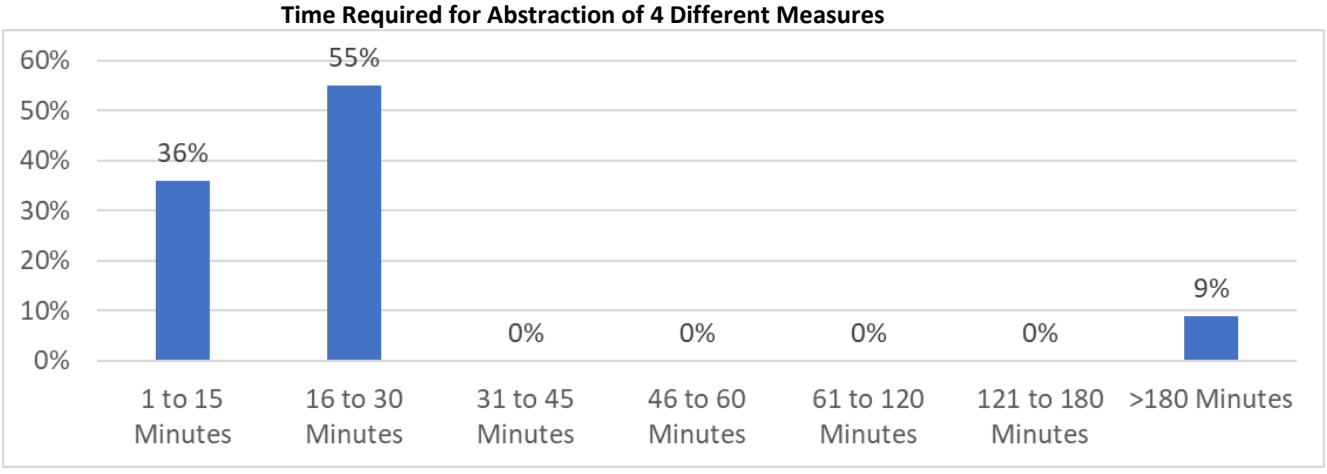
Table 3. Percent of cases meeting inclusion criteria from case reviews performed at 4 Utah hospitals over a 6-month time period.

*	# beds	# cases reviewed	# cases included	% included
Hospital 1	1000	217	89	41.0%
Hospital 2	502	130	51	39.2%
Hospital 3	90	115	50	43.5%
Hospital 4	132	83	50	60.2%

*Indicates the table cell left intentionally blank

Our case review form was utilized to review pneumonia cases in 4 Utah hospitals (90 to 1000 beds) retrospectively, over a 6-month period. The number of cases reviewed ranged from 83 to 217, and of those 39.2% to 60.2% met criteria for the inappropriate diagnosis of CAP measure.

When speaking to Infection Preventionists at included hospitals, the time for data collection of our measure was on par with other NHSN measures currently requiring case review (e.g., CAUTI, CLABSI, SSI, CDI, VAP). They all noted that feasibility improved for those measures over the years as electronic health record vendors built modules to reduce initial screening. The Joint Commission also provided comparative data during our Technical Expert Panel. They noted that 4 chart review measures abstracted across 11 sites had similar time requirements to our proposed measure.



Time Required for Abstraction of 4 Different Measures. The majority of abstractions (91%) took 30 minutes or less to complete (36% 1-15 minutes; 55% 16-30 minutes; 9% >180 minutes).

*Data provided by Dr. David Baker of The Joint Commission

During our technical expert panel, we surveyed our experts on measure feasibility via the following two questions:

1. How appropriate is the quantity of information collected for use in determining inappropriate diagnosis of CAP? (N=14 experts)

- 64% (9/14) responded it was the correct amount of data

2. Compared to other measures requiring chart review, how easy do you believe it would be for a hospital to collect the data needed to assess whether a case represents an inappropriate diagnosis of CAP? (N=14 experts)

- 71% (10/14) reported it would be "about the same as other measures"
- 14.3% reported it would be easier and 14.3% reported it would be more difficult than other measures

We also surveyed hospitals participating in HMS (N=40) to ask about their experiences with the feasibility of the inappropriate diagnosis of CAP measure.

