NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3403

Measure Title: Percentage of Prevalent Patients Waitlisted (PPPW)

Measure Steward: Centers for Medicare and Medicaid Services

Brief Description of Measure: This measure tracks the percentage of patients at each dialysis facility who were on the kidney or kidney-pancreas transplant waitlist. Results are averaged across patients prevalent on the last day of each month during the reporting year.

Developer Rationale: A measure focusing on the wait listing process is appropriate for improving access to kidney transplantation for several reasons. First, wait listing is a necessary step prior to potential receipt of a deceased donor kidney. Second, dialysis facilities exert substantial control over the process of waitlisting. This includes proper education of dialysis patients on the option for transplant, referral of appropriate patients to a transplant center for evaluation, assisting patients with completion of the transplant evaluation process, and optimizing the health and functional status of patients in order to increase their candidacy for transplant wait listing. These types of activities are included as part of the conditions for coverage for Medicare certification of ESRD dialysis facilities. In addition, dialysis facilities can also help maintain patients on the wait list through assistance with ongoing evaluation activities and by optimizing health and functional status. Finally, wide regional variations in wait listing rates highlight substantial room for improvement for this process measure [1,2,3].

This measure focuses specifically on the prevalent dialysis population, examining waitlisting status monthly for each patient. This allows evaluation and encouragement of ongoing waitlisting of patients beyond the first year of dialysis initiation who have not yet been listed. Patients may not be ready, either psychologically or due to their health status, to consider transplantation early after initiation of dialysis and many choose to undergo evaluation for transplantation only after years on dialysis. In addition, as this measure assesses monthly waitlisting status of patients, it also evaluates and encourages maintenance of patients on the waitlist. Maintenance of active status on the waitlist is important for increasing likelihood of transplantation [4] and thus by extension, is waitlisting overall. This is an important area to which dialysis facilities can contribute through ensuring patients remain healthy, and complete any ongoing testing activities required to remain on the wait list. In contrast to this measure, another waitlisting measure, the Standardized First Kidney Transplant Waitlist Ratio for Incident Dialysis Patients (SWR), focuses solely on new listing or living kidney donor transplantation within the first year after initiation of dialysis with the rationale of encouraging early access to transplantation or the wait list.

1. Ashby VB, Kalbfleisch JD, Wolfe RA, et al. Geographic variability in access to primary kidney transplantation in the United States, 1996-2005. American Journal of Transplantation 2007; 7 (5 Part 2):1412-1423.

Abstract:

This article focuses on geographic variability in patient access to kidney transplantation in the United States. It examines geographic differences and trends in access rates to kidney transplantation, in the component rates of wait-listing, and of living and deceased donor transplantation. Using data from Centers for Medicare and Medicaid Services and the Organ Procurement and Transplantation Network/Scientific Registry of Transplant Recipients, we studied 700,000+ patients under 75, who began chronic dialysis treatment, received their first living donor kidney transplant, or were placed on the waiting list pre-emptively. Relative rates of wait-listing and transplantation by State were calculated using Cox regression models, adjusted for patient demographics. There were geographic differences in access to the kidney waiting list and to a kidney transplant. Adjusted wait-list rates ranged from 37% lower to 64% higher than the national average. The living donor rate ranged from 57% lower to 166% higher, while the deceased donor transplant rate ranged from 60% lower to 150% higher than the national average. In general, States with higher wait-listing rates tended to have lower transplantation rates and States with lower wait-listing rates had higher transplant rates. Six States demonstrated both high wait-listing and deceased donor transplantation rates while six others, plus D.C. and Puerto Rico, were below the national average for both parameters.

2. Satayathum S, Pisoni RL, McCullough KP, et al. Kidney transplantation and wait-listing rates from the international Dialysis Outcomes and Practice Patterns Study (DOPPS). Kidney Intl 2005 Jul; 68 (1):330-337.

Abstract:

BACKGROUND: The international Dialysis Outcomes and Practice Patterns Study (DOPPS I and II) allows description of variations in kidney transplantation and wait-listing from nationally representative samples of 18- to 65-year-old hemodialysis patients. The present study examines the health status and socioeconomic characteristics of United States patients, the role of for-profit versus not-for-profit status of dialysis facilities, and the likelihood of transplant wait-listing and transplantation rates.

METHODS: Analyses of transplantation rates were based on 5267 randomly selected DOPPS I patients in dialysis units in the United States, Europe, and Japan who received chronic hemodialysis therapy for at least 90 days in 2000. Left-truncated Cox regression was used to assess time to kidney transplantation. Logistic regression determined the odds of being transplant wait-listed for a cross-section of 1323 hemodialysis patients in the United States in 2000. Furthermore, kidney transplant wait-listing was determined in 12 countries from cross-sectional samples of DOPPS II hemodialysis patients in 2002 to 2003 (N= 4274).

RESULTS: Transplantation rates varied widely, from very low in Japan to 25-fold higher in the United States and 75-fold higher in Spain (both P values <0.0001). Factors associated with higher rates of transplantation included younger age, nonblack race, less comorbidity, fewer years on dialysis, higher income, and higher education levels. The likelihood of being wait-listed showed wide variation internationally and by United States region but not by for-profit dialysis unit status within the United States.

CONCLUSION: DOPPS I and II confirmed large variations in kidney transplantation rates by country, even after adjusting for differences in case mix. Facility size and, in the United States, profit status, were not associated with varying transplantation rates. International results consistently showed higher transplantation rates for younger, healthier, better-educated, and higher income patients.

3. Patzer RE, Plantinga L, Krisher J, Pastan SO. Dialysis facility and network factors associated with low kidney transplantation rates among United States dialysis facilities. Am J Transplant. 2014 Jul; 14(7):1562-72.

Abstract:

Variability in transplant rates between different dialysis units has been noted, yet little is known about facility-level factors associated with low standardized transplant ratios (STRs) across the United States End-stage Renal Disease (ESRD) Network regions. We analyzed Centers for Medicare & Medicaid Services Dialysis Facility Report data from 2007 to 2010 to examine facility-level factors associated with low STRs using multivariable mixed models. Among 4098 dialysis facilities treating 305698 patients, there was wide variability in facility-level STRs across the 18 ESRD Networks. Four-year average STRs ranged from 0.69 (95% confidence interval [CI]: 0.64-0.73) in Network 6 (Southeastern Kidney Council) to 1.61 (95% CI: 1.47-1.76) in Network 1 (New England). Factors significantly associated with a lower Standardized Transplantation Ratio(STR) (p<0.0001) included for-profit status, facilities with higher percentage black patients, patients with no health insurance and patients with diabetes. A greater number of facility staff, more transplant centers per 10 000 ESRD patients and a higher percentage of patients who were employed or utilized peritoneal dialysis were associated with higher STRs. The lowest performing dialysis facilities were in the Southeastern United States. Understanding the modifiable facility-level factors associated with low transplant rates may inform interventions to improve access to transplantation.

4. Grams, M. E., Massie, A. B., Schold, J. D., Chen, B. P., & Segev, D. L. (2013). Trends in the inactive kidney transplant waitlist and implications for candidate survival. American Journal of Transplantation, 13(4), 1012-1018.

Abstract

In November 2003, OPTN policy was amended to allow kidney transplant candidates to accrue waiting time while registered as status 7, or inactive. We evaluated trends in inactive listings and the association of inactive status with transplantation and survival, studying 262,824 adult first-time KT candidates listed between 2000 and 2011. The proportion of waitlist candidates initially listed as inactive increased from 2.3% prepolicy change to 31.4% in 2011. Candidates initially listed as inactive were older, more often female, African American, and with higher body mass index. Postpolicy change, conversion from initially inactive to active status generally occurred early if at all: at 1 year after listing, 52.7% of initially inactive candidates had been activated; at 3 years, only 66.3% had been activated. Inactive status was associated with a substantially higher waitlist mortality (aHR 2.21, 95%CI:2.15-2.28, p<0.001) and lower rates of eventual transplantation (aRR 0.68, 95%CI:0.67-0.70, p<0.001). In summary, waitlist practice has changed significantly since November 2003, with a sharp increase in the number of inactive candidates. Using the full waitlist to estimate organ shortage or as a comparison group in transplant outcome studies is less appropriate in the current era.

Numerator Statement: Number of patient months in which the patient at the dialysis facility is on the kidney or kidney-pancreas transplant waitlist as of the last day of each month during the reporting year.

Denominator Statement: All patient-months for patients who are under the age of 75 in the reporting month and who are assigned to the dialysis facility according to each patient's treatment history as of the last day of each month during the reporting year.

Denominator Exclusions: Exclusions that are implicit in the denominator include:

- Patients who were at age 75 or older in the reporting month.
- Patient who were admitted to a skilled nursing facility (SNF) or a hospice during the month of evaluation were excluded from that month; patients who were admitted to a skilled nursing facility (SNF) at incidence or previously according to Form CMS-2728 were also excluded.

Measure Type: Process

Data Source: Claims, Registry Data

Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

Staff Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

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

The developer provides the following evidence for this measure:

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

Evidence Summary

- 2011 American Journal of Transplantation Systematic Review: Kidney Transplantation Compared With Dialysis In Clinically Relevant Outcomes
- A total of 110 studies were included in the review, representing over 1.9 million patients. All studies were either retrospective and/or prospective cohort observational study designs. No randomized clinical trials were available for inclusion.
- Individual studies indicate that kidney transplantation is associated with lower mortality and improved quality of life compared with chronic dialysis treatment.
- Results were not pooled because of expected diversity inherent to observational studies.

Exception to evidence - NA

Questions for the Committee:

- o *What is the relationship of this measure to patient outcomes?*
- o *How strong is the evidence for this relationship?*
- o *Is the evidence directly applicable to the process of care being measured?*
  - ▪ *Note that the evidence presented by the developer pertains primarily to the relationship between transplants and mortality; however, this measure assesses waitlisting of patients, rather than receipt of a transplant itself. Is there a close enough relationshp between waitlisting and receipt of a transplant for this measure to meet the evidence criterion?*

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3) → QQC presented (Box 4) → Quantity: high; Quality: moderate; Consistency: high (Box 5) → Moderate (Box 5b) → Moderate

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

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

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

- After applying all exclusion criteria, the PPPW performance score was evaluated for all dialysis facilities that had at least 11 patients and 2 expected events during 2013-2015.

N	Mean	Standard Deviation	25% Q1	Median	75% Q3	Max
6617	0.21	0.11	0.12	0.19	0.27	0.78

- The developer states the wide variation across facilities suggests there is substantial opportunity for improvement

Disparities

- Estimates, p-values and Hazard Ratios (HR) for race, sex and ethnicity based on the original model, 2013-2015
  - o Sex:
    - ▪ Male (Reference)
    - ▪ Female (estimate = -0.08, p-value = <.001)

- Race:
    - White (Reference)
    - Black (estimate = -0.08, p-value = <.001)
    - Asian (estimate = 0.38, p-value = <.001)
    - Native American (estimate = -0.31, p-value = <.001)
    - Other (estimate = -0.01, p-value = 0.93)
- Ethnicity:
    - Hispanic (Reference)
    - Non-Hispanic/Unknown (estimate = -0.04, p-value = 0.01)
- The developer states that there is evidence of significant differences in measure results by sex, race and ethnicity, however, there is no clear biological rationale for differences in waitlisting on the basis of sex, race or ethnicity to justify a need for adjustment.

Questions for the Committee:

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

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

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence:

- PPPW is a process measure examining transplant wait-listing in prevalent patients receiving dialysis. Wait-listing is a pre-requisite for deceased donor transplantation. The submitted evidence focuses on the outcomes associated with transplantation and does not match what is being measured (e.g., process of wait-listing). Evidence provided under performance gap states the following: ""States with higher wait-listing rates tended to have lower transplantation rates and States with lower wait-listing rates had higher transplant rates."" (Ashby et al. AJT 2007) Per 2017 USRDS ADR: 83,978 candidates on the wait-list and 18,805 kidney transplants in 2015. Note: correlation with STR examined in validity testing
- Evidence is the same as for #3402
- The measure focuses on the prevalent dialysis population and allows evaluation of ongoing waitlisting. It is known that there are geographical differences in access to kidney transplantation and a such it is important to measure the percentage of prevalent waitilisting as transplantation is known to be a better modality for patients for ESRD. The measure does demonstrate the target population outcome.
- Agree with the aim of the goal. However, not sure whether this is fully controllable by dialysis facility. There are multiple factors in addition to education from dialysis facility that may prevent listing, including acceptance by a transplant center, economic status, insurance availability, that may cause patient decision not to list.
- Yes
- This is a new measure. Evidence is not graded but strong quality with 110 studies. This is a measure for waitlisting so there is inferred relationship since waitlisting is the first step to achieving a transplant.
- Process measure. I do think that the committee should discuss carefully whether there is enough evidence that waitlisting in itself has been proven to improve outcomes after accounting for other factors that inform waitlisting.

