Specifications

Descriptive Information

De.1. Measure Type *(Patient-reported outcomes include HRQoL/functional status, symptom/burden, experience with care, health-related behavior.)**

undefined

De.2. Measure Title - *Measure titles should be concise yet convey who and what is being measured (see* What Good Looks Like*)**
Monthly Hemoglobin Measurement for Pediatric Patients

De.3. Brief description of 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*)
Percentage of patient months of pediatric (< 18 years old) in-center hemodialysis, home hemodialysis, and peritoneal dialysis patients who have monthly measures for hemoglobin during the reporting period.

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?

Measure Specifications

S.1. Measure-specific Web Page *(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.)*
N/A

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

- ○ This is an eMeasure
- ○ X This is not an eMeasure

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff. Provide descriptors for any codes. Use one file with multiple worksheets as needed.)*

X Available in attached Excel or csv file

No data dictionary/code table – all information provided in the submission form

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

S.2d. If this is an instrument-based measure, please indicate responder.
- ○ Patient
- ○ Family or other caregiver
- ○ Clinician
- ○ Not an instrument-based measure

N/A

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

Yes

No

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

There have been no changes to the specifications since the last endorsement in 2015.

S.4. Numerator Statement *(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.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Number of patient months of pediatric (< 18 years old) in-center hemodialysis, home hemodialysis, and peritoneal dialysis patients with a measurement of hemoglobin during the reporting period. The hemoglobin value reported for the end of each reporting month (end-of-month hemoglobin) is used for the calculation.

S.5. Numerator Details *(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 S.2b)*

IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk- adjusted outcome should be described in the calculation algorithm (S.14).
The numerator will be determined by counting all patient months in the denominator that include values for 'Hemoglobin' and 'Hemoglobin Collection Date.' A valid hemoglobin value is defined as between 5-20 g/dL

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

IF an OUTCOME MEASURE, state the target population for the outcome. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
All patient months for pediatric (less than 18 years old) in-center hemodialysis, home hemodialysis, and peritoneal dialysis patients under the care of the dialysis facility for the entire reporting month.

S.7. Denominator Details *(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 S.2b)*

IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome
A treatment history file is the data source for the denominator calculation used for the analyses supporting this submission. This file provides a complete history of the status, location, and dialysis treatment modality of an ESRD patient from the date of the first ESRD service until the patient dies or the data collection cutoff date is reached. For each patient, a new record is created each time he/she changes facility or treatment modality. Each record represents a time period associated with a specific modality and dialysis facility. CROWNWeb is the primary basis for placing patients at dialysis facilities and dialysis claims are used as an additional source of information in certain situations. Information regarding first ESRD service date, death, and transplant is obtained from CROWNWeb (including the CMS Medical Evidence Form (Form CMS-2728) and the Death Notification Form (Form CMS-2746)) and Medicare claims, as well as the Organ Procurement and Transplant Network (OPTN).

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
Exclusions that are implicit in the denominator definition include all patients >=18 years and those who have not been in the facility the entire reporting month (transient patients). There are no additional exclusions for this measure.

S.9. Denominator Exclusion Details *(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 S.2b)*
N/A

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including 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 with at S.2b)*
N/A

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)
N/A

S.12. Type of score:
Rate/proportion

S.13. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. Calculation Algorithm/Measure Logic *(Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period of data, aggregating data; risk adjustment; etc.)*

Patients are included in the facility calculation if "Admit Date" to the specified facility is prior or equal to the first day of the study period, AND the patient has not been discharged ("Discharge Date" is null or blank), OR "Discharge Date" from the facility is greater than or equal to the last day of the study period. The patient's age will be determined by subtracting the patient's date of birth from the first day of the reporting month. All in-center HD, home HD, and PD patients under the facility's care for the entire calendar month and are less than 18 years of age will be included in the denominator. The numerator will be determined by counting all patients in the denominator who have values for 'Hemoglobin' and 'Hemoglobin Collection Date.'

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

N/A

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)* Also, specify calculation of response rates to be reported with performance measure results.