1. **How easy is it for your hospital to collect the data needed to assess whether a case represents an inappropriate diagnosis of CAP? (N=40 hospitals)**

Table 4. Rating of difficulty to perform case abstractions for the inappropriate diagnosis of CAP measure.

Rating	Response; N (%)
Very Easy	8 (20.0%)
Easy	10 (25.0%)
Neither Easy nor Difficult	14 (35.0%)
Difficult	7 (17.5%)
Very Difficult	1 (2.5%)

The majority of respondents, 80.0% (32/40), reported it was very easy, easy, or neither easy nor difficult to perform case abstractions for the inappropriate diagnosis of CAP measure.

[Response Ends]

Consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

3.07. Detail any fees, licensing, or other requirements to use any aspect of the measure as specified (e.g., value/code set, risk model, programming code, algorithm),

Attach the fee schedule here, if applicable.

[Response Begins]
All measures are free to use. Data dictionaries and data collection templates are free and accessible at our website (https://mi-hms.org/inappropriate-diagnosis-community-acquired-pneumonia-cap-hospitalized-medical-patients).
[Response Ends]

Criteria 4: Use and Usability

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement, in addition to demonstrating performance improvement.

4a.01. Check all current uses. For each current use checked, please provide:

Name of program and sponsor

URL

Purpose

Geographic area and number and percentage of accountable entities and patients included

Level of measurement and setting

[Response Begins]
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)
[Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Please Explain]
Program: Michigan Hospital Medicine Safety Consortium (HMS)
Sponsor: Blue Cross and Blue Shield of Michigan
URL:
https://mi-hms.org/quality-initiatives/antimicrobial-use-initiative
Purpose: To improve outcomes of hospitalized patients with community-acquired pneumonia.
Geographic area: Acute care hospitals in the state of Michigan. From inception, this includes 49 hospitals and 18,625 patients treated for CAP.
Level of measurement and setting: We collect patient-level data which is evaluated for inappropriate diagnosis of CAP. Hospitals receive a list of all patients considered inappropriately diagnosed. In addition, aggregated data on inappropriate diagnosis of CAP from each hospital is presented quarterly and annually to hospitals to allow them to compare: a) performance in their own hospital over time and b) performance compared to other hospitals participating in HMS.

[Response Ends]

4a.02. Check all planned uses.

[Response Begins]
Payment Program
Regulatory and Accreditation Program
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)
Quality Improvement (internal to the specific organization)
Measure Currently in Use
[Response Ends]

4a.03. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing), explain why the measure is not in use.

For example, do policies or actions of the developer/steward or accountable entities restrict access to performance results or block implementation?

[Response Begins]
[Response Ends]

4a.04. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes: used in any accountability application within 3 years, and publicly reported within 6 years of initial endorsement.

A credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.

[Response Begins]
[Response Ends]

4a.05. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

Detail how many and which types of measured entities and/or others were included. If only a sample of measured entities were included, describe the full population and how the sample was selected.

[Response Begins]
Since 2017, the inappropriate diagnosis of CAP measure has been in use through the Michigan Hospital Medicine Safety Consortium (HMS) to measure and improve care for hospitalized patients with CAP. HMS is a collaborative quality initiative of 60+ hospitals across the state of Michigan whose purpose is to improve the care of hospitalized infections. As part of its Antimicrobial Use Initiative, data have been collected from a pseudo-random population of hospitalized patients treated for CAP. Every quarter, participating hospitals receive data on the proportion of patients treated for CAP at their hospital that are inappropriately diagnosed. In addition, each hospital receives data on how their performance compares to all other hospitals in HMS and how their performance has changed over time. Hospitals also receive a list of patients who were considered inappropriately diagnosed so that they can further evaluate inappropriate diagnosis and use those data to drive internal quality improvement efforts.
[Response Ends]

