

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (if previously endorsed): Click here to enter NQF number

Measure Title: Post-Discharge Appointment for Heart Failure Patients

Date of Submission: 12/23/2013

Type of Measure:

<input type="checkbox"/> Composite – STOP – use composite testing form	<input type="checkbox"/> Outcome (including PRO-PM)
<input type="checkbox"/> Cost/resource	<input checked="" type="checkbox"/> Process
<input type="checkbox"/> Efficiency	<input type="checkbox"/> Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. **If there is more than one set of data specifications or more than one level of analysis, contact NQF staff** about how to present all the testing information in one form.
- For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- For outcome and resource use measures,** section 2b4 also must be completed.
- If specified for **multiple data sources/sets of specifications** (e.g., claims and EHRs), section 2b6 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 (2b2-2b6) 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 20 pages (*including 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](#).

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

2a2. Reliability testing¹⁰ 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing¹¹ 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion;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

2b4. 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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; ^{14,15} and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For eMeasures, composites, and PRO-PMs (or other measures susceptible to missing data), 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.

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. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. 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.23)	Measure Tested with Data From:
<input type="checkbox"/> abstracted from paper record	<input type="checkbox"/> abstracted from paper record
<input type="checkbox"/> administrative claims	<input type="checkbox"/> administrative claims
<input checked="" type="checkbox"/> clinical database/registry	<input checked="" type="checkbox"/> clinical database/registry
<input type="checkbox"/> abstracted from electronic health record	<input type="checkbox"/> abstracted from electronic health record
<input type="checkbox"/> eMeasure (HQMF) implemented in EHRs	<input type="checkbox"/> eMeasure (HQMF) implemented in EHRs
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> other: Click here to describe

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

The data sample used for testing consisted of all eligible discharges in the American Heart Association (AHA) Get with the Guidelines-Heart Failure (GWTG-HF) registry during the study period. Hospitals that submitted fewer than 25 discharges during 2012 were excluded from the analyses.

Get With the Guidelines-Heart Failure (GWTG-HF) is the American Heart Association's collaborative performance improvement program, demonstrated to improve adherence to evidence-based care of patients hospitalized with heart failure. This program includes the measure we have submitted and is used in over 539 hospitals, with over 910,049 entered. Additional information on GWTG-HF is available at: http://www.heart.org/HEARTORG/HealthcareResearch/GetWithTheGuidelinesHFStrokeResus/Get-With-The-Guidelines-Heart-Failure_UCM_306087_SubHomePage.jsp

1.3. What are the dates of the data used in testing? The study sample included discharges from 1/1/2012 through 12/31/2012. In addition, we used data from 1/1/2011 through 12/31/2011 for temporal comparisons.

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.26)	Measure Tested at Level of:
<input type="checkbox"/> individual clinician	<input type="checkbox"/> individual clinician
<input type="checkbox"/> group/practice	<input type="checkbox"/> group/practice
<input checked="" type="checkbox"/> hospital/facility/agency	<input checked="" type="checkbox"/> hospital/facility/agency
<input type="checkbox"/> health plan	<input type="checkbox"/> health plan
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? *(identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample)*

For 2012: There were 432 hospitals included in the study sample, The volume of eligible patients at the hospitals included in the study sample ranged from 10 to 1798, with a mean of 204.09 (Standard Deviation, 177.42) and a median of 165 (interquartile range, 86-267). Additional details on the hospitals in the 2012 study sample are provided in the table below.