N/A

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED). If other, please describe in S.18.*

Claims ☐ Management Data

☐ Electronic Health Data

☐ Assessment Data X Registry data

☐ Paper Medical Records

☐ Electronic Health Records ☐ Instrument-based data

☐ Other

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s); and standard methods, modes, and languages of administration.
CROWNWeb

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix)*

○ Available at measure-specific web page URL identified in S.1

○ Available in attached appendix at A.1

X No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

☐ Other

☐ Integrated Delivery System

☐ Clinician : Individual

☐ Population : Community, County or City

☐ Clinician : Group/Practice

☐ Population : Regional and State

X Facility

☐ Health Plan

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

☐ Emergency Department and Services

☐ Outpatient Services

☐ Inpatient/Hospital

☐ Post-Acute Care

☐ Home

Other: Dialysis Facility

S.22. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules,or calculation of individual performance measures if not individually endorsed.)*

N/A

Importance

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three sub-criteria must be met to pass this criterion. See guidance on evidence.

Evidence (Measure evaluation criterion 1a)

1a. Attach evidence submission form (Click here to download Evidence Submission Form Template)

1a.1. For maintenance of endorsement:

Is there new evidence about the measure since the last update/submission?

Please use the most current version of the evidence attachment (V7.1). Please use red font to indicate updated evidence.

Yes

No

Performance Gap - Opportunity for Improvement (Measure evaluation criterion 1b)

1b.1. Briefly explain the rationale for this measure (e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

Prior studies show a high prevalence of anemia in the pediatric ESRD population. Studies suggest that among Chronic Kidney Disease (CKD) pediatric patients, anemia is associated with adverse outcomes including increased mortality risk and hospitalizations. Therefore, routine measurement of hemoglobin levels and early management of anemia if present, are critical in this population.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. (This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, 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 (4b1) under Usability and Use.

Based on 2017 CROWNWeb clinical data (Jan-Dec), there were 62 facilities with at least 11 eligible pediatric patients. The mean and median performance scores were 90% and 92%, respectively. The 25th percentile was 88.6% and the 75th percentile was 97.0%.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, 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.
N/A

1b.4. 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. (This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) 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 (4b1) under Usability and Use.
Disparity analyses were performed among the entire eligible pediatric population (n=1445) to examine the difference in performance scores by sex, race, ethnicity and insurance status.

In particular, for each facility, the percent of patient-months by demographic group (sex, race, ethnicity, insurance status) was calculated. Then, the facilities were divided into tertiles (Q1-Q3) based on the percentage of patient-months in the particular demographic category (i.e., a facility with percentage of females similar to the national median will be included in tertiles 3). The top 33.3% of facilities in terms of rank, based on the percentages of females, were classified as Q3, while the bottom 33.3% of facilities were classified as Q1. Average (mean) performance for the measure was calculated for each tertile, and the means were examined for trend across tertiles (Q1-Q3).

The mean performance scores for percent of patient-months with a pediatric hemoglobin measurement in each tertile, by demographic group, are presented below. Males, non-Black, non-White, non-Hispanic, non-Dual eligibility, are the respective reference categories. Based on the small sample size, we do not believe that the following results suggest a meaningful trend.

Range of Facility Level Tertiles by Population Group (Tertile 1-3):
Females (Q1=91.4%, Q2=90.0%, Q3=89.2%)
Black (Q1=91.36, Q2=90.0%, Q3= 89.4%)
White (Q1=88.5%, Q2=90.1%, Q3=92.0%)
Hispanic (Q1=91.4%, Q2=86.7%, Q3=92.6%)
Dual eligibility (Q1=88.7%, Q2=90.4%, Q3=91.4%)

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, 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 1b.4.
N/A

Scientific Acceptability

Testing Attachment

2. Attach measure testing form (Click to here to download the Measure Testing Form Template OR the Composite Measure Testing Form.)

2.1. For maintenance of endorsement:

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

- Yes
- No

2.2. For maintenance of endorsement:

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

- Yes
- No

2.3. For maintenance of endorsement:

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1, 2b3, and 2b5 in the Testing attachment and S.10 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

- Yes -Updated information is included.
- No - This measure is not risk-adjusted.

Feasibility

Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement.

Data Elements Generated as Byproduct of Care Processes (Measure evaluation criterion 3a)

3a.1. How are the data elements needed to compute measure scores generated? (Check all that apply)

Data used in the measure are:

X Generated "or collected" by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, "depression score")

☐ Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

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

☐ Other

Electronic Sources (Measure evaluation criterion 3b)

3b.1. To what extent are the specified data elements available electronically in defined fields (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for maintenance of endorsement.