4a.06. Describe the process for providing measure results, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

[Response Begins]
Tri-annual Collaborative Wide Meetings
Individuals from participating hospitals meet in person three times a year. We encourage hospitals to send their Clinical Data Abstractors, physician champions, and quality leads, as well as other individuals from their hospital that might be interested in participation. These meetings take place three times per year – in March, July, and November. Traditionally, meetings took place in-person at venues across Michigan. In 2020 and 2021, these meetings were hosted via an on-line format due to COVID-19.
The tri-annual meetings provide individuals from member hospitals with the opportunity engage with each other in a variety of formats. Each meeting includes a formal discussion of the data from each of the HMS initiatives—including data on inappropriate diagnosis of CAP—for the previous quarter, presentations from member hospitals and expert guests, breakout/work group sessions, and networking opportunities. These meetings allow individuals from member hospitals to network with individuals from other hospitals who have excelled in those areas to seek ideas on how to improve their performance. It also allows for an opportunity for feedback and to answer questions related to their measure performance.
Site-specific Reports on Measure Performance
Tri-annually, each participating hospital receives a printed and email version of a site-specific data report. These reports are also available daily within the database/registry (see below). These reports provide an in-depth look into the performance of each site. For example, we provide hospital data on the number of patients inappropriately vs.

appropriately diagnosed with CAP, details on antibiotic use and outcomes (e.g., adverse events), longitudinal performance, and data on how individual hospitals compare to other hospitals in the state in terms of inappropriate diagnosis. Hospitals also receive a list of all patients who were considered "inappropriately diagnosed with CAP" to enable them to return to their hospital and conduct case reviews of those patients. Each hospital is encouraged to review these cases with their local team to perform audit and feedback, identify trends, and assist with overall quality improvement. This also provides an opportunity for measure feedback—for example, hospitals might find an error in case classification. Early during measure development this case-specific feedback was critical for improving measure validity.

Live Database Reports

Each of the HMS databases are equipped with the ability to view live reports utilizing Business Objects software. These reports provide updated data every 24 hours regarding measures (site performance and collaborative performance), fallout case information, demographics, critical/non-critical data errors, completeness of abstracted cases, and case classification information.

Individuals who participate in the collaborative either as a Clinical Data Abstractor or a quality administrator have the ability to log into the HMS databases and view these reports at their leisure. The software that HMS utilizes also allows for these reports to be exported as Excel files or PDFs for hospital-specific customization. This information is often utilized by participating hospitals at committee meetings or for presentations to track progress and inform quality improvement efforts. They also assist the Clinical Data Abstractor to identify errors in their abstraction and resolve them in real time. These reports also allow hospitals to review individual fallout cases and their clinical scenarios to inform individual clinicians or groups of clinicians of their performance and provide targeted education.

[Response Ends]

4a.07. Summarize the feedback on measure performance and implementation from the measured entities and others. Describe how feedback was obtained.

[Response Begins]

Throughout measure development, we received feedback on the measure performance/validity through three mechanisms: 1) Expert Feedback from Data Design and Publications Committee and Michigan Hospital Medicine Safety (HMS) Consortium Hospital Experts/Representatives, 2) "Fall-out" Feedback, and 3) October 2021 Hospital Survey. Feedback from the Data, Design, and Publications Committee and "Fall-out" feedback has been described in the "validity" section. Briefly, measure performance feedback allowed us to refine the measures to the current version. The Data, Design, and Publications Committee approved the measures for use across HMS.

In October 2021, we systematically assessed the perceived use and usability of the inappropriate diagnosis of CAP measure by soliciting feedback from HMS hospitals participating at that time (N=40) via an online survey. Specifically, we asked all participating hospitals (N=40) to answer the following questions:

> Q1. Please briefly describe how you have used or are planning to use the [inappropriate diagnosis of CAP] measure to improve care.