Hospital Characteristics	Post Discharge Appointment for Heart Failure		
	All	Yes	No
Region			
Northeast	31.16%	43.53%	56.47%
Midwest	19.75%	62.72%	37.28%
South	33.35%	46.61%	53.39%
West	15.73%	35.15%	64.85%
States (with >=5 hospitals)			
AL	2.17%	51.25%	48.75%
CA	7.03%	43.34%	56.66%
CO	1.55%	40.92%	59.08%
CT	1.53%	64.30%	35.70%
FL	4.20%	42.10%	57.90%
GA	0.77%	29.85%	70.15%
IL	3.80%	54.66%	45.34%
IN	3.10%	49.69%	50.31%
KY	0.54%	61.89%	38.11%
LA	1.44%	42.54%	57.46%
MA	1.15%	63.31%	36.69%
MI	1.21%	61.50%	38.50%
MO	1.55%	71.48%	28.52%
MS	1.09%	63.89%	36.11%
NC	5.79%	45.76%	54.24%

Hospital Characteristics	Post Discharge Appointment for Heart Failure		
	All	Yes	No
2012			
NJ	7.43%	18.64%	81.36%
NV	1.65%	20.08%	79.92%
NY	9.66%	42.97%	57.03%
OH	5.59%	71.49%	28.51%
PA	10.65%	57.00%	43.00%
SC	3.98%	65.42%	34.58%
TN	2.88%	34.67%	65.33%
TX	4.65%	31.72%	68.28%
UT	0.72%	10.49%	89.51%
VA	1.97%	58.35%	41.65%
WA	1.77%	22.68%	77.32%
WI	2.36%	63.07%	36.93%
WV	1.06%	39.87%	60.13%

For 2011: There were 445 hospitals included in the 2011 study sample, The volume of eligible patients at the hospitals included in the study sample ranged from 22 to 2299, with a mean of 209.93 (Standard Deviation, 199.84) and a median of 162 (interquartile range, 80-267). Additional details on the hospitals in the 2012 study sample are provided in the table below.

	Post Discharge Appointment for Heart Failure		
	All	Yes	No
2011			
Region			
Northeast	30.70%	13.84%	86.16%
Midwest	19.42%	24.56%	75.44%
South	34.18%	16.87%	83.13%
West	15.69%	8.83%	91.17%
States (with >=5 hospitals)			
AL	1.82%	24.43%	75.57%

	Post Discharge Appointment for Heart Failure		
2011	All	Yes	No
AZ	0.66%	3.92%	96.08%
CA	7.02%	7.30%	92.70%
CO	1.49%	12.81%	87.19%
CT	1.05%	16.97%	83.03%
FL	4.06%	8.70%	91.30%
HI	1.53%	18.98%	81.02%
IL	2.96%	17.81%	82.19%
IN	3.43%	5.80%	94.20%
LA	1.76%	10.82%	89.18%
MA	1.55%	21.58%	78.42%
MI	1.07%	19.18%	80.82%
MO	1.79%	33.55%	66.45%
MS	1.27%	31.36%	68.64%
NC	5.61%	16.00%	84.00%
NJ	8.41%	5.10%	94.90%
NV	1.92%	3.23%	96.77%
NY	8.31%	16.25%	83.75%
OH	5.18%	31.92%	68.08%
PA	10.65%	17.89%	82.11%
SC	3.58%	27.88%	72.12%
TN	2.54%	15.87%	84.13%
TX	5.64%	11.52%	88.48%
UT	0.90%	. %	100.00 %
VA	2.32%	13.96%	86.04%
WA	1.47%	18.96%	81.04%
WI	2.16%	28.98%	71.02%
WV	0.82%	12.94%	87.06%

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)

2011: There were 93,417 hospital stays (representing 85,025 patients) included in the study sample.

2012: There were 88,168 hospital stays (representing 81,339 patients) included in the study sample.