○ ALL data elements are in defined fields in electronic health records (EHRs)

○ ALL data elements are in defined fields in electronic claim

○ ALL data elements are in defined fields in electronic clinical data (e.g., clinical registry, nursing home MDS, home health OASIS)

X ALL data elements are in defined fields in a combination of electronic sources

○ Some data elements are in defined fields in electronic sources

○ No data elements are in defined fields in electronic sources ○
Patient/family reported information (may be electronic or paper)

3b.2. 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 other than electronic sources. For maintenance of endorsement, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
N/A

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.
N/A

Data Collection Strategy (Measure evaluation criterion 3c)

3c.1. Required for maintenance of endorsement. 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.

IF instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Data collection is accomplished via CROWNWeb, a web-based and electronic batch submission platform maintained and operated by CMS contractors. While this measure is not currently publically reported, measures reported on DFC that are based on CROWNWeb data are reviewed on a regular basis by dialysis facility providers and rare instances of inaccurate or missing data are present based on comments reported in the DFC ticketing system.

3c.2. Describe 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)?
N/A

Usability and Use

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 performance improvement.

4a. Use
4.1. Current and Planned Use *(check all the current and planned uses; for any current uses that are checked, provide a program name and URL for the specific program)*

Intended Use	Specific Plan for Use	Current Use	For current use, provide Program Name and URL
a. Public Reporting	X		Dialysis Facility Compare http://www.medicare.gov/dialysisfacilitycompare/
b. Public Health/Disease Surveillance			
d. Regulatory and Accreditation Programs			
e. Professional Certification or Recognition Program			
f. Quality Improvement with Benchmarking (external benchmarking to multiple organizations)			
g. Quality Improvement (Internal to the specific organization)			

h. Not in use	
i. Use Unknown	

Accountability/Transparency (measure evaluation criterion 4a1)

4a1.1. For each CURRENT use, checked above (update for maintenance of endorsement), provide:

Name of program and sponsor Purpose

Geographic area and number and percentage of accountable entities and patients included Level of measurement and setting

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation*?)

The measure is not currently used in a CMS program at this time. CMS will decide if this measure is appropriate for inclusion in a public reporting program

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specific timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

N/A

Feedback on the measure by those being measured or by others (measure evaluation criterion 4a2)

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

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.

N/A

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

N/A

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

N/A

4a2.2.2. Summarize the feedback obtained from those being measured.

N/A

4a2.2.3. Summarize the feedback obtained from other users.

N/A

4a2.3. Describe how the feedback described in 4a2.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not

N/A

4b. Usability

Improvement (measure evaluation criterion 4b1)

4b1. Refer to data provided in 1b 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, what are the reasons? 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.

The measure is not currently publically reported, so improvement data are not available. Public reporting of this measure would encourage facilities to comply with the Kidney Disease Outcomes Quality Initiative (KDOQI) Clinical Practice Recommendations, which recommends that pediatric patients have monthly measurement of hemoglobin if they are treated with erythropoiesis-stimulating agents (ESAs).

Unexpected findings (measure evaluation criterion 4b2)

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.
N/A

4b2.2. Please explain any unexpected benefits from implementation of this measure.

N/A

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.

Relation to Other NQF-endorsed® Measures (Measure evaluation criterion 5)

5. Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures. (Can search and select measures.)

Yes X No

N/A

Harmonization of Related Measures (Measure evaluation criterion 5a)

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

Yes

No

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.
N/A

Competing Measure(s) (Measure evaluation criterion 5b)

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1424
Measure Title: Monthly Hemoglobin Measurement for Pediatric Patients
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title
Date of Submission: 4/2/2019

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one

step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)
Outcome

☐ Outcome: Click here to name the health outcome

☐ Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☒ Process: measurement of hemoglobin values for pediatric patients

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

1a.2 LOGIC MODEL Diagram or 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.

Large scale clinical trials have not been conducted in the pediatric population, however smaller scale observational and cohort studies have shown an association between anemia and poor outcomes including poor quality of life, cardiovascular disease, morbidity, and mortality.

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

N/A

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

N/A

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of 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. (IOM)

X Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Clinical Practice Guidelines for Anemia in Chronic Kidney Disease: 2006 KDOQI CPR for Pediatrics 3.1: Using ESAs
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	3.1.1.1 In the opinion of the Work Group, the frequency of hemoglobin monitoring in patients treated with ESAs should be at least monthly. Please note that these are clinical practice recommendations and are therefore defined in the KDOQI document as "expert opinion" based recommendations.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	N/A
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	The 2006 KDOQI CPR 3.1 was based on Work Group consensus; it was not graded using a formal grading system.
Provide all other grades and definitions from the recommendation grading system	N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	N/A
Estimates of benefit and consistency across studies	N/A
What harms were identified?	N/A