Responses: The 40 responses to this open-ended question largely fell into a few broad categories. These including using the measures to enable education/feedback, improve antibiotic use, change order sets, and none/no plans. Examples related to education/feedback include "have used it to provide feedback to clinicians in cases of inappropriate use, as one more tool discouraging antibiotic use", or "providing feedback to providers, including sharing our performance on the measure at our site meetings and at educational sessions, along with sharing articles and publications from HMS has improved care. We will be continuing to have education sessions." With regard to improving antibiotic use, examples include "less antibiotic use", "we have used to assess need for appropriate antibiotic use." In the area of changing order sets, examples include "put required stop field at 5 days to help enhance our provider compliance with the measure" or "order set changed for pneumonia."

> Q2. What perceived barriers do you see/foresee to using the [inappropriate diagnosis of CAP] measure to guide care improvement?

Responses: Nearly half (42.5%, 17/40) of hospitals indicated that they don't see/foresee any barriers. One-quarter (27.5%, 11/40) noted issues with physician pushback/buy in. Statements made here include "providers level of comfort in changing their practices" or "providers not receptive to the treatment guides and changes to care". There was also a broad category that can be characterized as lack of time/resources (15%, 6/40), which includes "time to re-educate clinicians on clinical findings of pneumonia," or "lack of resources to conduct audit with feedback on all pneumonia cases (notably hospitals in HMS are required to conduct many more cases than we are suggesting in our measure submission)". Finally, several participants included complexity of patients, such as "over-lapping symptoms/borderline cases with heart failure", or "not all patients align with the pneumonia measures used by HMS."

[Response Ends]

4a.08. Summarize the feedback obtained from those being measured.

[Response Begins]
Response: In summary, feedback from hospitals on measure performance was used to inform the development and refinement of the measures as currently submitted. In addition, feedback on measure implementation was broadly positive—that the measures were useful to guide care and improve diagnosis and antibiotic use. Based on feedback that time was a barrier to data collection, we limited the amount of data to be collected (average time for case collection 15-30 minutes) and decreased the number of cases we request be abstracted to still achieve a high reliability (N=73 cases). Thus, the measure submitted in this proposal should have even higher feasibility with similar usability as the measure tested in the Michigan Hospital Medicine Safety Consortium.
[Response Ends]

4a.09. Summarize the feedback obtained from other users.

[Response Begins]
In addition to the hospital feedback described above, we conducted a Patient Engagement Panel in order to understand patient perspectives about antibiotic treatment during hospitalization with infection. Seven individuals who were hospitalized or had a close family member who was hospitalized for an infection and received antibiotics participated in a 90-minute focus group. A discussion guide was used to assess participants' knowledge and perceptions about: how diagnoses are made and what information is needed; antibiotic risks and benefits; certainty of diagnosis and timing of treatment initiation; whether knowing how well a hospital accurately diagnoses infections would influence treatment choices. A brief summary of the Patient Engagement Panel is presented below.
Table 1. Summary of Patient Engagement Panel

Question/Topic	Responses	Impression
Understanding of how infection diagnosis was made	Patients were aware of the necessity of tests (e.g., chest X-rays), labs (e.g., urine and blood tests), and clinical signs and symptoms (e.g., fever, O_2 saturation, pain, cough) in determining the diagnosis of infection. They relied on physicians' knowledge, but in some instances understood there may disagreement.	Patients understood that a process is involved in diagnosis; that diagnosis is reliant on lab results (which take time); and that there may be some uncertainty and thus differing opinions of physicians.
Risks and Benefits of Antibiotics	Patients universally agreed that antibiotics are beneficial: quickly reducing symptoms and clearing infections; necessary for treatment of severe illness. The discussion of risks identified many concerns: antibiotic resistance, allergic reactions, disruptions to gut microbiome, side effects from drug:drug interactions.	Patients understood there were both benefits and risks of antibiotic treatment.
What does overdiagnosis mean? Under-diagnosis?	Patients expressed several ideas about what "over-diagnosis" is: "prescribing medication whether needed or not", "when a minor issue is overemphasized and overtreated", "antibiotics given without tests being done". The idea of "under-diagnosis" was expressed as: "settling on a routine diagnosis when something more significant is happening", "not utilizing antibiotics", "not enough concern when treating a routine" infection.	Patients understood that "over-diagnosis" relates to treatment that may not be necessary and that "under-diagnosis" involves the possibility of missing the diagnosis and not receiving the appropriate treatment.

