NQF questions:

This information is submitted on the NQF Testing Attachment

1) **Select the level of validity testing that was conducted**

- ☐ Patient or Encounter-Level (data element validity must address ALL critical data elements)
- ☐ Accountable Entity Level (e.g. hospitals, clinicians)
- ☐ 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)

2) **For each level of testing checked above, describe the method of validity testing and what it tests**

Evaluation Criteria:

2b. Validity 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) validity must be demonstrated for the data element level as well as for the computed performance score. For composite performance measures, validity must be demonstrated for the computed performance score by the time of endorsement maintenance; if empirical testing of the computed performance score is not feasible at the time of initial endorsement, acceptable alternatives include systematic assessment of content or face validity of the composite performance measure or demonstration that each of the component measures meet NQF subcriteria for validity (via either empirical testing or face validity) .

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.

Process	Description **(stage of measure)**	Results	Interpretation
During Measure Development	*	*	*
A. Face Validity- National Guidelines	Based on National Guidelines and literature review (**Early Measure**)	2019 IDSA Asymptomatic Bacteriuria Guidelines[1]	Initial basis for definitions
B. Face Validity- Expert Feedback	Data Design and Publications Committee and Michigan Hospital Medicine Safety Consortium (HMS) 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 inappropriately diagnosed cases reported to participating hospital **(Early Measure AND Current Measure as Specified)**	Minor adjustments based on feedback from real cases	Minor measure refinement

Process	Description **(stage of measure)**	Results	Interpretation
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 29 hospitals) to assess effect of any discrepancies on encounter-level validity **(Current Measure as Specified)**	Overall abstraction accuracy was 98.6%. Two cases changed classification due to discrepancies noted in audit. IRR: Kappa = 0.91 95% CI (0.78 – 1.00) Strong to "almost perfect" reliability	Encounter-level validity is high with a "strong" to "almost perfect" reliability. Data abstraction is typically accurate; what mistakes are made generally do not affect case classification.
E. Encounter-level Validity: Structured Implicit Case Review	25 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
F. Face Validity: Feedback from HMS hospitals (N=40 hospitals)	"Approximately, what percentage of cases called [inappropriate diagnosis of UTI] by HMS do you agree are [inappropriately diagnosed] (0-100%)?" **(Current Measure as Specified)**	Median: 90% IQR: 80% to 97%	Most participating hospitals believed the measure was highly accurate
G. Face Validity: National Expert Panel Feedback (N=11 experts)	Individuals representing 11 national organizations participated in 2-week online discussion of measure. **(Current Measure as Specified)**	Survey Question: "The inappropriate diagnosis of UTI measure as specified can be used to distinguish between better and worse quality hospitals." Likert (1=Strongly disagree, 5=Strongly agree) 9 respondents (82%) reported that they agreed/strongly agreed with this statement.	Measure with substantial face validity by TEP Additional feedback to improve validity was 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 UTI measure was valid and important
I. Empirical Validity: Evaluated association with other measures of diagnostic quality	Evaluated association at hospital level between UTI inappropriate diagnosis and inappropriate diagnosis of community acquired pneumonia (CAP). **(Current Measure as Specified)**	Hospitals with higher rates of inappropriate diagnosis of UTI also had higher rates of inappropriate diagnosis of CAP; R=0.53 (i.e., moderate positive correlation)	Hospitals performing better on this measure were also better at appropriately diagnosing CAP
J. Empirical Validity: Evaluated association of inappropriate diagnosis of UTI with outcomes	Characterized antibiotic use in patients inappropriately diagnosed with UTI and the association of antibiotic use with adverse events after hospital discharge **(Current Measure as Specified)**	Median (IQR) 7 (4-9) unnecessary antibiotic days Patients inappropriately diagnosed with UTI had an ~1 day longer length of stay after urine testing than those with asymptomatic bacteriuria (ASB) who were not treated with antibiotics (aRR: 1.37 [1.28-1.47]).	Inappropriate diagnosis of UTI is associated with unnecessary antibiotic use and longer hospitalizations

*Cells intentionally left empty

Alt-Text for above table:

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.