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The clinical TEP reviewed the body of evidence available for pediatric anemia measurement. An analysis of patients <18 years of age in the North American Pediatric Renal Transplant Cooperative Study (NAPRTCS) found that 68% of patients were anemic. Additionally, results showed anemia was associated with a 52% higher risk of death [1]. An observational study of adolescents on hemodialysis showed decreased risk for death (HR: 0.31; 95% CI: 0.14, 0.65) among patients with hemoglobin 11-12 compared to those with hemoglobin<10 [2]. Results from a small observational study of hemodialysis and peritoneal dialysis patients age 20 months to 22 years showed patients with severe left ventricular hyptertrophy (LVH) had a significantly lower hemoglobin (p=0.027) compared to those without LVH [3]. Finally, a small observational study found a significant association between anemia and lower quality of life among adolescent patients with CKD [4]. Additionally, the 2006 KDOQI Clinical Practice Recommendation (CPR) for pediatric patients states that hemoglobin should be measured at least monthly in patients treated with ESAs. 1. Warady B Ho M. Morbidity and mortality in children with anemia at initiation of dialysis. Pediatr Nephrol 18:1055-1062, 2003. 2. Amaral S, Hwang W, Fivush B, Neu A, Frankenfield D, Furth S. Association of mortality and hospitalization with achievement of adult hemoglobin targets in adolescents maintained on hemodialysis. J Am Soc Nephrol 17:2878-85, 2006. 3. Mitsnefes MM, Daniels SR, Schwartz SM, Meyer RA, Khoury P, Strife CF. Severe left ventricular hypertrophy in pediatric dialysis: Prevalence and predictors. Pediatr Nephrol 14:898-902, 2000. 4. Gerson A, et al. Anemia and health-related quality of life in adolescents with chronic

	kidney disease. Am J Kidney Dis 44:1017-1023, 2004.

_____

1a.4 OTHER SOURCE OF EVIDENCE

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

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

N/A

1a.4.2 What process was used to identify the evidence?

N/A

1a.4.3. Provide the citation(s) for the evidence.

N/A

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 1424
Measure Title: Monthly Hemoglobin Measurement for Pediatric Patients
Date of Submission: 1/7/2019
Type of Measure:

☐ Outcome (*including PRO-PM*)	☐ Composite – ***STOP – use composite testing form***
☐ Intermediate Clinical Outcome	☐ Cost/resource
☒ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	

Instructions

- 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.
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at 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 version 7.1 of the Measure Testing Attachment.

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.

2a2. Reliability testing [10] 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 [11] 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; [12]

AND
If patient preference (e.g., informed decisionmaking) 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). [13]

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 nonresponders) and how the specified handling of missing data minimizes bias.

Notes
10. 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).
11. 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.
12. 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.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions.
15. 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.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.17*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☒ registry	☒ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. 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*).

CROWNWeb

For the Spring 2019 Maintenance submission, CROWNWeb and Medicare claims data was used.

1.3. What are the dates of the data used in testing? January 2013 to December 2013

For the Spring 2019 Maintenance submission, data from January – December 2017 was used.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.20*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were 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)

Fifty-nine facilities that had at least 11 eligible pediatric patients during January 2013 – December 2013 were included in the analyses. Public reporting of this measure on DFC or in the ESRD QIP would be restricted to facilities with at least 11 eligible patients for the measure. We have applied this restriction to all the reliability and validity testing reported here.

For the 2019 Maintenance submission, 62 facilities that had at least 11 eligible pediatric patients during January 2017 – December 2017 were included in the analyses.

Public reporting of this measure on DFC or in the ESRD QIP would be restricted to facilities with at least 11 eligible patients for the measure to comply with restrictions on reporting of potentially patient identifiable information related to small cell size. We have applied this restriction to all the reliability and validity testing reported here.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

There were a total of 1,978 eligible pediatric patients. Among those patients, the average age was 10.5 years, 56% of patients were male, 63% were white, 28% were black, 5% were Asian/Pacific Islander, 29% were Hispanic, and 26% had glomerulonephritis as the primary cause of ESRD.

A subset of 1,280 pediatric patients who belonged to the facilities that had at least 11 eligible pediatric patients were included in the testing and analyses.

For the Spring 2019 Maintenance submission, there were a total of 1,924 eligible pediatric patients. Among those patients, the average age was 10.3 years, 58% of patients were male, 66% were white, 27% were black, 5% were Asian/Pacific Islander, 28% were Hispanic, and 31% had glomerulonephritis as the primary cause of ESRD.