See tables below for details on the gender, age, race and insurance status of the patients included in the study samples for each year

	Post Discharge Appointment for Heart Failure		
	All	Yes	No
2012			
Total N Eligible	88,168	41,463	46,705
Age			
Mean	70.32	69.94	70.65
STD	14.59	14.60	14.57
>=70	55.69%	46.15%	53.85%
<70	44.31%	48.13%	51.87%
Male	53.79%	49.34%	50.66%
Female	46.21%	48.21%	51.79%
Race			
White	59.01%	48.08%	51.92%
Black	20.65%	53.76%	46.24%
Hispanic	6.66%	42.78%	57.22%
Asian	1.74%	47.00%	53.00%
Other	11.94%	32.56%	67.44%
Insurance			
Medicare	23.53%	49.53%	50.47%
Medicaid	4.89%	59.91%	40.09%
Other	11.68%	56.78%	43.22%

	Post Discharge Appointment for Heart Failure		
	All	Yes	No
None/Not documented/UTD	59.90%	43.09%	56.91%

	Post Discharge Appointment for Heart Failure		
	All	Yes	No
2011			
Total N Eligible	93,417	15,105	78,312
Age			
Mean	70.37	69.25	70.59
STD	14.63	14.62	14.62
>=70	55.84%	15.26%	84.74%
<70	44.16%	17.32%	82.68%
Male	53.27%	17.76%	82.24%
Female	46.73%	16.45%	83.55%
Race			
White	55.03%	17.19%	82.81%
Black	19.99%	18.85%	81.15%
Hispanic	6.38%	17.53%	82.47%
Asian	1.82%	15.33%	84.67%
Other	16.77%	9.20%	90.80%
Insurance			
Medicare	28.79%	22.84%	77.16%
Medicaid	6.02%	27.49%	72.51%
Other	13.98%	23.88%	76.12%
None/Not documented/UTD	51.20%	8.98%	91.02%

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.

The study sample described above was used for all data analyses.

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

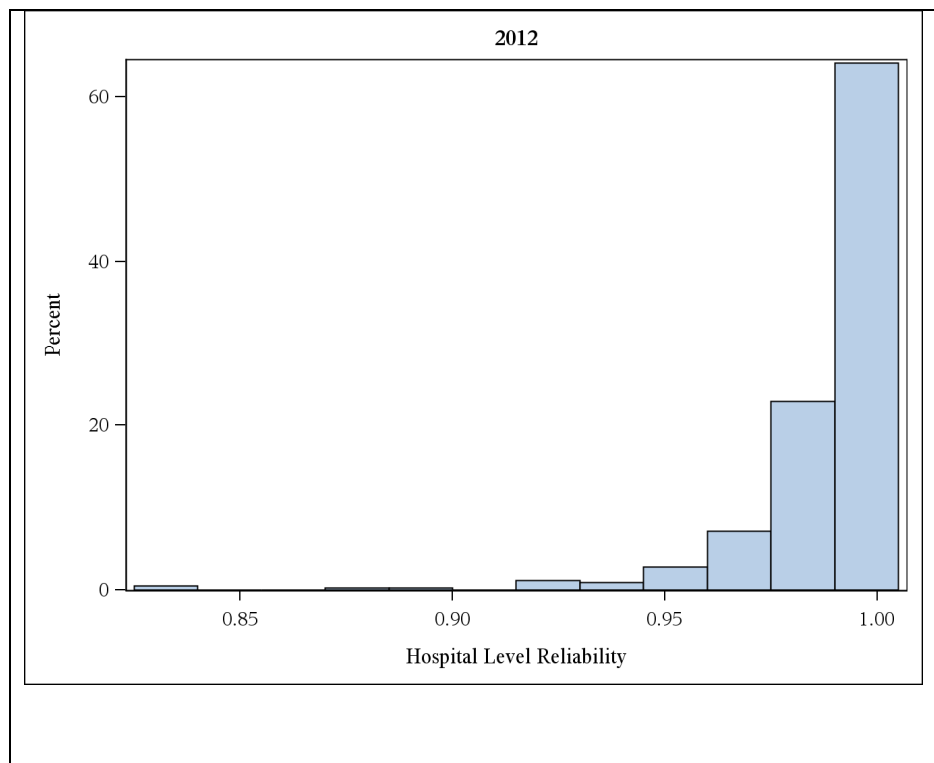
(describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used)

Reliability was established by estimating the Signal-to-Noise ratio* in the derivation cohort based on the GWTG database records from 2011 and 2012. After measure exclusions, a total of 181,711 hospital stays (representing 161,548 patients) were submitted by 492 facilities. An overall Signal-to-Noise ratio (SNR) was estimated among sites with at least 200 patients, as well as hospital-specific SNR estimates.