1b. Performance Gap:

- Facility-level variation: median 0.19 [0.12, 0.27] Regional variation based on the literature. Evidence provided also highlights how 2003 OPTN policy change impacted wait-listing (study by Gram et al. also examined variation in inactive wait-listing); do we understand to what extent the 2014 OPTN policy change has impacted the use of inactive wait-listing by transplant centers?

- Exclusion criteria is not extensive enough and therefore impacts the PPPW performance scores. Needs cancer and drug exclusions.
- There is a gap based upon race, sex and ethnicity across facilities without any biologic rationale for this. Data was submitted for the subgroups.
- Disparities in care are demonstrable. Disparities may also be linked to more economic factors as well.
- Yes
- Wide variation across facilities so this demonstrates an opportunity for improvement. And disparities among race, sex, and ethnicity.
- Yes

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability**;** Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

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

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

Evaluators: Susan White, Michael Stoto, J Matt Austin

Evaluation of Reliability and Validity (and composite construction, if applicable):

Review #1, Review #2, Review #3

Additional Information regarding Scientific Acceptability Evaluation (*if needed*):

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

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

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

Review #1: Scientific Acceptability

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the*** Measure Evaluation Criteria and Guidance document ***(pages 18-24) and the 2-page*** Key Points document ***when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   REFERENCE: "MIF_xxxx" document

   NOTE*: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

   TIPS*: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

3. Was reliability testing conducted with computed performance measure scores for each measured entity?

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

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

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

   ☒Yes (go to Question #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

5. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

   ☒High (go to Question #6)

   ☐Moderate (go to Question #6)

   ☐Low (please explain below then go to Question #6)

   ☐Insufficient (go to Question #6)

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   REFERENCE: Testing attachment, section 2a2.

   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☐Yes (go to Question #7)

   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

   REFERENCE: Testing attachment, section 2a2.2

   ***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

   REFERENCE: Testing attachment, section 2a2

   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

   ☐Insufficient (go to Question #9)

9. Was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

   REFERENCE: testing attachment section 2b1.

   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

   TIP:** You should answer this question <u>ONLY</u> if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

   ☐Yes (go to Question #10 and answer using your rating from <u>data element validity testing</u> – Question #23)

   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and <u>all</u> testing results:

    ☐High (NOTE: Can be HIGH <u>only if</u> score-level testing has been conducted)

    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has <u>not</u> been conducted)

    ☐Low (please explain below) [NOTE: Should rate <u>LOW</u> if you believe specifications are NOT precise, unambiguous, and complete]

    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is <u>not</u> required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

    REFERENCE: Testing attachment, section 2b2-2b6

    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

    ☒Yes (go to Question #12)

    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

12. Analysis of potential threats to validity: Any concerns with measure exclusions?

    REFERENCE: Testing attachment, section 2b2.

***TIPS**: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

ASSESSMENT OF MEASURE TESTING

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

    ☒Yes (go to Question #18)

    ☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

    ☒Yes (go to Question #19)

    ☐No (please explain below, then skip questions #19-20 and go to Question #21)

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

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

    ☒Yes (go to Question #20)

    ☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☒High (go to Question #21)

    ☐Moderate (go to Question #21)

    ☐Low (please explain below then go to Question #21)

    ☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Prior validity studies of the same data elements may be submitted*

    ☐Yes (go to Question #22)

    ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    ***REFERENCE:** Testing attachment, section 2b1.*

    ***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☐Moderate (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

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

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

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Review #2: Scientific Acceptability

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

RELIABILITY

27. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

REFERENCE: "MIF_xxxx" document

***NOTE**: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

***TIPS**: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

☒Yes (go to Question #2)

☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

28. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

29. Was reliability testing conducted with computed performance measure scores for each measured entity?

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

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

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

☒Yes (go to Question #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

31. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

REFERENCE: Testing attachment, section 2a2.2

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?

☐High (go to Question #6)

☒Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

32. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

REFERENCE: Testing attachment, section 2a2.2

***TIPS**: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

34. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

35. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

36. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

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

    ☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

    ☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

    ☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

37. Were potential threats to validity that are relevant to the measure empirically assessed ()?

    REFERENCE: Testing attachment, section 2b2-2b6

    TIPS*: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

    ☒Yes (go to Question #12)

    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

38. Analysis of potential threats to validity: Any concerns with measure exclusions?

    REFERENCE: Testing attachment, section 2b2.

    TIPS*: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

    ☐Yes (please explain below then go to Question #13)

    ☒No (go to Question #13)

    ☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

39. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

    REFERENCE: Testing attachment, section 2b3.

    13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

    13b. Are social risk factors included in risk model? ☐Yes ☒No

    13c. Any concerns regarding the risk-adjustment approach?

    TIPS*: Consider the following:* ***If measure is risk adjusted****: If the developer asserts there is* ***no conceptual basis*** *for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model* *discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for* ***not risk adjusting*** *provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

Unclear why the 'age knots' are different for the various measures submitted by this organization.

I am not questioning the need for an age adjustment, but the write up would be stronger with some explanation of how the age categories were selected.

40. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

41. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

42. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

ASSESSMENT OF MEASURE TESTING

43. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

44. Was validity testing conducted with computed performance measure scores for each measured entity?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

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

REFERENCE: Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

Validity testing based on correlation of PPPW with STR, SMR, SHR, SHR (ED) and SRR. The testing was appropriate. Since 6,000 to 6,400 observations are included, there is a potential for over-powering of the correlation test. For example, the correlation between PPPW and SHR is -0.03 (p<.001) – this may or may not be a practically significant result.

46. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #21)

    ☒Moderate (go to Question #21)

    ☐Low (please explain below then go to Question #21)

    ☐Insufficient (go to Question #21)

47. Was validity testing conducted with patient-level data elements?

    REFERENCE: Testing attachment, section 2b1.

    ***TIPS**: Prior validity studies of the same data elements may be submitted*

    ☐Yes (go to Question #22)

    ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

48. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    ***REFERENCE:** Testing attachment, section 2b1.*

    ***TIPS**: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

    ☐Yes (go to Question #23)

    ☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

49. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

    ☐Moderate (skip Questions #24-25 and go to Question #26)

    ☐Low (please explain below, skip Questions #24-25 and go to Question #26)

    ☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

50. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

51. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

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

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

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Review #3: Scientific Acceptability

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

RELIABILITY

53. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

    REFERENCE: "MIF_xxxx" document

    ***NOTE**: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

    ***TIPS**: Consider the following: Are all the data elements clearly defined? Are all appropriate codes included? Is the logic or calculation algorithm clear? Is it likely this measure can be consistently implemented?*

    ☐Yes (go to Question #2)

    ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

    ONLY BASIC SPECIFICATIONS WERE PROVIED AT THE TIME OF METHODS PANEL EVALUATION

54. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

    REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

    ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

    ☒Yes (go to Question #3)

    ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified **OR** there is no reliability testing (please explain below, skip Questions #3-8, then go to Question #9)

55. Was reliability testing conducted with computed performance measure scores for each measured entity?

    REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2

    ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

    ☒Yes (go to Question #4)

    ☐No (skip Questions #4-5 and go to Question #6)

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

    REFERENCE: Testing attachment, section 2a2.2

    ***TIPS**: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.*

    ☒Yes (go to Question #5)

    ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

57. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

    REFERENCE: Testing attachment, section 2a2.2

    ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

    ☒High (go to Question #6)

☐Moderate (go to Question #6)

☐Low (please explain below then go to Question #6)

☐Insufficient (go to Question #6)

58. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

REFERENCE: Testing attachment, section 2a2.

TIPS*: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

REFERENCE: Testing attachment, section 2a2.2

TIPS*: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements*

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☐Yes (go to Question #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

60. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

REFERENCE: Testing attachment, section 2a2

TIPS*: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

61. Was **empirical VALIDITY testing** of patient-level data conducted?

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP: *You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but* ***check with NQF staff before proceeding, to verify.***

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

62. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and all testing results:

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required, but check with NQF staff]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

63. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

TIPS: *Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

64. Analysis of potential threats to validity: Any concerns with measure exclusions?

REFERENCE: Testing attachment, section 2b2.

TIPS: *Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?*

☐Yes (please explain below then go to Question #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

65. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and “NOT APPLICABLE” is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☒Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☒No

13c. Any concerns regarding the risk-adjustment approach?

TIPS: *Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer’s decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a “clinical model only” if social risk factors are included in the final model? If a measure is NOT risk-adjusted, is a justification for **not risk***

***adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☒No (go to Question #14)

☐Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

66. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

67. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

68. Analysis of potential threats to validity: Any concerns regarding missing data?

REFERENCE: Testing attachment, section 2b6.

☒Yes (please explain below then go to Question #17)

☒No (go to Question #17)

ASSESSMENT OF MEASURE TESTING

69. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).*

☒Yes (go to Question #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

70. Was validity testing conducted with computed performance measure scores for each measured entity?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

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

REFERENCE: Testing attachment, section 2b1.

***TIPS**: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score*

☒Yes (go to Question #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

72. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #21)

    ☒Moderate (go to Question #21)

    ☐Low (please explain below then go to Question #21)

    ☐Insufficient (go to Question #21)

    The correlations were not very different than zero (i.e., no relationship).

73. Was validity testing conducted with patient-level data elements?

    REFERENCE: Testing attachment, section 2b1.

    TIPS: *Prior validity studies of the same data elements may be submitted*

    ☐Yes (go to Question #22)

    ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

74. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    REFERENCE: *Testing attachment, section 2b1.*

    TIPS: *For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.*

    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

    ☐Yes (go to Question #23)

    ☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

75. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

    ☐Moderate (skip Questions #24-25 and go to Question #26)

    ☐Low (please explain below, skip Questions #24-25 and go to Question #26)

    ☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

76. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

    NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

    REFERENCE: Testing attachment, section 2b1.

    TIPS: *Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.*

    ☐Yes (go to Question #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

77. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

REFERENCE: Testing attachment, section 2b1.

TIPS: *Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

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

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

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

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability-Specifications and 2a2. Reliability testing:

- Model is age-adjusted; why is the model not further case-mix adjusted? Exclusions include Age 75 or older, how was the age cut-point selected? IUR 0.80 but no information on IUR by facility size. Did the IUR differ by facility size? (data not provided). Recommend discussion of reliability by committee
- Transplant centers have different exclusion criteria which would impact the reliability of the data.
- Data elements are defined as well as exclusions. Data is reliable and needed.
- No reliability concerns about calculation of measure. Concerns are more about controllability of measure, especially as relates to dialysis facility v. transplant center.
- No
- No concerns about the reliability of the measure.
- Yes

2b1. Validity –Testing and 2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data):

- Correlation coefficient with STR 0.45;
- Correlation coefficients for SMR, SHR, SHR (ED), and SRR range from -0.22 to -0.03;

- SMP commented on whether correlation values were of practical significance and ""not very different than zero (i.e., no relationship)””
- The data shows wide variation across facilities but can be easily attributed to the variations in transplant center criteria.
- I do not have concerns about data and reilability as well as exclusion criteria.
- No concerns
- Since this is a measure of waitlisting, this is a list that is updated to dialysis clinics by the transplant programs so is easy to obtain and easily validated.
- No

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment):

- Recommend discussion of age adjustment approach and whether age-adjustment alone is sufficient for a measure of transplantation. Are exclusions sufficient given known contraindications to transplantation (e.g., recent malignancy, etc.)? Recommend discussion of SDS/SES findings. Recommend discussion of validity by committee.
- I do not have concerns with reliability testing.
- Does denominator of calculation need to be modified to further exclude those ineligible for transplant on permanent or temporary basis by co-morbidity or other health concern?
- No concerns
- The only group of patients excluded are patients over 75 years of age. Adding in hospice or palliative care patients would be appropriate. Also patients who have been turned down by transplant should also not be included.
- As with the other measure, there needs to be further elaboration upon considerations for adjustment of social risk factors as MAP group suggested.