A. Face Validity Indicated by Established UTI Guidelines

The initial definition of inappropriate diagnosis of UTI was derived from the "Clinical Practice Guideline for the Management of Asymptomatic Bacteriuria: 2019 Update by the Infectious Diseases Society of America."[1] Additional expert feedback and review helped refine measure development and design.

The 2019 Infectious Diseases Society of America Clinical Practice Guideline for the Management of Asymptomatic Bacteriuria (ASB) defines ASB as the following: "ASB is the presence of 1 or more species of bacteria growing in the urine at specified quantitative counts ($\geq 10^5$ colony-forming units [CFU]/mL or $\geq 10^8$ CFU/L), irrespective of the presence of pyuria, in the absence of signs or symptoms attributable to UTI."[1] This definition is consistent with our measure which defines inappropriate diagnosis of UTI as any patient treated for UTI that does not have signs or symptoms of a UTI. We also use their criteria of when to treat altered mental status as a UTI: 1) when altered mental status occurs with other symptoms or 2) when patient has "other systemic signs of infection (e.g., fever or hemodynamic instability)."[1] 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.[2]

[1] Nicolle LE, Gupta K, Bradley SF, et al. Clinical Practice Guideline for the Management of Asymptomatic Bacteriuria: 2019 Update by the Infectious Diseases Society of America. *Clin Infect Dis.* 2019;68(10):e83-e110. doi:10.1093/cid/ciy1121.

[2] 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 three 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=3197 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 UTI (appropriate or inappropriate). These cases included 29 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 daily symptoms/signs, urinary catheter 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. Using standard methods, an inter-rater reliability was obtained to assess difference between original case classification and true case classification after identifying data errors.

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

In 2020, we conducted structured implicit review of cases of inappropriate diagnosis of UTI by 2-4 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 altered mental status). 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 UTI 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=40 hospitals)

In October 2021 (after measure specifications had been finalized), we systematically assessed the perceived validity of the inappropriate diagnosis of UTI 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 UTI] by HMS do you agree are [inappropriately diagnosed] (0-100%)?"

G. Face Validity: National Expert Panel Feedback (N=11 experts)

Throughout measure development, we obtained expert and stakeholder input. In October 2021, we obtained formal expert feedback by holding a series of meetings over two-weeks with a national Technical Expert Panel (TEP). This TEP included representatives from societies and organizations who would potentially be impacted by the measure to provide feedback on the measure.

In alignment with the CMS Measures Management System guidance on TEPs,[3] 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, urologists, 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
American Urological Association (AUA)	Micheal Liss

The eleven TEP panelists and their organizations are listed.

Following the zoom expert panel, all participants filled out 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 UTI 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 UTI measure?

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

Finally, we solicited patient feedback through a Patient Engagement Panel in order to understand patient perspectives on the inappropriate diagnosis of UTI measure. 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 an infection 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 UTI 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 UTI with inappropriate diagnosis of CAP at the hospital level.

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

We also assessed the association of inappropriate diagnosis with antibiotic-associated adverse events. First, we characterized antibiotic use in patients inappropriately diagnosed with UTI using descriptive statistics. Because duration was skewed, we report median (IQR/inter-quartile range) duration of antibiotic therapy.

Next, we compared outcomes in patients inappropriately diagnosed with UTI vs. those who had ASB but were not unnecessarily treated with antibiotics. Outcomes assessed included: 30-day mortality, 30-day hospital readmission, 30-day emergency department visit, discharge to post–acute care settings, *Clostridioides difficile* infection at 30 days, and duration of hospitalization after urine testing. The association of inappropriate diagnosis with outcomes was assessed using logistic generalized estimating equation models, inverse probability of treatment weighted by baseline covariates identified to be significant in the bivariate and/or multivariate analysis, and other factors potentially associated with the outcome.