A subset of 1,415 pediatric patients who belonged to the facilities that had at least 11 eligible pediatric patients were included in the testing and analyses. Among those patients, the average age was 9.8 years, 57% of patients were male, 68% were white, 25% were black, 5% were Asian/Pacific Islander, 31% were Hispanic, and 30 % had glomerulonephritis as the primary cause of ESRD.

1.7. 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 reported below.

N/A

1.8 What were 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.

N/A

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level 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*)

We used January 2013 – December 2013 CROWNWeb data to calculate facility level monthly and annual performance scores. Fifty-nine facilities that had at least 11 eligible patients were included in the testing. There were a total of 1,280 patients.

We assessed reliability by calculating facility-level Pearson correlation coefficients between the current performance month and the preceding month for reporting months during January 2013 – December 2013.

In addition, we calculated inter-unit reliability (IUR) for each reporting month and the overall 12 months. The monthly based measure was a simple average across individuals in the facility. The NQF-recommended approach for determining measure reliability is a one-way analysis of variance (ANOVA), in which the between and within facility variation in the measure is determined. The inter-unit reliability (IUR) measures the proportion of the measure variability that is attributable to the between-facility variance. The yearly based measure, however, is not a simple average and we instead estimate the IUR using a bootstrap approach, which uses a resampling scheme to estimate the within facility variation that cannot be directly estimated by ANOVA. We note that the method for calculating the IUR was developed for measures that are approximately normally distributed across facilities. Since this measure is not normally distributed, the IUR value should be interpreted with some caution.

For the Spring 2019 Maintenance submission, we used January 2017 – December 2017 CROWNWeb data to calculate facility level monthly and annual performance scores. Sixty-two facilities that had at least 11 eligible patients were included in the testing. There were a total of 1,415 patients.

We assessed reliability by calculating facility-level Pearson correlation coefficients between the current performance month and the preceding month for reporting months during January 2017 – December 2017. We also calculated inter-unit reliability using the same methodology as the previous submission.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The Pearson correlation coefficients of each pair of the current and the preceding months ranged from 0.78 to 0.98. All were statistically significant ($p<0.0001$), indicating this measure is reliable over time.

For reliability we calculated the monthly and annual IUR across the 12 reporting months. As explained above, the method for calculating the IUR was developed for measures that are approximately normally distributed across facilities. IUR=0.96, which is high and suggests 96% of variation in the measure is attributed to between facility variation.

For the Spring 2019 Maintenance submission, the median of Pearson correlation coefficients of each pair of the current and the preceding months was 0.84, with a range of 0.28 to 0.92. All were statistically significant ($p<0.05$), indicating this measure is reliable over time.

In a sensitivity analysis, we identified an outlier facility that drastically influenced the correlation between February and March. After removing the outlier facility, the range of correlation coefficients of each pair of the current and the preceding months narrowed down to be from 0.47 and 0.92, with a median of 0.85. All correlations were statistically significant ($p<0.0001$),

For reliability we calculated the monthly and annual IUR across the 12 reporting months. As explained above, the method for calculating the IUR was developed for measures that are approximately normally distributed across facilities. IUR=0.82, which is high and suggests 82% of variation in the measure is attributed to between facility variation.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The Pearson correlation coefficients were moderate to strong, indicating this measure is reliable over time periods of measurement.

The IUR suggest this measure is reliable. However, since the distribution of performance scores is skewed, the IUR value should be interpreted with some caution.

For the Spring 2019 Maintenance submission, the Pearson correlation coefficients were moderate to strong, indicating this measure is reliable over time periods of measurement.

The IUR again suggests this measure is reliable. However, since the distribution of performance scores is skewed, the IUR value should be interpreted with some caution.

_____

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

- ☐ **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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

2b1.2. 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)*

We used January 2013 – December 2013 CROWNWeb data to calculate facility level monthly and annual performance scores. Fifty-nine facilities that had at least 11 eligible patients and included 1,280 patients in total were included in the testing.

We computed the Spearman correlation to assess the association between the annual performance scores and the NQF endorsed (0369) standardized mortality ratio (SMR) using the 2013 SMR. The data source and the methodology in SMR calculations are attached.

This measure is being maintained on the basis of face validity. The measurement of hemoglobin as a dialysis quality measure was initially developed and approved by a Clinical TEP, which agreed that this quality measure is important in the assessment of the quality of care for pediatric dialysis patients.