Criterion 3. Feasibility

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

- Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
- ALL data elements are in defined fields in a combination of electronic sources

Questions for the Committee:

o *Are the required data elements routinely generated and used during care delivery?*
o *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
o *Is the data collection strategy ready to be put into operational use?*

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

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3. Feasibility:

- Data elements are routinely generated as part of health care delivery and are electronically available. Proposed strategy should be ready for operational use.
- Data on transplant status on waitlists is generated by the transplant centers - not the facilities. Data may or may not be provided in a timely manner. Patients can be put on hold for a variety of reasons out of the control of the centers.
- Elements are feasible.

- Question whether additional exclusionary data could be easily collected.
- No concerns
- The dialysis clinics already receive a list from transplant centers which of their patients are on transplant list. This will be easily converted to a required data element as yes or no. I am assuming in transplant workup will not be captured.
- Yes

Criterion 4: Usability and Use

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

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

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details The measure has gone through the process of being recommended for Dialysis Facility Compare (DFC), and will go through a Dry Run for DFC in July 2018, with the intention that the measure will be publicly reported in October 2019.

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

Feedback on the measure by those being measured or others: Measure is not currently in use.

Additional Feedback: The NQF Measure Application Partnership (MAP) reviewed this measure during the 2017-2018 Pre-Rulemaking session. MAP acknowledged that this measure addresses an important quality gap for dialysis facilities; however, it discussed a number of factors that should be balanced when implementing this measure. MAP reiterated the critical need to help patients receive kidney transplants to improve their quality of life and reduce their risk of mortality. MAP also noted there are disparities in the receipt of kidney transplants and there is a need to incentivize dialysis facilities to educate patients about wait listing processes and requirements. On the other hand, MAP also acknowledged concerns and public comments about the locus of control of the measure, where dialysis facilities may not be able to adequately influence this measure as well as transplant centers. MAP also noted the need to ensure the measure is appropriately risk-adjusted and recommended the exploration of adjustment for social risk factors and proper risk model performance. MAP ultimately supported the measure with the condition that it is submitted for NQF review and endorsement. Specifically, the MAP recommended that this measure be reviewed by NQF's Scientific Methods Panel as well the Renal Standing Committee. MAP recommended the endorsement process examine the validity of the measure, particularly the risk adjustment model and if it appropriately accounts for social risk. Finally, the MAP noted the need for the Disparities Standing Committee to provide guidance on potential health equity concerns.

Questions for the Committee:

- o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

o *How has the measure been vetted in real-world settings by those being measured or others?*

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

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

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

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

Improvement results: Measure is not currently in use.

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

Unexpected findings (positive or negative) during implementation: Measure is not currently in use.

Potential harms Measure is not currently in use.

Additional Feedback: See MAP feedback above.

Questions for the Committee:

o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
o *Do the benefits of the measure outweigh any potential unintended consequences?*

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

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a. Use and 4b. Usability:

- Recommend committee discussion of usability and use
- This is a better measure than #3402 for facilities.
- Currently measure is not reported. measure can be used to further the goal of high quality of care for patients with ESRD by ensuring that patients are tracked in terms of listing for transplant. Listing dos not ensure transplant-however if the patient is not listed for transplant - the patient will definitely not be transplanted. The measure increases compliance with referral to transplant programs who should be the ultimate deciders in patient suitability for transplantation.
- Significantly question the controllability of this measure at the dialysis facility level. External sources including transplant centers, private insurers, other health factors and economics are at least equally as significant.
- No concerns
- Is being recommended for Dialysis Facility Compare with public reporting to start in Oct 2019. I do not think the harms outweigh the benefits. The only unintended consequence may be that a patient not choose a facility based on this (and that clinic may have a high proportion of patients who are hospice or have been turned down by transplant).
- I have a few additional concerns for this measure. Is this measure designed to be paired with the other incident transplant measure? I ask because part of the rationale for this measure is that 'patients in their first year of dialysis may not be psychologically ready or physically healthy' enough to embark on a transplant evaluation. I agree. This argument weakens the incident measure's validity. One strategy could be including both as a package so as to best represent the importance of kidney transplant waitlist as an indicator of facility quality of care. peritoneal dialysis (PD). PD patients may have distinct geographic barriers to kidney transplant evaluation compared with in-center HD patients, especially in rural states. It was not clear to me that 'patients attributed to a facility' includes only HD or both HD and PD patients. If it is the latter, I would consider looking at these populations separately for differences.

Criterion 5: Related and Competing Measures

Related or competing measures

- No related or competing measures were identified.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: June 6, 2018

- The National Kidney Foundation supports this measure as it is very meaningful for patients. This measure would incentivize greater care coordination by the dialysis facility with the transplant center. Many transplant centers have dialysis outreach programs to better educate facility staff and patients about the transplant process and the patient and dialysis facility role in the process. However, gaps in patients getting waitlisted remain. Patients continue to report that they were not fully informed about transplant or were provided misinformation that led them not to not pursue transplant. Holding dialysis facilities accountable for ensuring their patient population is knowledgeable about transplant and supporting patients to maintain their status on the waitlist will help address this current gap in care. Dialysis facilities can help support patients in maintaining their active status on the waitlist for routine antibody and other periodic testing. However, ultimately, the decision on whether a patient is listed for a transplant is made by the transplant center that evaluated the patient (and the patient's desire for a transplant). These are complex decisions that consider many factors and vary by transplant center and geographic region, which would make nationwide comparisons of waitlist percentages difficult to interpret. The effect of this variance in transplantation policy on dialysis facility performance on this measure should be considered prior to implementation.
- KCP recognizes the tremendous importance of improving transplantation rates for patients with ESRD, but does not support the attribution to dialysis facilities of successful/unsuccessful waitlisting. KCP believes that while a referral to a transplant center, initiation of the waitlist evaluation process, or completion of the waitlist evaluation process may be appropriate facility-level measures that could be used in ESRD quality programs, the Percentage of Prevalent Patients Waitlisted (PPPW) and Standardized First Kidney Transplant Waitlist Ratio for Incident Dialysis Patients (SWR) are not. Waitlisting per se is a decision made by the transplant center and is beyond a dialysis facility's locus of control. In reviewing the PPPW measure, we offer the following comments:
    - FACILITY ATTRIBUTION. KCP appreciated the Measure Applications Partnership (MAP) Hospital Workgroup's recommendation that the Waitlist measures also be reviewed by NQF's Attribution Expert Panel to assess KCP's and other stakeholders' concerns about the measures' attribution models. However, we strongly object to attributing successful/unsuccessful placement on a transplant waitlist to dialysis facilities and believe this is a fatal structural flaw. The transplant center decides whether a patient is placed on a waitlist, not the dialysis facility. One KCP member who is a transplant recipient noted there were many obstacles and delays in the evaluation process with multiple parties that had nothing to do with the dialysis facility—e.g., his private pay insurance changed the locations where he could be evaluated for transplant eligibility on multiple occasions, repeatedly interrupting the process mid-stream. Penalizing a facility each month through the PPPW and SWR for these or other delays is inappropriate; such misattribution is fundamentally misaligned with NQF's first "Attribution Model Guiding Principle", which states that measures' attribution models should fairly and accurately assign accountability.[2] KCP emphasizes our commitment to improving transplantation access, but we believe other measures with an appropriate sphere of control should be pursued.
    - AGE AS THE ONLY SOCIODEMOGRAPHIC RISK VARIABLE. KCP appreciated the MAP Workgroup's recommendation that the Waitlist measures also be reviewed by NQF's Disparities Standing Committee to assess KCP's and other stakeholders' concerns about the measures' risk of potentiating existing health inequities. KCP strongly believes age as the only sociodemographic risk variable is insufficient. We believe other biological and demographic variables are important, and not accounting

for them is a significant threat to the validity of both measures. Transplant centers assess a myriad of demographic factors—e.g., family support, ability to adhere to medication regimens, capacity for follow-up, insurance-related issues, etc. Given transplant centers consider these types of sociodemographic factors, any waitlisting measure risk model should adjust for them. Of note, like the Access to Kidney Transplantation TEP, KCP does not support adjustment for waitlisting based on economic factors or by race or ethnicity.

- Geography, for instance, should be examined, since regional variation in transplantation access is significant. Waitlist times differ regionally, which will ultimately change the percentage of patients on the waitlist and impact performance measure scores. That is, facilities in a region with long wait times will "look" better than those in a region with shorter wait times where patients come off the list more rapidly—even if both are referring at the same rate.
- Additionally, criteria indicating a patient is "not eligible" for transplantation can differ by location—one center might require evidence of an absence of chronic osteomyelitis, infection, heart failure, etc., while another may apply them differently or have additional/ different criteria. The degree to which these biological factors influence waitlist placement must be accounted for in any model for the measure to be a valid representation of waitlisting.
- HOSPICE EXCLUSION. We note that an exclusion for patients admitted to hospice during the month of evaluation has been incorporated into both measures. KCP agrees that the transplantation access measures should not apply to persons with a limited life expectancy and so is pleased to see this revision.
- RISK MODEL FIT. KCP appreciates the MAP Hospital Workgroup's recommendation that the Waitlist measures also be reviewed by NQF's Scientific Methods Panel to assess KCP's and other stakeholders' concerns about the measures' risk models. We note that risk model testing yielded an overall C-statistic of 0.72 for the PPPW and 0.67 for the SWR, raising concerns that the models will not adequately discriminate performance. Smaller units, in particular, might look worse than their actual performance. We reiterate our long-held position that a minimum C-statistic of 0.8 is a more appropriate indicator of a model's goodness of fit, predictive ability, and validity to represent meaningful differences among facilities.
- STRATIFICATION OF RELIABILITY RESULTS BY FACILITY SIZE. CMS has provided no stratification of reliability scores by facility size for either measure; we are thus unable to discern how widely reliability varies across the spectrum of facility sizes. We are concerned that the reliability for small facilities might be substantially lower than the overall IURs, as has been the case, for instance, with other CMS standardized ratio measures. This is of particular concern with the SWR, for which empiric testing has yielded an overall IUR of only 0.6—interpreted as "moderate" reliability by statistical convention.[3] To illustrate our concern, the Standardized Transfusion Ratio for Dialysis Facilities (STrR) measure (NQF 2979) also was found to have an overall IUR of 0.60; however, the IUR was only 0.3 ("poor" reliability) for small facilities (defined by CMS as <=46 patients for the STrR). Without evidence to the contrary, KCP is thus concerned that SWR reliability is similarly lower for small facilities, effectively rendering the metric meaningless for use in performance measurement in this group of providers. KCP believes it is incumbent on CMS to demonstrate reliability for all facilities by providing data by facility size.
- MEANINGFUL DIFFERENCES IN PERFORMANCE. We note that with large sample sizes, as here, even statistically significant differences in performance may not be clinically meaningful. A detailed description of measure scores, such as distribution by quartile, mean, median, standard deviation, outliers, should be provided to allow stakeholders to assess the measure and allow for a thorough review of the measures' performance
- ADDITIONAL LANGUAGE RELATED TO EXCLUSIONS. We note that since KCP reviewed these measures and provided comment to CMS in 2016, one PPPW exclusion has been altered with the following boldface text: Patients admitted to a skilled nursing facility or hospice during the month of evaluation

are excluded from that month; patients admitted to a skilled nursing facility at incidence or previously according to Form CMS 2728 are also excluded. Similarly, one SWR exclusion has been altered with the following boldface/strikeout text: Preemptive patients: Patients at the facility who had the first transplantation prior to the start of ESRD treatment orPatients at the facility whowere listed on the kidney or kidney-pancreas transplant waitlist prior to the start of dialysis.

- KCP supports these changes, but notes that the testing forms submitted by the developer do not provide information on the impact of these exclusions on performance, as required by NQF. We recommend the appropriate, required testing be reported.
- PROCESS VS. INTERMEDIATE OUTCOME MEASURE. The Measure Submission Form identified the PPPW as a process measure. KCP believes the PPPW is an intermediate outcome measure and recommends it be indicated as such.
- In sum and for the reasons stated above, KCP does not believe that the PPPW measure is appropriate for NQF endorsement.
- KCP again thanks you for the opportunity to comment on this important work. If you have any questions, please do not hesitate to contact Lisa McGonigal, MD, MPH (lmcgon@msn.com or 203.530.9524).

- **Of the one NQF members who have submitted a support/non-support choice:**
  - One support the measure

1. Evidence and Performance Gap – Importance to Measure and Report

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

PPPW_NQF_Evidence_form.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 3403

Measure Title: Percentage of Prevalent Patients Waitlisted (PPPW)

Date of Submission: 4/2/2018

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:

☐Patient-reported outcome (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: kidney or kidney-pancreas transplant waitlisting

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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.