Question/Topic	Responses	Impression
How do you know if a hospital is doing a good job? What would help you to know?	Patients were aware that hospitals are rated on certain performance measures. They also expressed some skepticism about these due to: not knowing what the ratings are based on, variations in individual physicians (e.g., a top-rated hospital could still have a low-rated physician and vice versa), concern that hospitals could "game" the system of measurement. Even so, patients expressed interest in being able to access ratings of performance for aspects of healthcare.	Patients were receptive to information about hospital performance measures, especially if they had some assurances that they could be trusted. They were interested in measures of diagnostic performance as a way to make informed decisions about hospitals.

Based on the focus group discussion, the measure is consistent with their understanding and expectations of diagnosis and treatment of infection.

Summary of patient feedback: Based on the focus group discussion, the measure is consistent with their understanding and expectations of diagnosis and treatment of infection. There were no issues or concerns raised that would necessitate modifications of the measure.

[Response Ends]

4a.10. Describe how the feedback described has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

[Response Begins]

Feedback from HMS hospitals, the technical expert panel, and the patient engagement panel was all used to refine the measure. Major changes include: a) simplification of measure, b) refinement of measure specifications, c) streamline/decrease in amount of data requested for assessment, and d) defining minimum cases necessary for abstraction to decrease number of cases required to be submitted. We also received feedback on the naming of the measure. When we first began measure development, the measure had been named "over-diagnosis of pneumonia" which we changed to "inappropriate diagnosis of CAP" based on feedback from diagnostic error experts in our technical expert panel and to avoid confusion as "over-diagnosis" has alternate meanings in the diagnostic error community.

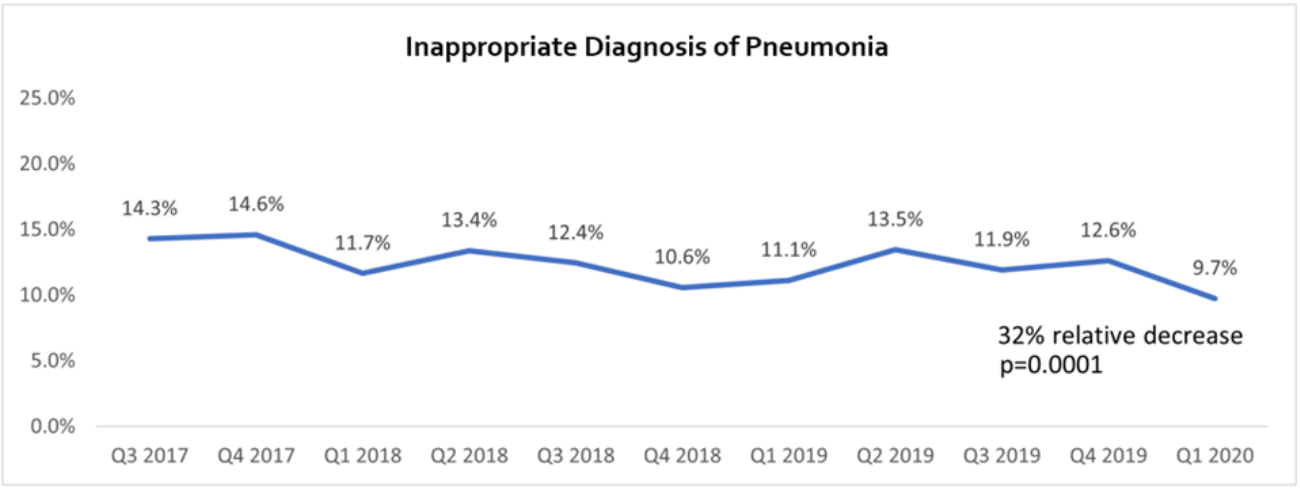
[Response Ends]