*Adams JL, Mehrotra A, McGlynn EA, *Estimating Reliability and Misclassification in Physician Profiling*, Santa Monica, CA: RAND Corporation, 2010. Available at: <http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html> Accessed December 22, 2013.

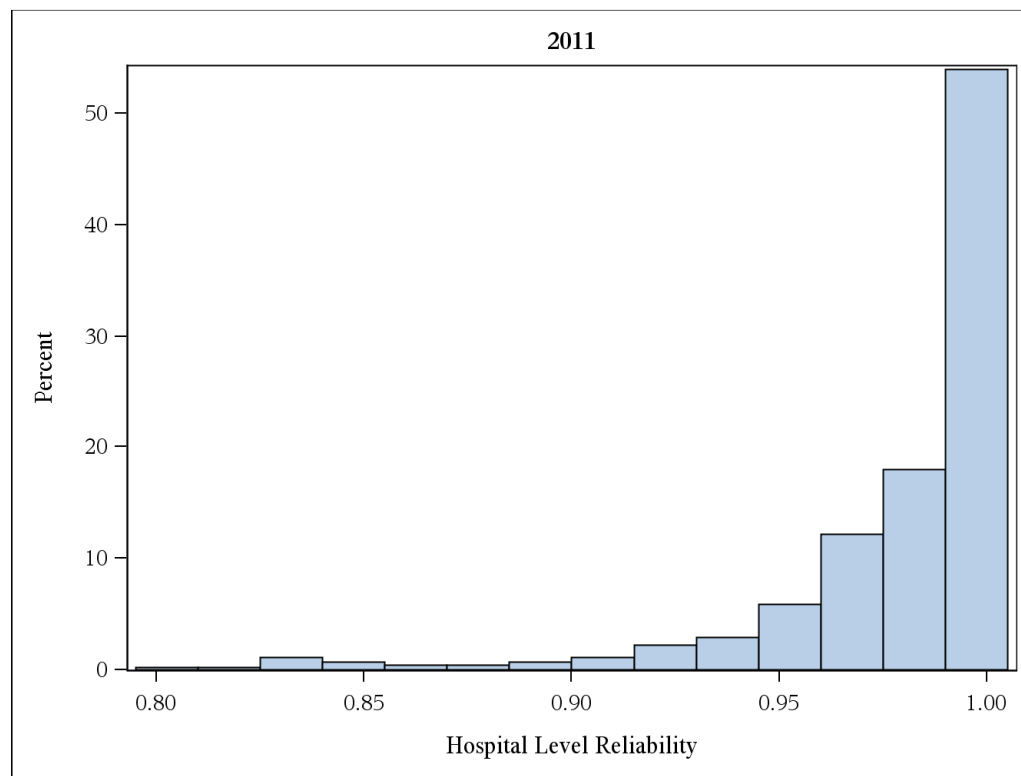
2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis)

2012 Signal-to-noise ratio: Overall Estimate (at 200 patients per site) = 0.988



Quantile	Estimate
100% Max	1.000000
99%	1.000000
95%	1.000000
90%	1.000000
75% Q3	0.997011
50% Median	0.993512
25% Q1	0.985460
10%	0.970530
5%	0.957419
1%	0.916090
0% Min	0.826882

2011: Signal-to-noise ratio: Overall Estimate (at 200 patients per site) = 0.985



Quantile	Estimate
100% Max	1.000000
99%	1.000000
95%	1.000000
90%	1.000000
75% Q3	0.999600
50% Median	0.991218
25% Q1	0.972543
10%	0.944815
5%	0.920100
1%	0.832819
0% Min	0.805809

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

The SNR estimates are high, which was somewhat surprising. However, we had considerable variability in the measure rates across hospitals (see distribution across hospitals in figures and tables below for 2011 and 2012. Information on 2011-2012 overall is available in the full testing report in Appendix A-1 in the supplemental materials. This means that we should be able to estimate reliability well and the parameters from the beta-binomial model used to estimate the SNR estimates fit the observed data pretty well.