The intended objective of this measure is to increase access to kidney transplantation among patients on dialysis. To access transplantation from a deceased donor, the patient must first be accepted on to the kidney transplant wait list. In contradistinction to the Standardized Waitlist Ratio (SWR), which measures placement on the kidney or kidney pancreas transplant waitlist (or receiving a living donor transplant), this measure will assesses ongoing placement on the kidney or kidney-pancreas transplant wait list among prevalent dialysis patients. This is a necessary first step prior to receipt of a deceased donor transplant. The process flow for the steps involved is diagrammed below:

Patients with ESRD on dialysis → Patients not already on the wait list are assessed for eligibility for transplant referral by a nephrologist at the dialysis facility→ Patients are referred to a transplant center for evaluation of candidacy for kidney or kidney-pancreas transplantation → Dialysis facility assists patient with completion of the transplant evaluation process and in optimizing their health and functional status → Patients deemed to be candidates for transplantation are placed on the waitlist. Some with compatible living donors may receive living donor transplants and thus may or may not be placed on the wait list → Dialysis facility helps patient maintain status on the wait list through involvement in ongoing evaluation activities and by optimizing health and functional status → Patients on the wait list have the potential to receive a deceased donor transplant if a compatible one becomes available → Increase in access to transplantation.

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)

☐ 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**	**Tonelli M, Wiebe N, Knoll G, et al. Systematic review: kidney transplantation compared with dialysis in clinically relevant outcomes. American Journal of Transplantation 2011 Oct; 11(10): 2093-2109** **http://onlinelibrary.wiley.com/doi/10.1111/j.1600-6143.2011.03686.x/abstract;jsessionid=61798BDADCD756C587A21D0CE92E60B6.f03t04**

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.	Individual studies indicate that kidney transplantation is associated with lower mortality and improved quality of life compared with chronic dialysis treatment. We did a systematic review to summarize the benefits of transplantation, aiming to identify characteristics associated with especially large or small relative benefit. Results were not pooled because of expected diversity inherent to observational studies. Risk of bias was assessed using the Downs and Black checklist and items related to time-to-event analysis techniques. MEDLINE and EMBASE were searched up to February 2010. Cohort studies comparing adult chronic dialysis patients with kidney transplantation recipients for clinical outcomes were selected. We identified 110 eligible studies with a total of 1 922 300 participants. Most studies found significantly lower mortality associated with transplantation, and the relative magnitude of the benefit seemed to increase over time (p < 0.001). Most studies also found that the risk of cardiovascular events was significantly reduced among transplant recipients. Quality of life was significantly and substantially better among transplant recipients. Despite increases in the age and comorbidity of contemporary transplant recipients, the relative benefits of transplantation seem to be increasing over time. These findings validate current attempts to increase the number of people worldwide that benefit from kidney transplantation.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	No formal grading was used by the authors of the systematic review. However, evaluation of the quality of the studies was performed (described in more detail below). The authors concluded based on the consistent beneficial effect noted on mortality for transplantation versus a range of dialysis modalities that kidney transplantation is the preferred modality of treatment for patients requiring renal replacement therapy.
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	N/A
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?	A total of 110 studies were included in the review, representing over 1.9 million patients. All studies were either retrospective and/or prospective cohort observational study designs. No randomized clinical trials were available for inclusion.
Estimates of benefit and consistency across studies	Due to heterogeneity, results were not formally pooled. However, the majority of studies (76%) demonstrated a survival advantage for kidney transplantation. Among those studies with the best design for reducing selection bias, including multivariable adjustment and a comparison group consisting of waitlisted dialysis patients, 94% of tested comparisons demonstrated a lower mortality with transplantation (with hazard ratios ranging from 0.16-0.73). Similarly, the vast majority of studies demonstrated better quality of life scores on the SF-36 for kidney transplant patients versus those on dialysis.

What harms were identified?	No harms were examined.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	More recent studies published after this review also confirm the survival benefits of kidney transplantation over dialysis and none substantively affect the conclusions of the systematic review [1,2,3,4,5,6,7,8 below].

1. Reese PP, Shults J, Bloom RD, et al. Functional Status, Time to Transplantation, and Survival Benefit of Kidney Transplantation Among Wait-Listed Candidates. Am J Kidney Dis. 2015 Jul 7. pii: S0272-6386(15)00844-6

Abstract:

BACKGROUND: In the context of an aging end-stage renal disease population with multiple comorbid conditions, transplantation professionals face challenges in evaluating the global health of patients awaiting kidney transplantation. Functional status might be useful for identifying which patients will derive a survival benefit from transplantation versus dialysis.

STUDY DESIGN: Retrospective cohort study of wait-listed patients using data for functional status from a national dialysis provider linked to United Network for Organ Sharing registry data.

SETTING & PARTICIPANTS: Adult kidney transplantation candidates added to the waiting list between 2000 and 2006.

PREDICTOR: Physical Functioning scale of the Medical Outcomes Study 36-Item Short Form Health Survey, analyzed as a time-varying covariate.

OUTCOMES: Kidney transplantation; survival benefit of transplantation versus remaining wait-listed.

MEASUREMENTS: We used multivariable Cox regression to assess the association between physical function with study outcomes. In survival benefit analyses, transplantation status was modeled as a time-varying covariate.

RESULTS: The cohort comprised 19,242 kidney transplantation candidates (median age, 51 years; 36% black race) receiving maintenance dialysis. Candidates in the lowest baseline Physical Functioning score quartile were more likely to be inactivated (adjusted HR vs highest quartile, 1.30; 95% CI, 1.21-1.39) and less likely to undergo transplantation (adjusted HR vs highest quartile, 0.64; 95% CI, 0.61-0.68). After transplantation, worse Physical Functioning score was associated with shorter 3-year survival (84% vs 92% for the lowest vs highest function quartiles). However, compared to dialysis, transplantation was associated with a statistically significant survival benefit by 9 months for patients in every function quartile.

LIMITATIONS: Functional status is self-reported.

CONCLUSIONS: Even patients with low function appear to live longer with kidney transplantation versus dialysis. For wait-listed patients, global health measures such as functional status may be more useful in counseling patients about the probability of transplantation than in identifying who will derive a survival benefit from it.

2. Lloveras J, Arcos E, Comas J, Crespo M, Pascual J. A paired survival analysis comparing hemodialysis and kidney transplantation from deceased elderly donors older than 65 years. Transplantation. 2015 May; 99(5):991-6.

Abstract:

BACKGROUND: Kidney transplantation from deceased donors aged 65 years or older is associated with suboptimal patient and graft survival. In large registries, survival is longer after kidney transplantation than when remaining on dialysis. However, whether recipients of these old grafts survive longer than their dialysis counterparts is unknown.

METHODS: We retrospectively assessed the outcomes of 5,230 recipients of first deceased donor grafts transplanted during the period of 1990 to 2010 in Catalonia, 915 of whom received grafts from donors 65 years or older. In a match-pair analysis, we aimed to pair each of 915 eligible cases with one control (1:1 ratio). Each pair had the same characteristics at the time of entering dialysis program: age, sex, primary renal disease, period of dialysis onset, and cardiovascular comorbidities. We found 823 pairs.

RESULTS: Patient survival of 823 recipients of elderly donors was significantly higher than that of their 823 matched dialysis waitlisted nontransplanted partners (91.6%, 74.5%, and 55.5% vs. 88.8%, 44.2%, and 18.1%, respectively at 1, 5, and 10 years; P<0.001). The probability of death after the first year was similar (8.1% transplant vs 10.3% dialysis; P=0.137); however, analyzing the whole period, the adjusted proportional risk of death was 2.66 (95% confidence interval, 2.21-3.20) times higher for patients remaining on dialysis than for transplanted patients (P<0.001).

CONCLUSION: Our study demonstrates that despite the fact that kidney transplantation from elderly deceased donors is associated with reduced graft and patient survival, their paired counterpart patients remaining on dialysis have a risk of death 2.66 times higher.

3. Schold JD, Buccini LD, Goldfarb DA, et al. Association between kidney transplant center performance and the survival benefit of transplantation versus dialysis. Clin J Am Soc Nephrol. 2014 Oct 7; 9(10):1773-80.

Abstract:

BACKGROUND AND OBJECTIVES: Despite the benefits of kidney transplantation, the total number of transplants performed in the United States has stagnated since 2006. Transplant center quality metrics have been associated with a decline in transplant volume among low-performing centers. There are concerns that regulatory oversight may lead to risk aversion and lack of transplantation growth.

DESIGN, SETTING, PARTICIPANTS, & MEASUREMENTS: A retrospective cohort study of adults (age≥18 years) wait-listed for kidney transplantation in the United States from 2003 to 2010 using the Scientific Registry of Transplant Recipients was conducted. The primary aim was to investigate whether measured center performance modifies the survival benefit of transplantation versus dialysis. Center performance was on the basis of the most recent Scientific Registry of Transplant Recipients evaluation at the time that patients were placed on the waiting list. The primary outcome was the time-dependent adjusted hazard ratio of death compared with remaining on the transplant waiting list.

RESULTS: Among 223,808 waitlisted patients, 59,199 and 32,764 patients received a deceased or living donor transplant, respectively. Median follow-up from listing was 43 months (25th percentile=25 months, 75th percentile=67 months), and there were 43,951 total patient deaths. Deceased donor transplantation was independently associated with lower mortality at each center performance level compared with remaining on the waiting list; adjusted hazard ratio was 0.24 (95% confidence interval, 0.21 to 0.27) among 11,972 patients listed at high-performing centers, adjusted hazard ratio was 0.32 (95% confidence interval, 0.31 to 0.33) among 203,797 patients listed at centers performing as expected, and adjusted hazard ratio was 0.40 (95% confidence interval, 0.35 to 0.45) among 8039 patients listed at low-performing centers. The survival benefit was significantly different by center performance (P value for interaction <0.001).

CONCLUSIONS: Findings indicate that measured center performance modifies the survival benefit of kidney transplantation, but the benefit of transplantation remains highly significant even at centers with low measured quality. Policies that concurrently emphasize improved center performance with access to transplantation should be prioritized to improve ESRD population outcomes.

4. Tennankore KK, Kim SJ, Baer HJ, Chan CT. Survival and hospitalization for intensive home hemodialysis compared with kidney transplantation. J Am Soc Nephrol. 2014 Sep; 25(9):2113-20.

Abstract:

Canadian patients receiving intensive home hemodialysis (IHHD; ≥16 hours per week) have survival comparable to that of deceased donor kidney transplant recipients in the United States, but a comparison with Canadian kidney transplant recipients has not been conducted. We conducted a retrospective cohort study of consecutive, adult IHHD patients and kidney transplant recipients between 2000 and 2011 at a large Canadian tertiary care center. The primary outcome was time-to-treatment failure or death for IHHD patients compared with expanded criteria, standard criteria, and living donor recipients, and secondary outcomes included hospitalization rate. Treatment failure was defined as a permanent switch to an alternative dialysis modality for IHHD patients, and graft failure for transplant recipients. The cohort comprised 173 IHHD patients and 202 expanded criteria, 642 standard criteria, and 673 living donor recipients. There were 285 events in the primary analysis. Transplant recipients had a reduced risk of treatment failure/death compared with IHHD patients, with relative hazards of 0.45 (95% confidence interval [95% CI], 0.31 to 0.67) for living donor recipients, 0.39 (95% CI, 0.26 to 0.59) for standard criteria donor recipients, and 0.42 (95% CI, 0.26 to 0.67) for expanded criteria donor recipients. IHHD patients had a lower hospitalization rate in the first year of treatment compared with standard criteria donor recipients and in the first 3 months of treatment compared with living donor and expanded criteria donor recipients. In this cohort, kidney transplantation was associated with superior treatment and patient survival, but higher early rates of hospitalization, compared with IHHD.

5. Gill JS, Lan J, Dong J, et al. The survival benefit of kidney transplantation in obese patients. Am J Transplant. 2013 Aug; 13(8):2083-90.

Abstract:

Obese patients have a decreased risk of death on dialysis but an increased risk of death after transplantation, and may derive a lower survival benefit from transplantation. Using data from the United States between 1995 and 2007 and multivariate non-proportional hazards analyses we determined the relative risk of death in transplant recipients grouped by body mass index (BMI) compared to wait-listed candidates with the same BMI (n = 208 498). One year after transplantation the survival benefit of transplantation varied by BMI: Standard criteria donor transplantation was associated with a 48% reduction in the risk of death in patients with BMI ≥ 40 kg/m(2) but a ≥ 66% reduction in patients with BMI < 40 kg/m2. Living donor transplantation was associated with ≥ 66% reduction in the risk of death in all BMI groups. In sub-group analyses, transplantation from any donor source was associated with a survival benefit in obese patients ≥ 50 years, and diabetic patients, but a survival benefit was not demonstrated in Black patients with BMI ≥ 40 kg/m(2). Although most obese patients selected for transplantation derive a survival benefit, the benefit is lower when BMI is ≥ 40 kg/m(2), and uncertain in Black patients with BMI ≥ 40 kg/m(2).