4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included). If no improvement was demonstrated, provide an explanation. If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

[Response Begins]

Since 2017, when the measure first began being reported to 49 participating HMS hospitals, we have seen a 32% relative decrease ($P<0.001$) in the percentage of patients inappropriately diagnosed with CAP (see Figure). This represents an improvement in diagnosis, reduction in unnecessary antibiotic use, and improved care.

Figure 1. Percent of inappropriate diagnosis of CAP cases in 49 HMS hospitals from 2017 to 2020



The percent of inappropriate diagnosis of CAP cases in 49 participating HMS hospitals decreased significantly from 14.6% in 2017 to 9.7% in 2020 (P<0.001).

In addition, since 2017, we have seen a statistically significant (though minor) decrease in antibiotic duration for patients inappropriately diagnosed with CAP (driven mostly by fewer cases treated with antibiotic durations longer than 9 days). Though no antibiotic therapy is ideal for this patient population, there is often diagnostic uncertainty that drives brief empiric therapy. Stopping this therapy as soon as possible can reduce the risk of harm. In fact, we found that each additional day of unnecessary antibiotic use places patients at higher risk of antibiotic-associated adverse events reported by patients 30 days after hospitalization.

Table 1. Association of antibiotic use with outcomes for hospitalized patients inappropriately diagnosed with CAP, N=2,286 patients

N = 2,286 patients	Unadjusted Odds ratio per day of antibiotic use (95% CI)	Unadjusted P Values	Adjusted* Odds ratio per day of antibiotic use (95% CI)	Adjusted P Values
C Diff Infection	0.94 (0.79, 1.14)	0.55	0.95 (0.79, 1.14)	0.61
Physician Reported Adverse Event	0.98 (0.91, 1.05)	0.61	0.99 (0.92, 1.06)	0.71
Patient Reported Adverse Event	**1.05 (1.01, 1.09)**	**0.005**	**1.05 (1.01, 1.08)**	**0.01**

Each additional day of unnecessary antibiotic use was associated with a significantly higher risk of antibiotic-associated adverse events reported by patients 30 days after hospitalization.

Furthermore, we found that as inappropriate diagnosis of CAP decreased over time, outcomes improved for all patients treated for CAP in HMS hospitals.

Table 2. Adverse events in patients treated for CAP in HMS hospitals in 2017 vs 2020.

Outcomes	2017 (N=6405)	2020 (N=4961)
30-day Composite Outcome[a]	26.9% (1723)	25.4% (1260)
Death	3.5% (221)	2.9% (145)
Adverse-Antibiotic Event	4.8% (306)	3.0% (147)

The proportion of adverse events in patients treated for CAP in HMS hospitals decreased from 2017 to 2020.

[a]Includes readmission, ED visit, death, *C. difficile*, and physician or patient-reported antibiotic-associated adverse event.

[Response Ends]

4b.02. Explain any unexpected findings (positive or negative) during implementation of this measure, including unintended impacts on patients.

[Response Begins]

There were no unexpected findings. Expected findings included decreased rates of inappropriate diagnosis of CAP, decreased unnecessary antibiotic use, decreased antibiotic-associated adverse events.