Demographic comparisons between the derivation and validation cohorts were essentially comparable. Thus, despite different years of submission, the GWTG-HF patient population eligible for this measure is highly consistent over time. This observation, as well as the large number of records submitted by many institutions in both years, supports the contention that this population is likely to be highly representative of the HF population. Because we had considerable variability in the measure rates across hospitals, we should be able to estimate reliability well and the results of this analysis suggest that the measure is likely reliable for this population.

Reference:

Adams JL, Mehrotra A, McGlynn EA, *Estimating Reliability and Misclassification in Physician Profiling*, Santa Monica, CA: RAND Corporation, 2010. www.rand.org/pubs/technical_reports/TR863. (Accessed on December 22, 2013)

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (may be one or both levels)

- ☐ **Critical data elements** (data element validity must address ALL critical data elements)
- ☐ **Performance measure score**
 - ☐ **Empirical validity testing**
 - ☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance)

2b2.2. 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)

Content validity for this measure was systematically assessed by expert work group members during the development process during extensive discussion and a final confidential vote. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and concurrent

formal peer review process. Additionally, comments were solicited from a panel of consumer, purchaser, and patient representatives convened by the AMA-PCPI specifically for this purpose. All comments received were reviewed by the expert work group and the measures were adjusted as needed. Additionally, the measure underwent review and approval by the Board of Trustees of the ACC and the Science Advisory and Coordinating Committee of the AHA, as well as review and voting by the PCPI membership. Members of the expert work group that developed the measure included: Robert O. Bonow, MD, MACC, FAHA, FACP (Co-Chair) (cardiology); Theodore G. Ganiats, MD (Co-Chair) (family medicine; measure methodology); Craig T. Beam, CRE (patient representative); Kathleen Blake, MD (cardiac electrophysiology); Donald E. Casey, Jr., MD, MPH, MBA, FACP, FAHA (internal medicine); Sarah J. Goodlin, MD (geriatrics, palliative medicine); Kathleen L. Grady, PhD, APN, FAAN, FAHA (cardiac surgery); Randal F. Hundley, MD, FACC (cardiology, health plan representative); Mariell Jessup, MD, FACC, FAHA, FESC (cardiology, heart failure); Thomas E. Lynn, MD (family medicine, measure implementation); Frederick A. Masoudi, MD, MSPH (cardiology); David Nilasena MD, MSPH, MS (general preventive medicine, public health, measure implementation); Ileana L. Piña, MD, FACC (cardiology, heart failure); Paul D. Rockswold, MD, MPH (family medicine); Lawrence B. Sadwin (patient representative); Joanna D. Sikkema, MSN, ANP-BC, FAHA (cardiology); Carrie A. Sincak, PharmD, BCPS (pharmacy); John Spertus, MD, MPH (cardiology); Patrick J. Torcson, MD, FACP, MMM (hospital medicine); Elizabeth Torres, MD (internal medicine); Mark V. Williams, MD, FHM (hospital medicine); John B Wong, MD (internal medicine).

Face validity of the measure score was systematically assessed as follows:

After the measure was fully specified, members of three existing committees, one at the ACC, one at AHA and one joint ACC/AHA, with expertise in general cardiology, interventional cardiology, heart failure, electrophysiology and quality improvement, outcomes research, informatics and performance measurement who were not involved in development of the measure, were asked to review the measure specifications and rate their agreement with the following statement:

“The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.” The respondents recorded their rating on a scale of 1-5, where 1= Strongly Disagree; 3=Neither Agree nor Disagree; 5= Strongly Agree

There were 17 committee members who completed the survey.