6. Ingsathit A, Kamanamool N, Thakkinstian A, Sumethkul V. Survival advantage of kidney transplantation over dialysis in patients with hepatitis C: a systematic review and meta-analysis.Transplantation. 2013 Apr 15; 95(7):943-8.

Abstract:

BACKGROUND: The clinical outcomes of hepatitis C infection in kidney transplantation and maintenance dialysis patients remain controversial. Here, we conducted a systematic review and meta-analysis that aimed at comparing 5-year mortality rates between waiting list and kidney transplantation patients with hepatitis C infections.

METHODS: We searched Medline, EMBASE, and Scopus databases published since inception to June 2011 and found nine studies with 1734 patients who were eligible for pooling. Eligible studies were cohort studies that analyzed adult end-stage renal disease patients with hepatitis C virus infection and compared death rates between waiting list and kidney transplantation. The crude risk ratio of death along with its 95% confidence interval was estimated for each study. Data were independently extracted by two reviewers.

RESULTS: The pooled risk ratio of death at 5 years by using a random-effect model was 2.19 (95% confidence interval, 1.50-3.20), which significantly favored the kidney transplantation when compared with the waiting list. There was evidence of heterogeneity of death rates across studies ($\chi(2) = 22.6$; df = 8; P = 0.004). From the metaregression model, age and male gender could be the source of heterogeneity or variation of treatment effects. A major cause of death in the waiting list was cardiovascular diseases, whereas infection was a major cause in the transplant group. There was no evidence of publication bias suggested by an Egger test.

CONCLUSIONS: This systematic review suggested that hepatitis C virus-infected patients who remain on dialysis are at higher risk of death when compared with those who received kidney transplantations.

7. De Lima JJ, Gowdak LH, de Paula FJ, et al. Which patients are more likely to benefit from renal transplantation? Clin Transplant. 2012 Nov-Dec; 26(6):820-5.

Abstract:

BACKGROUND: We evaluated whether the advantages conferred by renal transplantation encompass all individuals or whether they favor more specific groups of patients.

METHODS: One thousand and fifty-eight patients on the transplant waiting list and 270 receiving renal transplant were studied. End points were the composite incidence of CV events and death. Patients were followed up from date of placement on the list until transplantation, CV event, or death (dialysis patients), or from the date of transplantation, CV event, return to dialysis, or death (transplant patients).

RESULTS: Younger patients with no comorbidities had a lower incidence of CV events and death independently of the treatment modality (log-rank=0.0001). Renal transplantation was associated with better prognosis only in high-risk patients (p=0.003).

CONCLUSIONS: Age and comorbidities influenced the prevalence of CV complications and death independently of the treatment modality. A positive effect of renal transplantation was documented only in high-risk patients. These findings suggest that age and comorbidities should be considered indication for early transplantation even considering that, as a group, such patients have a shorter survival compared with low-risk individuals.

8. Wong G, Howard K, Chapman JR, et al. Comparative survival and economic benefits of deceased donor kidney transplantation and dialysis in people with varying ages and co-morbidities. PLoS One. 2012; 7(1):e29591.

Abstract:

BACKGROUND: Deceased donor kidneys for transplantation are in most countries allocated preferentially to recipients who have limited co-morbidities. Little is known about the incremental health and economic gain from transplanting those with co-morbidities compared to remaining on dialysis. The aim of our study is to estimate the average and incremental survival benefits and health care costs of listing and transplantation compared to dialysis among individuals with varying co-morbidities.

METHODS: A probabilistic Markov model was constructed, using current outcomes for patients with defined co-morbidities treated with either dialysis or transplantation, to compare the health and economic benefits of listing and transplantation with dialysis.

FINDINGS: Using the current waiting time for deceased donor transplantation, transplanting a potential recipient, with or without co-morbidities achieves survival gains of between 6 months and more than three life years compared to remaining on dialysis, with an average incremental cost-effectiveness ratio (ICER) of less than $50,000/LYS, even among those with advanced age. Age at listing and the waiting time for transplantation are the most influential variables within the model. If there were an unlimited supply of organs and no waiting time, transplanting the younger and healthier individuals saves the most number of life years and is cost-saving, whereas transplanting the middle-age to older patients still achieves substantial incremental gains in life expectancy compared to being on dialysis.

CONCLUSIONS: Our modelled analyses suggest transplanting the younger and healthier individuals with end-stage kidney disease maximises survival gains and saves money. Listing and transplanting those with considerable co-morbidities is also cost-effective and achieves substantial survival gains compared with the dialysis alternative. Preferentially excluding the older and sicker individuals cannot be justified on utilitarian grounds.

____

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

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

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 answer the composite questions.

A measure focusing on the wait listing process is appropriate for improving access to kidney transplantation for several reasons. First, wait listing is a necessary step prior to potential receipt of a deceased donor kidney. Second, dialysis facilities exert substantial control over the process of waitlisting. This includes proper education of dialysis patients on the option for transplant, referral of appropriate patients to a transplant center for evaluation, assisting patients with completion of the transplant evaluation process, and optimizing the health and functional status of patients in order to increase their candidacy for transplant wait listing. These types of activities are included as part of the conditions for coverage for Medicare certification of ESRD dialysis facilities. In addition, dialysis facilities can also help maintain patients on the wait list through assistance with ongoing evaluation activities and by optimizing health and functional status. Finally, wide regional variations in wait listing rates highlight substantial room for improvement for this process measure [1,2,3].

This measure focuses specifically on the prevalent dialysis population, examining waitlisting status monthly for each patient. This allows evaluation and encouragement of ongoing waitlisting of patients beyond the first year of dialysis initiation who have not yet been listed. Patients may not be ready, either psychologically or due to their health status, to consider transplantation early after initiation of dialysis and many choose to undergo evaluation for transplantation only after years on dialysis. In addition, as this measure assesses monthly waitlisting status of patients, it also evaluates and encourages maintenance of patients on the waitlist. Maintenance of active status on the waitlist is important for increasing likelihood of transplantation [4] and thus by extension, is waitlisting overall. This is an important area to which dialysis facilities can contribute through ensuring patients remain healthy, and complete any ongoing testing activities required to remain on the wait list. In contrast to this measure, another waitlisting measure, the Standardized First Kidney Transplant Waitlist Ratio for Incident Dialysis Patients (SWR), focuses solely on new listing or living kidney donor transplantation within the first year after initiation of dialysis with the rationale of encouraging early access to transplantation or the wait list.

1. Ashby VB, Kalbfleisch JD, Wolfe RA, et al. Geographic variability in access to primary kidney transplantation in the United States, 1996-2005. American Journal of Transplantation 2007; 7 (5 Part 2):1412-1423.

Abstract:

This article focuses on geographic variability in patient access to kidney transplantation in the United States. It examines geographic differences and trends in access rates to kidney transplantation, in the component rates of wait-listing, and of living and deceased donor transplantation. Using data from Centers for Medicare and Medicaid Services and the Organ Procurement and Transplantation Network/Scientific Registry of Transplant Recipients, we studied 700,000+ patients under 75, who began chronic dialysis treatment, received their first living donor kidney transplant, or were placed on the waiting list pre-emptively. Relative rates of wait-listing and transplantation by State were calculated using Cox regression models, adjusted for patient demographics. There were geographic differences in access to the kidney waiting list and to a kidney transplant. Adjusted wait-list rates ranged from 37% lower to 64% higher than the national average. The living donor rate ranged from 57% lower to 166% higher, while the deceased donor transplant rate ranged from 60% lower to 150% higher than the national average. In general, States with higher wait-listing rates tended to have lower transplantation rates and States with lower wait-listing rates had higher transplant rates. Six States demonstrated both high wait-listing and deceased donor transplantation rates while six others, plus D.C. and Puerto Rico, were below the national average for both parameters.

2. Satayathum S, Pisoni RL, McCullough KP, et al. Kidney transplantation and wait-listing rates from the international Dialysis Outcomes and Practice Patterns Study (DOPPS). Kidney Intl 2005 Jul; 68 (1):330-337.

Abstract:

BACKGROUND: The international Dialysis Outcomes and Practice Patterns Study (DOPPS I and II) allows description of variations in kidney transplantation and wait-listing from nationally representative samples of 18- to 65-year-old hemodialysis patients. The present study examines the health status and socioeconomic characteristics of United States patients, the role of for-profit versus not-for-profit status of dialysis facilities, and the likelihood of transplant wait-listing and transplantation rates.

METHODS: Analyses of transplantation rates were based on 5267 randomly selected DOPPS I patients in dialysis units in the United States, Europe, and Japan who received chronic hemodialysis therapy for at least 90 days in 2000. Left-truncated Cox regression was used to assess time to kidney transplantation. Logistic regression determined the odds of being transplant wait-listed for a cross-section of 1323 hemodialysis patients in the United States in 2000. Furthermore, kidney transplant wait-listing was determined in 12 countries from cross-sectional samples of DOPPS II hemodialysis patients in 2002 to 2003 (N= 4274).

RESULTS: Transplantation rates varied widely, from very low in Japan to 25-fold higher in the United States and 75-fold higher in Spain (both P values <0.0001). Factors associated with higher rates of transplantation included younger age, nonblack race, less comorbidity, fewer years on dialysis, higher income, and higher education levels. The likelihood of being wait-listed showed wide variation internationally and by United States region but not by for-profit dialysis unit status within the United States.

CONCLUSION: DOPPS I and II confirmed large variations in kidney transplantation rates by country, even after adjusting for differences in case mix. Facility size and, in the United States, profit status, were not associated with varying transplantation rates. International results consistently showed higher transplantation rates for younger, healthier, better-educated, and higher income patients.

3. Patzer RE, Plantinga L, Krisher J, Pastan SO. Dialysis facility and network factors associated with low kidney transplantation rates among United States dialysis facilities. Am J Transplant. 2014 Jul; 14(7):1562-72.

Abstract:

Variability in transplant rates between different dialysis units has been noted, yet little is known about facility-level factors associated with low standardized transplant ratios (STRs) across the United States End-stage Renal Disease (ESRD) Network regions. We analyzed Centers for Medicare & Medicaid Services Dialysis Facility Report data from 2007 to 2010 to examine facility-level factors associated with low STRs using multivariable mixed models. Among 4098 dialysis facilities treating 305698 patients, there was wide variability in facility-level STRs across the 18 ESRD Networks. Four-year average STRs ranged from 0.69 (95% confidence interval [CI]: 0.64-0.73) in Network 6 (Southeastern Kidney Council) to 1.61 (95% CI: 1.47-1.76) in Network 1 (New England). Factors significantly associated with a lower Standardized Transplantation Ratio(STR) (p<0.0001) included for-profit status, facilities with higher percentage black patients, patients with no health insurance and patients with diabetes. A greater number of facility staff, more transplant centers per 10 000 ESRD patients and a higher percentage of patients who were employed or utilized peritoneal dialysis were associated with higher STRs. The lowest performing dialysis facilities were in the Southeastern United States. Understanding the modifiable facility-level factors associated with low transplant rates may inform interventions to improve access to transplantation.

4. Grams, M. E., Massie, A. B., Schold, J. D., Chen, B. P., & Segev, D. L. (2013). Trends in the inactive kidney transplant waitlist and implications for candidate survival. American Journal of Transplantation, 13(4), 1012-1018.

Abstract

In November 2003, OPTN policy was amended to allow kidney transplant candidates to accrue waiting time while registered as status 7, or inactive. We evaluated trends in inactive listings and the association of inactive status with transplantation and survival, studying 262,824 adult first-time KT candidates listed between 2000 and 2011. The proportion of waitlist candidates initially listed as inactive increased from 2.3% prepolicy change to 31.4% in 2011. Candidates initially listed as inactive were older, more often female, African American, and with higher body mass index.