In October 2021, we systematically assessed the perceived use and usability of the inappropriate diagnosis of CAP measure by soliciting feedback from HMS hospitals participating at that time (N=40). Via online survey, we asked all hospitals to answer the following questions:

1. What unintended consequences do you see/foresee to using the [inappropriate diagnosis of CAP] measure to guide care improvement? (Q561)
   Most respondents said none/unknown (31/40). Of those that did note an unintended consequence, five noted lack of appropriate treatment "decreased antibiotic days, slow improvement of patients or readmission." Other respondents noted strained physician relationships or need to get physician cooperation, or the time to do feedback after the fact to validate accuracy.

[Response Ends]

4b.03. Explain any unexpected benefits realized from implementation of this measure.

[Response Begins]
Generally, there were no "unexpected benefits." Expected benefits included decreased rates of inappropriate diagnosis of CAP, decreased unnecessary antibiotic use, decreased antibiotic-associated adverse events.
In October 2021, we systematically assessed the perceived use and usability of the inappropriate diagnosis of CAP measure by soliciting feedback from HMS hospitals participating at that time (N=40). Via online survey, we asked all hospitals to answer the following question:

1. If you have already started work based on the [inappropriate diagnosis of CAP] measure, what unexpected benefits have been realized from implementing this measure?
   Responses: 10: N/A, 7: none, 4: unsure/not sure; 3: formal work not yet started
   A number of respondents (5) identified improved antibiotic use, either in terms of decrease in excess antibiotics on discharge, reduced use of fluoroquinolones, or an increase of patients appropriately receiving five days of antibiotic therapy. Two individuals indicated that because of their work on pneumonia, they have improved awareness of duration of therapy for other infections. Two hospitals made changes in their order sets, such as removal of default duration on discharge prescriptions or a general update of their pneumonia order sets/pathways.

[Response Ends]

Criteria 5: Related and Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

If you are updating a maintenance measure submission for the first time in MIMS, please note that the previous related and competing data appearing in question 5.03 may need to be entered in to 5.01 and 5.02, if the measures are NQF endorsed. Please review and update questions 5.01, 5.02, and 5.03 accordingly.

5.01. Search and select all NQF-endorsed related measures (conceptually, either same measure focus or target population).

(Can search and select measures.)
[Response Begins]
0468: Hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following pneumonia hospitalization
[Response Ends]

5.02. Search and select all NQF-endorsed competing measures (conceptually, the measures have both the same measure focus or target population).

(Can search and select measures.)
[Response Begins]
[Response Ends]

5.03. If there are related or competing measures to this measure, but they are not NQF-endorsed, please indicate the measure title and steward.

[Response Begins]
N/A
[Response Ends]

5.04. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s), indicate whether the measure specifications are harmonized to the extent possible.

[Response Begins]
No
[Response Ends]

5.05. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.

[Response Begins]
NQF 0468 estimates hospital 30-day, all-cause, risk-standardized mortality rate (RSMR) following pneumonia hospitalization in adults aged 65 years and older. There is overlap in target populations for NQF 0468 and our measure, which includes adults aged 18 years and older with a discharge diagnosis of pneumonia. Both measures use ICD-10 codes to identify pneumonia cases; our measure contains 6 codes not listed in NQF 0468 (see below), but this difference is not likely to have a significant effect on case identification. NQF 0468 is a claims-based measure of mortality, while the Inappropriate Diagnosis of CAP measure is based on medical record abstraction; therefore, the data collection burdens for both measures do not overlap.
Codes included in Inappropriate Diagnosis of CAP, not in 0468:
J17 (pneumonia is disease classification elsewhere)
J84.111 (idiopathic interstitial pneumonia)
J84.116 (Cryptogenic organizing pneumonia)
J84.117 (Desquamative interstitial pneumonia)

J84.2 (Lymphoid interstitial pneumonia)
J69.8 (pneumonitis due to inhalation of other solids or liquids)

[Response Ends]

5.06. Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality). Alternatively, justify endorsing an additional measure.

Provide analyses when possible.

[Response Begins]
N/A
[Response Ends]