2b2.3. What were the statistical results from validity testing? (e.g., correlation; t-test)

The results of the expert panel rating of the validity statement were as follows:

N = 16; Mean rating = 3.94 and 69% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality

Frequency Distribution of Ratings

1 - 0 (Strongly Disagree)

2 - 1

3 - 4 (Neither Agree nor Disagree)

4 - 6

5 - 5 (Strongly Agree)

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (*i.e., what do the results mean and what are the norms for the test conducted?*)

The measure was judged to have moderate to good face validity by both its clinical importance and the group of experts asked to rate it. The majority of experts agreed that the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — **skip to section 2b4**

2b3.1. 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*)

As specified, only patients discharged to ambulatory care (home/self care) or home health care with a principal discharge diagnosis of heart failure are eligible for this measure, since there may be valid reasons that a post-discharge appointment for HF was not scheduled for patients discharged to other settings, e.g., a nursing home or another acute care hospital. Therefore in the calculation of the measure, all patients discharged to other settings are excluded.

In the context of physician performance measurement, exceptions are the mechanism used to remove patients from the denominator of a performance measure when a patient does not receive a therapy or service AND that therapy or service would not be appropriate due to specific reasons for which the patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment and individual patient characteristics. For this measure, the following exceptions are allowed:

- Medical reason(s) for not documenting that a follow up appointment was scheduled (eg, patients who expired, patients who left against medical advice (AMA) or discontinued care).
- Patient reason(s) for not documenting that a follow up appointment was scheduled (eg, international patients, patients from state and/or local corrections facilities for whom scheduling the appointment is prohibited)

We examined the overall incidence of exceptions in the study samples for 2011 and 2012.

2b3.2. 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*)

In 2012, a medical or patient exception was reported in only 1.32% of discharges and in 2011, a medical or patient exception was reported in only 0.42% of discharges for patients who would otherwise be eligible for the measure.

2b3.3. 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

The overall frequency of exceptions is extremely low and indicates that providers are not using them to game the measure. Given the low rates, exceptions are not likely to bias performance results. We believe both of the exceptions must stay in the measure to allow for those circumstances where it is unnecessary, unwanted by the patient, or impossible for other valid reasons to schedule a post-discharge appointment.

2b4. 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 2b5.

2b4.1. What method of controlling for differences in case mix is used?

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

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient 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 and not related to disparities)

2b4.4. What were the statistical results of the analyses used to select risk factors?

2b4.5. 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 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (e.g., c-statistic, R-squared):

2b4.7. Statistical Risk Model Calibration Statistics (e.g., Hosmer-Lemeshow statistic):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. 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)

2b4.11. 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*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. 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)

We examined variation in hospital performance on this measure and provide additional information about potential disparities based on sex, age, race and a number of other patient factors. The full testing report with information on all patient characteristics is available in Appendix A-1 and summarized below.

2b5.2. 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)

2011

	White		Black		Hispanic		Asian		Other	
Description	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure
N	391	391	360	360	302	302	191	191	364	364
Mean	131.49	19.40%	51.87	17.33%	19.74	16.46%	8.91	13.43%	43.04	14.85%
Std Deviation	139.53	23.37%	74.04	23.74%	46.99	26.73%	17.24	25.40%	80.46	25.05%
100% Max	1427	100.00%	493	100.00%	593	100.00%	126	100.00%	524	100.00%
99%	637	100.00%	332	100.00%	193	100.00%	94	100.00%	408	100.00%
95%	404	66.67%	211	65.65%	96	90.00%	53	70.00%	208	69.23%

	White		Black		Hispanic		Asian		Other	
Description	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure
90%	286	52.17%	158	50.48%	50	50.00%	20	50.00%	148	50.00%
75% Q3	178	31.11%	61	28.73%	18	25.00%	8	18.87%	40	23.44%
50% Median	90	9.76%	22	6.68%	5	0.00%	3	0.00%	7	0.00%
25% Q1	40	0.91%	7	0.00%	2	0.00%	1	0.00%	2	0.00%
10%	18	0.00%	2	0.00%	1	0.00%	1	0.00%	1	0.00%
5%	8	0.00%	1	0.00%	1	0.00%	1	0.00%	1	0.00%
1%	1	0.00%	1	0.00%	1	0.00%	1	0.00%	1	0.00%
0% Min	1	0.00%	1	0.00%	1	0.00%	1	0.00%	1	0.00%