The results of this analysis were published in *JAMA Internal Medicine* in 2019 and are also shown below.[3]

[3] Petty LA, Vaughn VM, Flanders SA, et al. Risk Factors and Outcomes Associated With Treatment of Asymptomatic Bacteriuria in Hospitalized Patients. *JAMA Intern Med.* 2019. doi:10.1001/jamainternmed.2019.2871. PCMID: PMC6714039.

2b.03 NQF question: Provide the statistical results from validity testing.

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

In 2021, 50 cases were chronologically selected for detailed audit. Overall data element abstraction accuracy was 98.6%. When errors found through the data audit were corrected, there were two changes in case classification.

Table 3. Accuracy of abstractor vs auditor classification

Abstractor Classification (original)	Auditor Classification (updated)	Number (n=50)
Inappropriate Diagnosis of UTI	Inappropriate Diagnosis of UTI	14
UTI	UTI	34
Inappropriate Diagnosis of UTI	UTI	1
UTI	Inappropriate Diagnosis of UTI	1

Two cases changed classification due to discrepancies noted in audit. Thus, the IRR or Kappa was 0.91 (95% CI : 0.78 – 1.00) indicating strong to "almost perfect" reliability.

Alt-text for Table 3: A series of 50 cases selected for detailed audit resulted in agreement between abstractors and auditors in 48/50 cases (34/35 UTI and 14/15 Inappropriate Diagnosis of UTI cases).

E. Encounter-level Validity: Structured Implicit Case Review

In 2020, 25 cases of inappropriate diagnosis of UTI underwent structured implicit case review by 2-4 physicians. **In 92% of cases (23/25) there was 100% agreement by reviewers that the cases represented 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 feedback during this case review process, we made minor refinements to our measure specifications including refining our inclusion definitions. Specifically, two groups of patients would no longer be included: a) those who were never treated for a UTI even if symptomatic (because they are not inappropriately diagnosed), b) those who received antibiotics only outside of our symptom collection window (symptoms may have occurred later). We also added "hypogastric" as a synonym for "suprapubic" to ensure hypogastric pain was included as a UTI symptom.

F. Face Validity: Feedback from HMS hospitals (N=40 hospitals)

We systematically assessed the perceived validity (after finalization of measure specifications) of the inappropriate diagnosis of UTI measure by soliciting feedback from all participating HMS hospitals (N=40 hospitals) via the following question: "Approximately, what percentage of cases called ASB by HMS do you agree are inappropriately diagnosed with ASB (0-100%)." All hospitals (40/40) responded. Respondents were local leaders or quality champions for the measures.

Median: 90% Inter-quartile range: 80% to 97%

G. Face Validity: National Expert Panel Feedback

Based on conversations held during our two-week online TEP, the 11 national experts who attended our TEP generally agreed with the face validity and operationalization of the overdiagnosis of UTI measure as currently specified. They believed that patients we identified as being inappropriately diagnosed were, in fact, inappropriately diagnosed. There were also some concerns about the use of the word "over-diagnosis" in the measure name. As a result, we changed the measure name to "inappropriate diagnosis" of UTI. There were no changes to measure specifications suggested by the TEP.

TEP Survey results:

Table 4. Distribution of TEP responses to ***Question #1***: "*The inappropriate diagnosis of UTI measure as specified can be used to distinguish between better and worse quality hospitals*."

Rating	# of Responses (N=11)	Percent (%)	Cumulative Percent (%)
5 (Strongly agree)	1	9.1%	9.1%
4 (Agree)	8	72.7%	81.8%
3 (Neutral)	1	9.1%	90.9%
2 (Disagree)	0	0.0%	90.9%
1 (Strongly disagree)	1	9.1%	100.0%

We measured agreement on a 5-point scale Likert scale: 1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree

Alt-text for Table 4: The majority (81.8%) of experts on the TEP responded “Agree” or “Strongly agree” (8/11 and 1/11, respectively). There was one response each for “Neutral” and “Strongly disagree”.