Postpolicy change, conversion from initially inactive to active status generally occurred early if at all: at 1 year after listing, 52.7% of initially inactive candidates had been activated; at 3 years, only 66.3% had been activated. Inactive status was associated with a substantially higher waitlist mortality (aHR 2.21, 95%CI:2.15-2.28, p<0.001) and lower rates of eventual transplantation (aRR 0.68, 95%CI:0.67-0.70, p<0.001). In summary, waitlist practice has changed significantly since November 2003, with a sharp increase in the number of inactive candidates. Using the full waitlist to estimate organ shortage or as a comparison group in transplant outcome studies is less appropriate in the current era.

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.*

After applying all exclusion criteria, we evaluated the PPPW performance scores for all dialysis facilities that had at least 11 patients in 2016. The Percentage of Prevalent Patients Waitlisted (PPPW) varies considerably across facilities. The mean value of PPPW was 0.21 (i.e. 21% of prevalent patients were waitlisted). The interquartile range (Q3-Q1) is around 0.15, with the bottom quartile of facilities having 12% or less of prevalent patients waitlisted versus the top quartile of facilities having 27% or more of their prevalent patients waitlisted.

Mean standard deviation and quartiles of PPPW:

N= 6617

Mean = 0.21

Standard Deviation = 0.11

0% Min = 0.00

25% Q1 = 0.12

50% Median = 0.19

75% Q3 = 0.27

100% Max = 0.78

Descriptive statistics by decile are reported in the Appendix.

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.

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.*

Below we show the parameter estimates for the race, sex and ethnicity variables based on a model that included these variables along with original covariates. There is evidence of significant differences in measure results by sex, race, and ethnicity. However, there is no clear biological rationale for differences in waitlisting on the basis of sex, race or ethnicity to justify a need for adjustment.

Sex: Male (Reference), Female (estimate = -0.08, p-value = <.001)

Race: White (Reference), Black (estimate = -0.08, p-value = <.001), Asian (estimate = 0.38, p-value = <.001), Native American (estimate = -0.31, p-value = <.001), Other (estimate = -0.01, p-value = 0.93)

Ethnicity: Hispanic (Reference), Non-Hispanic/Unknown (estimate = -0.04, p-value = 0.01)

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

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

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 not an eMeasure **Attachment:**

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)*

Attachment **Attachment:** PPPW_DataDictionary.xlsx

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.

No, this is not an instrument-based measure **Attachment:**

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

Not an instrument-based measure

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.

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.

N/A

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 in which the patient at the dialysis facility is on the kidney or kidney-pancreas transplant waitlist as of the last day of each month during the reporting year.

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).

To be included in the numerator for a particular month, the patient must be on the kidney or kidney-pancreas transplant waitlist as of the last day of the month during the reporting year.

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

All patient-months for patients who are under the age of 75 in the reporting month and who are assigned to the dialysis facility according to each patient's treatment history as of the last day of each month during the reporting year.

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 should be described in the calculation algorithm (S.14).

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. Information regarding first ESRD service date, death, waitlist status 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) and the Social Security Death Master File. For denominator exclusions, the Nursing Home Minimum Dataset and the Questions 17u and 22 on CMS Medical Evidence Form are used to identify patients in skilled nursing facilities. Additionally, a separate CMS file that contains final action claims submitted by Hospice providers was used to determine the hospice status.

The model is currently age-adjusted, with age updated each month.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

Exclusions that are implicit in the denominator include:

- Patients who were at age 75 or older in the reporting month.
- Patient who were admitted to a skilled nursing facility (SNF) or a hospice during the month of evaluation were excluded from that month; patients who were admitted to a skilled nursing facility (SNF) at incidence or previously according to Form CMS-2728 were also excluded.

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.)*

The Nursing Home Minimum Dataset and the Questions 17u and 22 on CMS Medical Evidence Form are used to identify patients in skilled nursing facilities. For hospice patients, a separate CMS file that contains final action claims submitted by Hospice providers was used to determine the hospice status.

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)

Statistical risk model

If other:

S.12. Type of score:

Rate/proportion

If other:

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 for data, aggregating data; risk adjustment; etc.)*

See flowchart in Appendix.

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.)*

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, Registry Data

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 is the primary data source we used for denominator, risk adjustment (age) and exclusion of patients at 75 year-old or older (see information provided under "denominator details"). Organ Procurement and Transplant Network (OPTN) is the data source for numerator (waitlisting). The Nursing Home Minimum Dataset and Questions 17u and 22 on the CMS Medical Evidence Form are used to identify SNF patients. A separate CMS file that contains final action claims submitted by Hospice providers was used to determine the hospice status.

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

No data collection instrument provided

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

Facility

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

Other

If other: Dialysis Facility

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

N/A

2. Validity – See attached Measure Testing Submission Form

PPPW_NQF_TestingForm_20180402.docx

2.1 <u>For maintenance of endorsement</u>

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.

2.2 <u>For maintenance of endorsement</u>

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.

2.3 <u>For maintenance of endorsement</u>

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, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 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.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 3403
Measure Title: Percentage of Prevalent Patients Waitlisted (PPPW)
Date of Submission: <u>4/2/2018</u>

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 <u>all</u> measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For <u>outcome and resource use</u> 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:	☐ other:

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*).

2016 data derived from a combination of CROWNWeb, the Nursing Home Minimum Data Set, transplant registries (OPTN, SRTR), the CMS Medical Evidence Form (CMS Form-2728) and hospice claims from CMS.

1.3. What are the dates of the data used in testing? January-December 2016

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:	☐ other:

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*)

Using 2016 data, there were 6,617 facilities included in these analyses, after restricting to facilities that had ≥11 eligible patients.

1.6. How many and which patients 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 are 4,283,227 patient-months (449,110 patients) in total. Among all patient-months in 2016, the average age was 56.5 years old, 41.8% of patient-months were female, 55.2% were White, 37.3% were Black, 5.7% were Asian/Pacific Islander, 1.3% American Indian/Alaskan Native, 0.4% were other/Multi-racial/unknown/missing and 20.0 % were Hispanic. At patient level, the mean age was 56.5 years old and 41.9% were female. Of these 56.7% were White, 36.0% were Black, 5.6% were Asian/Pacific Islander, 1.3% were American Indian/Alaskan Native, and the rest 0.4% were other/Multiracial/unknown/missing. 19.2% patients were of Hispanic ethnicity, while 80.4% were non-Hispanic and 0.4% were unknown or missing.

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.

Patient level

- Sex
- Employment status 6 months prior to ESRD
- Race

- Ethnicity
- Medicare coverage*

**Assessed at a specific time point (e.g., at the reporting month). Medicare coverage in model was defined as:*

1. Medicare as primary and Medicaid

2. Medicare as primary and NO Medicaid

3. Medicare as secondary or Medicare HMO (e.g. Medicare Advantage)

4. Non-Medicare/missing

Data on patient level SDS/SES factors obtained from Medicare claims and administrative data.

ZIP code level – Area Deprivation Index (ADI) from 2014 Census data.

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 2016 – December 2016 data to calculate facility-level annual performance scores. The NQF-recommended approach for determining measure reliability is a one-way analysis of variance (ANOVA), in which the between-facility variation (σ_b^2) and the within-facility variation ($\sigma_{t,w}^2$) in the measure is determined. The inter-unit reliability (IUR) measures the proportion of the total variation of a measure (i.e., $\sigma_b^2 + \sigma_{t,w}^2$) that is attributable to the between-facility variation, the true signal reflecting the differences across facilities. We assessed reliability by calculating inter-unit reliability (IUR) for the annual performance scores. A small IUR (near 0) reveals that most of the variation of the measure between facilities is driven by random noise, indicating the measure would not be a good characterization of the differences among facilities, whereas a large IUR (near 1) indicates that most of the variation between facilities is due to the real difference between facilities.

Here we describe our approach to calculating IUR. Let $T_1,\ldots,T_N$ be the Percentage of Prevalent Patients Waitlisted (PPPW) for N facilities. Since the variation in $T_1,\ldots,T_N$ is mainly driven by the estimates of facility-specific intercepts ($\alpha_1,\ldots,\alpha_N$), we use their asymptotic distributions to estimate the within-facility variation in PPPW. Applying the delta method, we estimate the variance of T_i and denote the estimate as S_i^2. Calling on formulas from the one-way ANOVA, the within-facility variance in PPPW can be estimated by

$$s_{t,w}^2 = \frac{\sum_{i=1}^{N}\left[(n_i - 1)S_i^2\right]}{\sum_{i=1}^{N}(n_i - 1)},$$

and the total variation in PPPW can be estimated by

$$s_t^2 = \frac{1}{n'(N-1)} \sum_{i=1}^{N} n_i (T_i - \bar{T})^2 ,$$

where n_i is the number of subjects in the ith facility, $\bar{T} = \Sigma\, n_i T_i / \Sigma\, n_i$, and

$$n' = \frac{1}{N-1} \left(\sum n_i - \sum n_i^2 \Big/ \sum n_i\right)$$

is approximately the average facility size (number of patients per facility). Thus, the IUR = $\sigma_b^2 / (\sigma_b^2 + \sigma_{t,w}^2)$ can be estimated by $(s_t^2 - s_{t,w}^2)/s_t^2$.

The reliability of PPPW calculation only included facilities with at least 11 patients during the entire year.

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 IUR value is 0.80. Facilities with <11 eligible patients were excluded from this calculation.

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?*)

This value of IUR indicates that about four-fifths of the variation in the PPPW can be attributed to the between-facility differences (signal) and about one-fifth to within-facility variation (noise). This value of IUR implies a high degree of reliability.

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 <u>performance measure score</u> 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)*

Systematic Assessment of Face Validity: The primary purpose of this measure is to increase access to kidney transplantation for patients on chronic dialysis. Because waitlisting is a crucial, necessary step prior to potential receipt of a deceased donor kidney, a measure which assesses waitlisting of patients by dialysis facilities has face validity as a measure of access to transplantation. Furthermore, a Technical Expert Panel (TEP), of 11 members consisting of transplant nephrologists, social workers, administrators and nurses with transplant process, policy and research expertise was convened. The TEP was charged with development of potential dialysis facility measures directed at improving access to transplantation. Although not unanimous, there was majority (by formal vote of 8-3) support for a dialysis facility measure related to waitlisting, on the basis that dialysis facilities importantly contribute to waitlisting of patients by helping them to navigate the process from referral through completion of the transplant evaluation, ensuring that all necessary testing as part of the evaluation process is done in a timely manner, and contributing to their overall health and therefore suitability for transplantation.

Empirical validity testing - validation of performance measure scores: We assessed empirical validity of the measure by calculating Spearman correlations. Spearman correlation was selected because the data are rank-ordered (non-parametric data). Correlations were calculated to assess the association of the PPPW with other outcome quality measures. First, to demonstrate the relationship between PPPW and the anticipated outcome of increasing transplantation rates for patients at the facility, we examined the correlation of facility ranking with respect to the measure and the Standardized Transplant Ratio (STR, 2013-2016). The STR is the ratio of the actual number of first transplants to the expected number of first transplants for the facility, given the age composition of the facility's patients in 2013-2016. There are 4,857 facilities available for comparison. We expected to find that the PPPW and STR would be positively correlated.

We further examined the relationship between PPPW and a number of measures reflecting the quality of overall health care delivered to dialysis patients by facilities. These include the 2013-2016 Standardized Mortality Ratio (SMR), 2016 Standardized Hospitalization Ratio (SHR), 2016 Standardized Hospitalization Ratio (ED visits), and 2016 Standardized Readmission Ratio (SRR). We anticipated that facilities with higher PPPW would also have lower rates of adverse health outcomes, reflecting that maintenance of good health status by dialysis facilities increases the likelihood of waitlisting, and remaining on the waitlist. Therefore we expected to find that PPPW and these measures would be negatively correlated.

To summarize, we expected the following correlations of PPPW to the above quality measures:

STR: We anticipated a positive correlation between PPPW and the STR.

SMR: We anticipated a negative correlation with PPPW.

SHR: We anticipated a negative correlation with PPPW.

SHR (ED): We anticipated a negative correlation with PPPW.

SRR: We anticipated a negative correlation with PPPW.

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

The Spearman correlation coefficient between facility waitlist rate and STR was significant: rho=0.45, p<.0001. There is also significant correlation between PPPW and the SMR (n=6,086, r=-0.11, p<.001), SHR (admissions) (n=6,400, r=-0.03, p<.001), SHR (ED visits) (n=6,400, r=-0.22, p<.001), and SRR (n=6,375, r=-0.03, p<.001).

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?*)

All results were as expected. Percentage of Prevalent Patients Waitlisted (PPPW) is positively correlated with STR, suggesting that facilities with higher waitlisting rates also have higher transplant rates. The Spearman correlation between PPPW and other measures indicates that higher waitlisted rate is associated with lower mortality rate, lower hospitalization rate and lower readmission rate.