2012

	White		Black		Hispanic		Asian		Other	
Description	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure	Volume	Post-Discharge Appointment for Heart Failure
N	395	395	363	363	292	292	212	212	333	333
Mean	131.72	46.39%	50.16	44.88%	20.10	46.08%	7.23	45.60%	31.62	41.76%
Std Deviation	124.83	32.89%	71.64	34.37%	42.22	40.36%	14.55	40.57%	69.24	38.43%
100% Max	1094	100.00%	469	100.00%	457	100.00%	136	100.00%	521	100.00%
99%	612	100.00%	376	100.00%	185	100.00%	72	100.00%	332	100.00%
95%	388	98.37%	194	100.00%	84	100.00%	31	100.00%	163	100.00%
90%	292	90.91%	139	92.31%	61	100.00%	16	100.00%	89	100.00%
75% Q3	178	75.51%	60	75.00%	17	89.79%	6	98.15%	20	75.00%
50% Median	95	46.88%	22	45.83%	6	41.05%	2	50.00%	6	35.58%
25% Q1	47	14.96%	6	9.09%	2	0.00%	1	0.00%	2	0.00%
10%	19	0.90%	2	0.00%	1	0.00%	1	0.00%	1	0.00%
5%	10	0.00%	1	0.00%	1	0.00%	1	0.00%	1	0.00%
1%	1	0.00%	1	0.00%	1	0.00%	1	0.00%	1	0.00%
0% Min	1	0.00%	1	0.00%	1	0.00%	1	0.00%	1	0.00%

2b5.3. 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?)

Demographic comparisons between the derivation and validation cohorts were essentially comparable. Thus, despite different years of submission, the GWTG-HF patient population eligible for

this measure is highly consistent over time. This observation, as well as the large number of records submitted by many institutions in both years, supports the contention that this population is likely to be highly representative of the HF population. Because we had considerable variability in the measure rates across hospitals, the variations can be assumed to represent true differences in performance. We believe these differences are clinically meaningful as hospitals in the lowest quartile showed substantial and meaningful differences in their performance on this measure when compared with hospitals in the top quartile.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This criterion is directed 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). **If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of 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*)

2b6.2. 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*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of 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*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. 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*)

In GWTG-HF, missing values are interpreted as 'No' for most of variables. The overall incidence of missing data is very low, so we assume that there is no missing data. However we did exclude discharges

where the discharge date was missing, since we could not confirm in these cases that the date of follow up appointment was after the date of discharge. For 2011, we excluded 125 discharges for this reason and, in 2012, just 51. It's challenging to distinguish real missing vs 'No'. While this is a logical approach for handling missing data, there are likely missing data (i.e. poor documentation) across hospitals. However, this is not a threat to the validity of the measure, as the ability to schedule and more clearly document post-discharge appointment for eligible patients is clearly under the locus of control of the hospital. We believe that if this is an endorsed measure and is used to assess the quality of HF care, that there will be increasing pressure on providers both to ensure that a follow up appointment is scheduled and to clearly document it, facilitating the transition from the inpatient to outpatient setting and improving the quality of care provided to patients.

2b7.2. 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*)

Given our assumptions, noted above, we did not conduct an empirical analysis of the frequency or distribution of missing data. For this measure, missing data represents a failure to schedule a post-discharge appointment.

2b7.3. 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*)

Our assumption is that there is no missing data and that providers with a very low rate of post-discharge appointments either failed to schedule them, or failed to document that they did so. Either way, we do not believe that any biases are introduced in assessing hospital performance and endorsement of this measure would lead to improved care.