Table 5. TEP responses to ***Question #2*****.** “*What additional data would you like to see captured related to the inappropriate diagnosis of UTI? (free text)*” N=11 respondents (free text question)

% of Responses N=11	Response	Our Action/Response to Comment
72.3% (8/11)	None or N/A	None. Confirmed validity of measurement.
9.1% (1/11)	Duration of Antibiotic Treatment	Added data on duration of antibiotic treatment for patients inappropriately diagnosed with UTI to measure submission. **Patients inappropriately diagnosed with UTI received a median (IQR) 7 (4-9) antibiotic days, all of which were unnecessary.[3]**
9.1% (1/11)	Balancing Measure	Added additional resources on studies of underdiagnosis to measure submission
9.1% (1/11)	Length of stay data	Added data on length of stay for patients inappropriately diagnosed with UTI to measure submission. **Patients inappropriately diagnosed with UTI has a median (IQR) length of stay of 5 (4-7) days.** Compared to patients with ASB not treated with antibiotics, **patients inappropriately diagnosed with UTI had a longer duration of hospitalization after urine testing** (4 vs. 3 days, adjusted relative risk 1.37).[3]

Atl-text for Table 5. The majority (72.3%) of experts on the TEP indicated that no additional data were needed. Suggestions from 3 TEP panelists (1 each) included: a) duration of antibiotic treatment, b) balancing measure, and c) length of stay data. We addressed each of these in our measure submission.

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.”**

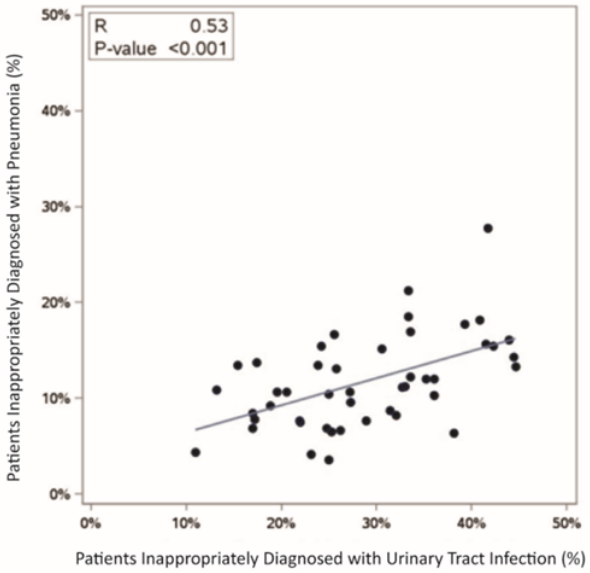
"If you were to have a measure of more correct diagnosis and incorrect diagnosis, and I would do it on the hospital scale, … I feel like if you were to get the correct diagnosis… I would automatically assume that you are getting the correct dose of medicine."

"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."

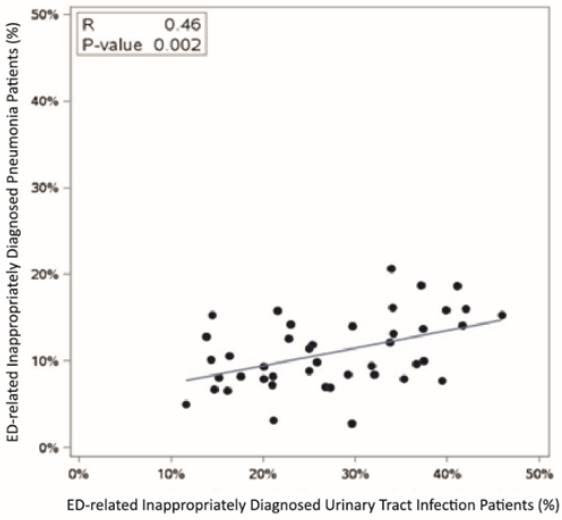
I. Empirical Validity: Association with Other Measures of Diagnostic Quality