____

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*)

In order to see the differences with and without excluding nursing home patients, the number of patient-months before and after exclusion were compared (Table 3). In Figure 1, we show a histogram of patient-months excluded by facility. Additionally, in Table 4 we compare the quantiles of crude percentage waitlisted (before versus after exclusion).

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*)

Table 3. Patient-months before and after excluding nursing home and hospice patients, 2016

	Before exclusion	**After exclusion**	**Percentage excluded**
Numbers of Patient-months	4,594,717	4,283,227	6.8%

Figure 1. Histogram of patient-months excluded, at facility level, 2016

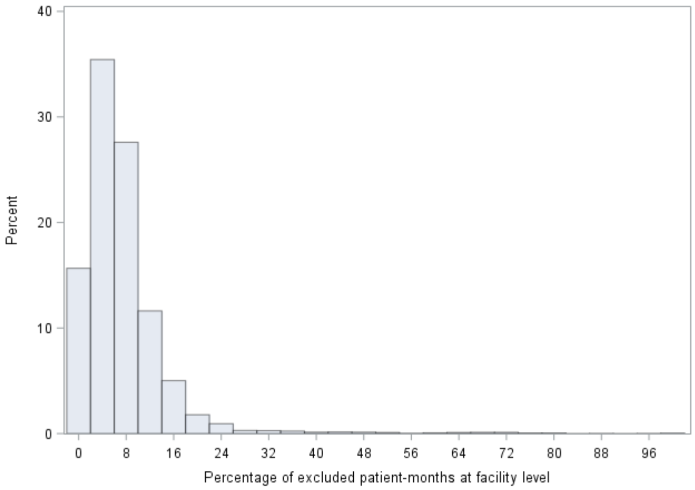


Table 4. Quantiles of crude waitlist rates before and after exclusion, 2016

	Mean (Std)	**Q1 (25%)**	**Q2 (50%)**	**Q3 (75%)**	**Q4 (100%)**
Before exclusion	0.19 (0.12)	0.11	0.18	0.26	1.00
After exclusion	0.20 (0.12)	0.12	0.19	0.27	1.00

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*)

Figure 1 reveals variation in the percent of excluded patients across facilities and Table 4 shows some change in the distribution of scores, supporting the need for exclusion to prevent distortion in performance results across facilities.

____

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** age (knots at 15, 55 and 70) as the **risk factors**

☐ **Stratification by _risk categories**

☐ **Other,**

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.

We assume a logistic regression model for the probability that a prevalent patient is wait-listed. Consider patient *i* at facility *j* during calendar month *k*; we set the response variate to Y_{ijk} =1 if the patient is on the wait list and Y_{ijk} 0 if not. The model is adjusted for age,

$$\text{logit}(p_{ijk}) = \alpha_j + \beta A_{ij},$$

coded as a linear spline with empirically determined knots at ages 15, 55 and 70. As such, the only factors in the logistic model are age and i and the facility indicators. The model is fitted using Generalized Estimating Equations (GEE; Liang and Zeger, 1986) in order to account for the correlation within-patient across months. With over 6,000 facilities, it is difficult to estimate all parameters (i.e., including the facility indicators) simultaneously. Therefore, we break the fitting process into stages. At the first stage, we estimate the β vector by averaging 10 subgroups of approximately 600 facilities each. At the second stage, we then estimate the α_j (j=1, .., 6000) by fitting facility-specific intercept-only GEE models, with the linear predictor from the first stage, $\beta A_{ij,}$ serving as an offset. Per well-established GEE results (e.g., Liang and Zeger, 1986), the estimator of α_j is consistent for its target value, and follows a Normal distribution with standard error given by the robust 'sandwich' estimator computed via GEE. We can then compute $PPPW_j$ for each facility *j* as follows:

$$PPPW_j = \sum_i\sum_k exp(\alpha_j+ \beta A_{il}) / \{1 + exp(\alpha_j + \beta A_{il})\}. / n,$$

where *n* = total number of patient-months included in the overall study sample. The standard error of $PPPW_j$ is estimated through the Delta method; i.e., $SE(PPPW_j)=d_j \times SE(\alpha_j)$, where $d_j = \sum_i\sum_k exp(\alpha_j+ \beta A_{il}) / \{1 + exp(\alpha_j + \beta A_{il})\}^2 / n$.

We then carry out a two-sided Wald test (0.05 significance level) that $PPPW_j$=*PPPW*, where *PPPW* equals the national average percentage waitlisted. Note that Wald the test is based on the logit of $PPPW_j$, which is much more likely to follow a Normal distribution than $PPPW_j$ itself, due to the symmetry and lack of range restrictions of the transformed version.

Variable	**Primary Data Source**
Facility CCN #	CMS data sources[*1]
Reporting year and month	CROWNWeb
Waitlist status	Organ Procurement and Transplant Network (OPTN)
Date of Birth	CMS data sources[*1]
Date of First ESRD	Medical Evidence Form (CMS-2728)
Nursing home status on the Medical Evidence Form [*2]	Medical Evidence Form (CMS-2728) Question 17u and 22
Nursing home status in the current month [*2]	CMS Long Term Care Minimum Data Set (MDS)
Hospice status in the current month [*2]	CMS Hospice file

*1. CROWNWeb is the primary basis for placing patients at dialysis facilities and dialysis claims are used as an additional source. Information regarding first ESRD service date, death, waitlist status 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) and the Social Security Death Master File. For denominator exclusions, the Nursing Home Minimum Dataset and the Questions 17u and 22 on CMS Medical Evidence Form are used to identify patients in skilled nursing facilities. Additionally, a separate CMS file that contains final action claims submitted by Hospice providers was used to determine the hospice status. *2. Exclusion factors

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.

N/A

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?

Age adjustment was deemed necessary on clinical grounds. Although age alone is not a contraindication to transplantation, older patients are likely to have more comorbidities and be generally more frail thus making them potentially less suitable candidates for transplantation and therefore some may be appropriately excluded from waitlisting for transplantation. This may affect waitlisting rates for facilities with a substantially older age composition than the average.

A linear spline was used to model the effect of (continuous) age. The spline's knots were determined empirically using standard techniques. Specifically, as an initial step, we categorized age into as many groups as the data would sustain (15 groups). We then estimated the effect of categorical age, then plotted the age-category-specific parameter estimates against their respective category-specific median ages. The shape of this plot indicates age intervals within which the slope is approximately constant, and similarly suggests ages at which the slope changes. Using this procedure and examining the plot in Figure 2, knots at 15, 55 and 70 were suggested.

In response to the requirements for NQF's Trial Period for the incorporation of sociodemographic factors into quality measures, we investigated several patient and zip code level data elements (see list in 1.8). Sociodemographic factors included in the analysis were based on conceptual criteria and empirically demonstrated findings in the literature, which have shown that barriers to waitlisting exist among racial minorities, women and the poor. In addition, the particular patient and area level variables chosen were based on availability of data for the analyses. We were able to acquire individual area-level variables included in the Area Deprivation Index (ADI) developed by Singh and colleagues at the University of Wisconsin[1].

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?

Table 5. Coefficients and p-value in final PPPW model (note: a_+=max(a,0)), 2016

Covariate	**Coefficient**	**p-value**
Age	0.06	<.001
(age-15)$_+$	-0.08	<.001
(age-55)$_+$	-0.03	<.001
(age-70)$_+$	-0.23	<.001

[1] Singh, GK. Area deprivation and widening inequalities in US mortality, 1969–1998. Am J Public Health. 2003;93(7):1137–1143.

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.**

The table below shows the parameter estimates for model including all SDS/SES variables along with original covariates.

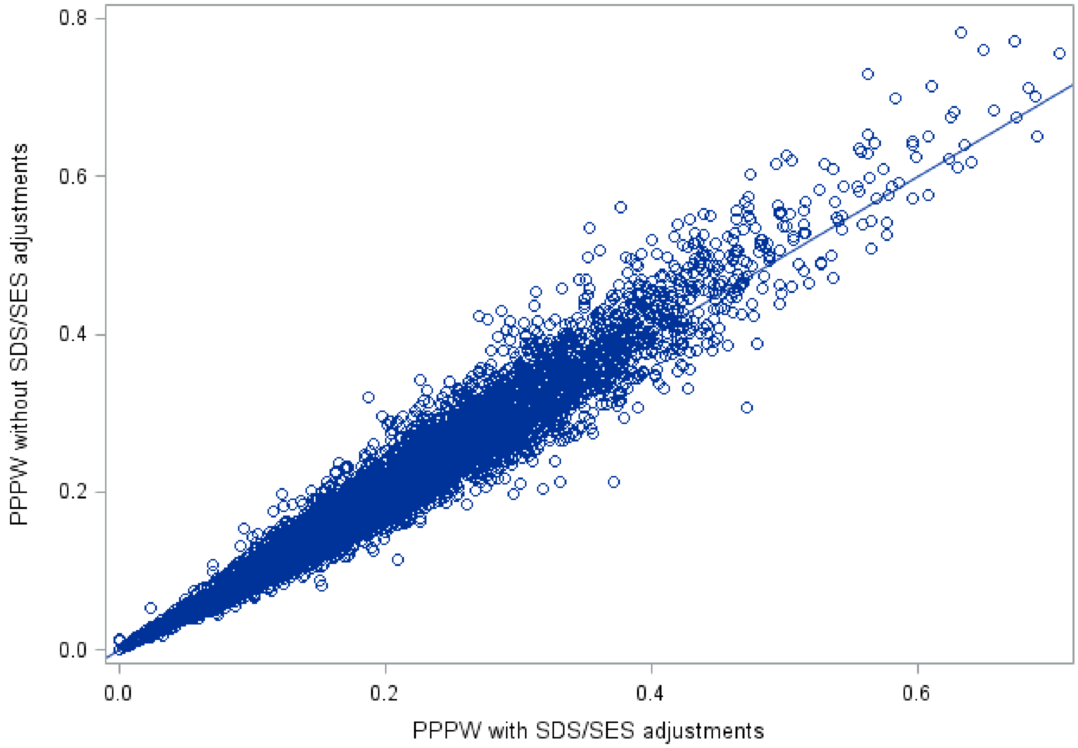
Table 6. Estimate and p-value of SES/SES variables, 2016

Covariate	**Estimate**	**P**
Sex		
Male	Reference	
Female	-0.08	<.001
Race		
White	Reference	
Non-White	0.03	0.008
Ethnicity		
Hispanic	0.11	<.001
Non-Hispanic	Reference	
Employment status		
Employed	0.66	<.001
Unemployed	-0.01	0.347
Retired/ Missing	Reference	
Medicare coverage		
Medicare as primary with Medicaid	Reference	
Medicare as primary without Medicaid	0.37	<.001
Medicare as secondary	0.29	<.001
Non-Medicare/missing	-0.63	<.001
ADI index	-1.03	<.001

Patient-level SDS/SES: Compared to male, female patients were less likely to be waitlisted (OR=0.92, p<.001). Hispanic patients were more likely to get waitlisted compared with non-Hispanic (OR=1.12, p<.001). Compared to retired/missing employment status patients, employed patients were more likely to get waitlisted (OR=1.93, p<.001); contrarily, unemployed patients were less likely to be waitlisted though the effect was not significant (OR=0.99, p=0.347). For insurance coverage, compared with Medicare as primary with Medicaid, patients with Medicare as primary without Medicaid and Medicare as secondary were more likely to be waitlisted (OR=1.45, p<.001; OR=1.34, p<.001), the non-Medicare/ missing group were less likely to get waitlisted (OR=0.53, p<.001).

Area-level SDS/SES: Patients in higher area-level deprivation (ADI), i.e. more deprived area, were less likely to be waitlisted (OR=0.36, p<.001).

Correlation between PPPWs with and without SDS/SES adjustments



The standard and SDS/SES-adjusted PPPW were highly correlated at 0.98 (*p*<.001).

Table 7. Flagging rates between original PPPW and PPPW adjusted for SES/SDS, 2016*

Standard PPPW	PPPW with SDS/SES adjustment			Total
	Better than expected	As expected	Worse than expected	
Better than expected	793	181	0	974 (14.75%)
As expected	91	5350	22	5463 (82.72%)
Worse than expected	0	44	123	167 (2.53%)
Total	884 (13.39%)	5575 (84.42%)	145 (2.20%)	6604

* Facilities with less than 11 patients were excluded.