For the Spring 2019 Maintenance submission, we used January 2017 – December 2017 CROWNWeb data to calculate facility level monthly and annual performance scores. Sixty-two facilities that had at least 11 eligible patients and included 1,415 patients in total were included in the testing.

We computed the Spearman correlation to assess the association between the annual performance scores and the NQF endorsed (0369) standardized mortality ratio (SMR) using the 2017 SMR. We expect the correlation to be negative, insofar as facilities with successful processes for monitoring clinically important intermediate outcomes of care would be expected to have better primary outcomes, including lower mortality. In addition, if consistent, effective monitoring of hemoglobin outcomes contributes to a higher percentage of hemoglobin values in the facility's target range, then patient primary outcomes would likely be improved (lower mortality), assuming the facility's target ranges were appropriate.

This measure is being maintained on the basis of face validity. The measurement of hemoglobin as a dialysis quality measure was initially developed and approved by a Clinical TEP, which agreed that this quality measure is important in the assessment of the quality of care for pediatric dialysis patients.

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Spearman correlation coefficient was -0.20, p=0.13.

For the Spring 2019 Maintenance submission, the spearman correlation coefficient was 0.07, p=0.55.

2b1.4. What is 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 result suggests that facilities with a higher percentage of pediatric patients (calculated as patient months) with hemoglobin measured is associated with a lower risk of mortality relative to facilities with a lower percentage of pediatric patients with hemoglobin measured. The result is however not statistically significant. This may be due to the small sample size (e.g., small number of facilities eligible for this measure).

The result does not suggest a statistically significant association between the measure (calculated as patient months) and mortality. We maintain the measure based on face validity as determined by the Technical Expert Panel that initially developed the measure.

_____

2b2. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b3*

2b2.1. Describe the method of testing exclusions and what it tests (*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*)

N/A

2b2.2. What were 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*)
N/A

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., 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*)
N/A

_____

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

2b3.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.
N/A

2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.
Risk adjustment is not required for this measure. While trend analysis suggests by disparities by race, the conservative interpretation would be that these differences reflected in the trend analysis reflect disparities in care for certain subpopulations. In the absence of biological effects explaining these differences, risk adjustment for these factors would potentially mask disparities in care.

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
Also discuss any "ordering" of risk factor inclusion; for example, are social risk factors added after all clinical factors?
N/A

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:

- ☐ **Published literature**
- ☐ **Internal data analysis**
- ☐ **Other (please describe)**

2b3.4a. What were the statistical results of the analyses used to select risk factors?
N/A

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**
N/A

2b3.5. 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 controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

N/A

2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
N/A

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
N/A

2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
N/A

2b3.9. Results of Risk Stratification Analysis:

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

N/A

2b3.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

_____

2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b4.1. 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 related to performance gap in 1b)*

Given that the number of facilities included in the calculation in 1.5 is small, the sample was determined to be too small to display useful data on meaningful differences.

For the Spring 2019 submission, we followed the methodology described below to using data from January – December 2017.

Testing was performed on the yearly based performance score. We used a re-sampling based exact test, with re-sampling generated from the population distribution of the patient level outcomes. Note that a one-sided test with significance level 0.025 is used (corresponding to cutoff=0.05 in a two-sided test) due to non-symmetric structure of the measure's distribution. To calculate the p-value, we compute the probability that the facility would experience a number of events (i.e., percentage with hypercalcemia) more extreme than that observed if the null hypothesis were true, with the null hypothesis being that the facility's distribution of hypercalcemia will follow the overall national distribution.

2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *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*)

N/A

For the Spring 2019 Maintenance submission, we did not identify statistically significant in performance scores across 62 facilities that had at least 11 eligible patients. This is because of the size of patient population is relatively small within the each facility (mean=23, median=20) and the performance score of measure is distributed uniformly across facilities (mean= 90.2%, standard deviation=11%).

2b4.3. What is 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? (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

N/A

For the Spring 2019 maintenance submission, as described above we did not identify statistically significant differences in performance scores for this measure, due to the small patient population in each facility.

_____

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

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 eMeasures). 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.***

2b5.1. 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; what statistical analysis was used*)

N/A

2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i*.e., what do the results mean and what are the norms for the test conducted*)

N/A

_____

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to 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 nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Missing is the outcome and this measure is reporting the percentage of non-missing. Thus, the missing data does not cause bias in this measure.

2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

N/A

2b6.3. What is 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 nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., 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, provide rationale for the selected approach for missing data*)

N/A