To address whether inappropriate diagnosis of UTI was correlated with other domains of quality, we assessed whether inappropriate diagnosis of UTI (as currently specified) was related to inappropriate diagnosis of CAP. This manuscript was published in *BMJ Quality & Safety*.[4] 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 UTI is moderately correlated with inappropriate diagnosis of CAP at the hospital level:



Alt-text for above figure: The percent of patients with inappropriate diagnosis of UTI (N=10,398) is moderately correlated with the percent of patients with inappropriate diagnosis of CAP (N=14,085) 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.



Alt-text for above figure: 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 UTI with Outcomes

There are three main harms associated with inappropriate diagnosis of UTI: delayed time to true diagnosis, antibiotic-associated adverse events, and antibiotic resistance.

In a paper published in *JAMA Internal Medicine*, we analyzed outcomes associated with antibiotic treatment in 2,733 hospitalized patients with ASB (i.e., inappropriate diagnosis of UTI).[3] Patients inappropriately diagnosed with UTI were treated with a median (IQR) 7 (4-9) days of antibiotic therapy, all of which was unnecessary.

Outcomes of patients inappropriately diagnosed vs. those who had ASB and did not receive antibiotics are shown in the table below. Notably, patients inappropriately diagnosed with UTI who were treated with antibiotics had an ~1 day longer length of stay after date of urine testing than those who were not treated with antibiotics (aRR: 1.37 [1.28-1.47]).

Table 6. Outcomes for Treatment vs No Treatment for Asymptomatic Bacteriuria (N = 2733)

Outcome[a]	Antibiotics (n=2259)	No Antibiotics (n=474)	Unadjusted Odds Ratio (95% CI)	Unadjusted *P* Value	Adjusted Odds Ratio (95% CI)	Adjusted *P* Value
30-d Postdischarge mortality[b], N (%)	63 (2.8)	11 (2.3)	1.22 (0.66-2.26)	0.53	1.34 (0.72-2.49)	0.35
30-d Postdischarge readmission[b], N (%)	362 (16.0)	66 (13.9)	1.16 (0.87-1.56)	0.31	1.29 (0.92-1.81)	0.14
30-d Postdischarge ED Visit[b], N (%)	272 (12.0)	62 (13.1)	0.91 (0.70-1.18)	0.48	0.90 (0.66-1.24)	0.52
Discharge to post-acute care facility[b,c], N (%)	811 (35.9)	102 (21.5)	1.98 (1.58-2.48)	<0.001	1.19 (0.90-1.57)	0.22
Clostridioides difficile infection[d], N (%)	14 (0.6)	2 (0.4)	1.39 (0.41-4.68)	0.59	0.88 (0.20-3.86)	0.86
Duration of hospitalization, median (IQR) d[e]	4 (3-6)	3 (2.5)	1.37 (1.28-1.47)[f]	<0.001	1.37 (1.28-1.47)[f]	<0.001

Alt-text for Table 6. Analysis of 2,733 patients inappropriately diagnosed with UTI and treated with antibiotics had an ~1 day longer length of stay after date of urine testing than those who were not treated with antibiotics (aOR:1.37 [1.28-1.47]).

2b.04 Question: **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?)**

The validity of the inappropriate diagnosis of UTI 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 TEP members, patients, and end-users (hospitals); (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. As shown above, 82% of TEP members agreed with the statement: "The inappropriate diagnosis of UTI measure as specified can be used to distinguish between better and worse quality hospitals."

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 UTI 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), the low effect of abstraction errors on case classification, and by the long-standing general agreement by hospital experts with case classification during data feedback.

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 UTI measure and measures of inappropriate diagnosis of other infections, namely CAP. As expected, we found hospitals that performed worse on one measure also performed worse on the other. Thus, the inappropriate diagnosis of UTI 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 UTI and outcomes.