After adjustment for SDS/SES, 338 facilities (5.1%) changed performance categories; 203 (3.1 %) performed worse after SDS/SES adjustment.

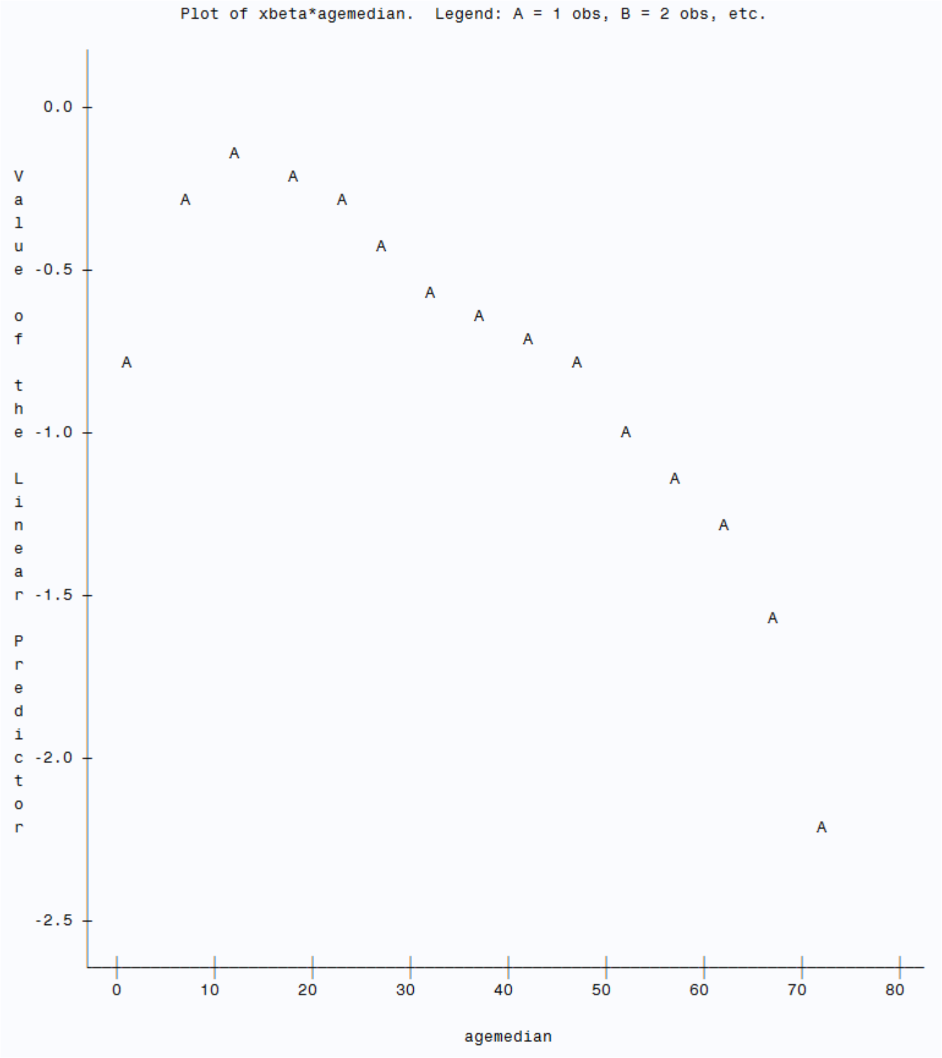
Patient level-variables such as employment, ethnicity, and health insurance had significant effects on waitlisting, as well as area-level variables. Although SDS/SES does affect waitlisting rates these were not included in the measure specification on biological/clinical grounds. Namely, there is no biological or clinical rationale to exclude patient groups on the basis of race, sex or economic status from transplantation as these groups still stand to substantially benefit from transplantation. Although barriers exist to waitlisting in these groups, it is expected that facilities should work towards helping such patients overcome those issues.

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**

Figure 2. Plot of age trend (linear predictor versus median of age)



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

The C-statistic (also known as the Index of Concordance) was 0.72. This indicates that the model correctly ordered 72% of the pairs of patient-months that were discordant with respect to the response variate. Month-specific C statistics were computed, in order to identify any trends by month in the model's discriminatory ability, and for computational ease.

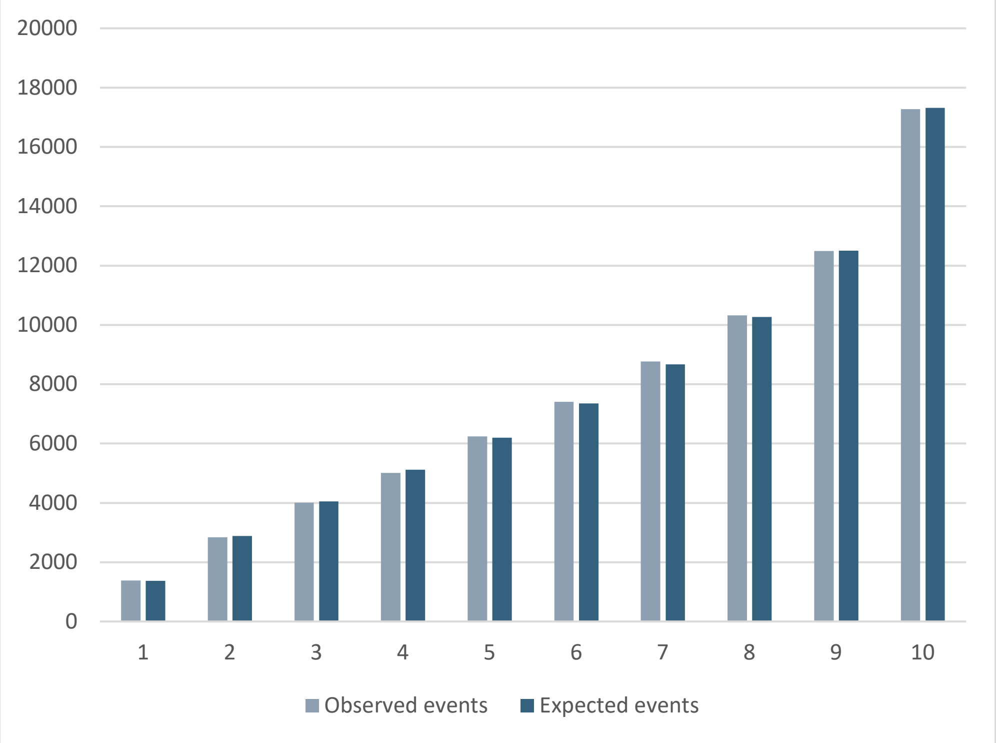
2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

The Hosmer-Lemeshow (H-L) statistic is defined strictly for independent trials, and months within-patient are expected to be highly correlated. We therefore chose to compute the H-L statistic in a month-specific fashion. No evidence of model mis-fit was detected for any month, with the p values being generally quite high (e.g., p=0.53 for January).

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

In Figure 3, we plot the key components of the Hosmer-Lemeshow test; namely, the observed and expected number of patients waitlisted by risk decile.

Figure 3. Observed and expected waitlist counts by risk decile



2b3.9. Results of Risk Stratification Analysis:

N/A

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*)

The plot in Figure 3 reveals that in no decile is there a practically important discrepancy between the observed number of waitlisted patients in a decile and that predicted by the model.

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)*

Since the distribution of waitlist rates are slightly skewed, logit transformation was used to reduce the skewness. Denote as the estimated waitlist rate for each facility, *j*=1,2, …,N. Set $\hat{g}_j = \log\frac{\hat{\varphi}_j}{1-\hat{\varphi}_j}$. So the formula for Z scores would be

$$\hat{Z}_j^g = \frac{\hat{g}_j - \overline{g(\hat{\varphi}_j)}}{SE\{\hat{g}_j\}}$$

where $g(\hat{\varphi}_j)$ is the average of the $\hat{g}_j$ national PPPW, and

$$SE\{\hat{g}_i\} = \frac{1}{\hat{\varphi}_j(1-\hat{\varphi}_j)} SE\{\hat{\varphi}_j\},$$

is the standard error after transformation and $SE\{\hat{\varphi}_j\}$, is obtained through the Delta method.

Then two-sided test with significant level 0.05 was used. Note that the reference distribution was Efron's empirical null, which essentially re-scales the critical value for the test statistic. The rescaling multiple is estimated by the slope (estimated via robust regression) correlating the empirical and theoretical Z score quantiles (e.g., with a multiple of 1 indicating that in fact no rescaling is required). Facilities are flagged if they have outcomes that are extreme when compared to the variation in national waitlist rate.

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*)

Table 8. Number and percentage of facilities by classification of the Waitlist Rate.*

Classification	N (%)	Median of PPPW
Better than expected	974 (14.7%)	0.37
As expected	5476 (82.8%)	0.17
Worse than expected	167 (2.5%)	0.04
Total	6617 (100%)	0.19

* Facilities with less than 11 patients were excluded.

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?*)

As is evident in Table 8, most facilities (82.8%) had a PPPW that was "As expected". Approximately 14.7% of facilities had a PPPW that was "Better than expected", while nearly 2.5% were "Worse than expected". This analysis demonstrates both practical and statistically significant differences in performance across facilities based on their proportion of patients placed on the transplant waitlist.

____

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*)

Age is the only adjustment variable in the PPPW measure. Since age was calculated using the date of birth and the reporting month, and date of birth was required in our Standard Analysis Data Files, no missing value in age was identified in the patient population.

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

3. Feasibility

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

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

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

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

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 a combination of electronic sources

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).

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.

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

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.

N/A

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

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Payment Program	

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

N/A

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 has gone through the process of being recommended for Dialysis Facility Compare (DFC), and will go through a Dry Run for DFC in July 2018, with the intention that the measure will be publicly reported in October 2019.

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 specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

As mentioned above, the measure will go through a Dry Run in July 2018, with the intention that it will be reported on DFC beginning on October 2019. The measure has also been reviewed by the NQF Measure Application Partnership, which is a precursor to being used in a payment program.

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.

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

N/A

4a2.3. Describe how the feedback described in 4a2.2.1 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

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

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 yet implemented in a public report program, so improvement could not be evaluated. CMS currently anticipates implementation of the Percentage of Prevalent Patients Waitlisted (PPPW). Once implemented, facility performance on the measure can be evaluated to determine if the measure has supported and detected quality improvement in promoting waitlisting for the prevalent population.

4b2. Unintended Consequences

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

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

5. Comparison to Related or 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.

5. Relation to Other NQF-endorsed Measures

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.

No

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

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?

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

N/A

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

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

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment Attachment: PPPW_Appendix-636582801711780633.pdf

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare and Medicaid Services

Co.2 Point of Contact: Sophia, Chan, sophia.chan@cms.hhs.gov

Co.3 Measure Developer if different from Measure Steward: University of Michigan Kidney Epidemiology and Cost Center

Co.4 Point of Contact: Jennifer, Sardone, jmsto@med.umich.edu, 734-936-5711-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

According to the CMS Measure Management System Blueprint, TEPs are advisory to the measure contractor. In this advisory role, the primary duty of the TEP is to suggest candidate measures and related specifications, review any existing measures, and determine if there is sufficient evidence to support the proposed candidate measures.

Stephen Pastan, MD

Emory University, Atlanta, GA

Amy Waterman, PhD

David Geffen School of Medicine, University of California, Los Angeles (UCLA), Los Angeles, CA

Todd Pesavento, MD

Comprehensive Transplant Center, Ohio State University, Columbus, OH

Sandra Amaral, MD, MHS

University of Pennsylvania, The Children´s Hospital of Philadelphia, Philadelphia PA

Ranjan Chanda, MD, MPH

Centennial Kidney Transplant Center, Nashville, TN

Mary Beth Callahan, ACSW, LCSW

Dallas Transplant Institute, Dallas, TX

Duane Dunn, MSW

DaVita Healthcare Partners Inc., Columbia, SC

Linda Wright, DrNP, RN, CNN, CCTC

Thomas Jefferson University Hospital, Philadelphia, PA

Robert Teaster, RN, MBA, CPTC, CPT

University of Virginia Medical Center, Charlottesville, VA

Chris Elrod, CCHT

Dialysis Clinic, Inc. (DCI) Chattanooga, TN

Nancy Scott

Dialysis Patient Citizens Education Center Washington, DC

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2018

Ad.3 Month and Year of most recent revision: 04, 2018

Ad.4 What is your frequency for review/update of this measure? Annually

Ad.5 When is the next scheduled review/update for this measure? 04, 2019

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: