**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** NQF 0139

**Measure Title**: National Healthcare Safety Network Central Line Associated Bloodstream Infection

**Date of Submission**: 1/23/2019

**Type of Measure:**

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

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * 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**](#Note10) 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**](#Note11) 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**](#Note12)  **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**](#Note13)  **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**](#Note14)**,**[**15**](#Note15) 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**](#Note16) **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: NHSN Primary BSI collection form; NHSN Denominator for ICU form; NHSN Denominator for NICU form; NHSN Denominator for Specialty Care Area/Oncology Form | 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 NHSN data set that was used was drawn from the NHSN database, which is an aggregation of data that healthcare facilities throughout the US submit, much of which is data required for submission to NHSN by state and federal mandates or both.

CDC NHSN used 2015 **healthcare-associated infection (**HAI) incidence and risk factor data to develop new predictive models for CLABSI and other HAI’s. The number of facilities in 2015 reporting CLABSI data includes: 3550 acute care hospitals (ACH), 489 long term acute care hospitals (LTACH), 662 inpatient rehabilitation facilities (IRF) throughout the US national database.

Please refer to the SIR Guide (p.4) at:

[**https://www.cdc.gov/nhsn/pdfs/ps-analysis-resources/nhsn-sir-guide.pdf**](https://www.cdc.gov/nhsn/pdfs/ps-analysis-resources/nhsn-sir-guide.pdf)

**1.3. What are the dates of the data used in testing**? January 2015 – December 31, 2015

**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*)

CLABSI data is reported to NHSN from over 4701 facilities in all 50 states, the District of Columbia, and several US territories. In 2015, 26% of hospitals reporting CLABSI data have fewer than 50 beds, 39% have between 51 and 200 beds, and 34% have more 200 beds.

**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*)   
Facilities reporting CLABSI data to NHSN do not report a count of patients under surveillance. The number of central line days is reported, as described in the measure submission. In 2015, 23,367,134 central line days were reported by acute care participating facilities. The central line day counts are reported by patient care location in the hospital and are not stratified by patient level factors such as age, race, and sex. All ages, all races and all diagnoses are included.

**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**.

Reliability testing is provided via 2015-2016 state studies; validity testing provided via **working group of the Healthcare Infection Control Practices Advisory Committee** (HICPAC) which developed a definition of the Mucosal Barrier Injury (MBI) instituted as part of the CLABSI surveillance definitions to improve construct validity; sample used to test CLABSI risk models consist 4701 number of facilities reporting CLABSIs in 2015.

**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.

No patient-level sociodemographic variables are used in the measure and none were available for analysis. There is no compelling evidence available supporting association between social risk factors and CLABSIs. Instead, they provide evidence that social risk factors are associated with an increased risk of chronic disease conditions, suboptimal care for those conditions, compromised functional status, exposure to nursing homes, and colonization with bacterial pathogens. While these associations may be meaningful, they do not establish a direct relationship between social risk factors and HAIs.

**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*)

See section 2b1 for validity testing of data elements.

As per NQF email “…data element validity testing may serve as a demonstration of data element reliability.” Please see section 2b1.2 for demonstration of data element reliability.

**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*)

As per NQF email “…data element validity testing may serve as a demonstration of data element reliability.” Please see section 2b1.3 for statistical results from reliability testing.

**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?*)

As per NQF email “…data element validity testing may serve as a demonstration of data element reliability.” Please see section 2b1.4 for interpretation of data element 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 performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**CLABSI definition and criteria are unchanged from prior submission, with the exception of the addition of the MBI exclusion, and definition and criteria were reviewed by the Healthcare Infection Control Practices Advisory Committee (HICPAC) panel using Delphi process which culminated with definition and criteria.**

MBI exclusion was established using Delphi process by **the Healthcare Infection Control Practices Advisory Committee (HICPAC)** Subject Matter Expert (SME) panel with 100 % consensus on the MBI definition and criteria.

Validity of inclusionary criteria, namely lab confirmed BSI was not retested for validity purposes.

**The Healthcare Infection Control Practices Advisory Committee (HICPAC) is a federal advisory committee chartered to provide advice and guidance to the Centers for Disease Control and Prevention (CDC) and the Secretary of the Department of Health and Human Services (HHS) regarding the practice of infection control and strategies for surveillance, prevention, and control of healthcare-associated infections, antimicrobial resistance and related events in United States healthcare settings.**

**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)*

Validity testing of critical data elements is performed by many of the state health departments that have implemented mandatory reporting of CLABSI data to the state using NHSN as the data entry system and the source of case definitions and surveillance methodology. Trained state health department validators apply NHSN CLABSI definition criteria in medical record reviews of records that were compiled during the stay in which patients reportedly met criteria of the CLABSI definition. The validator’s determination of whether or not the patient in question had a CLABSI is compared to the facility’s determination. Sensitivity, specificity, positive predictive value, and negative predictive value are then calculated. As part of the validation process, some state health departments validate counts of central line days through structured interviews with personnel who collect and report these data to NHSN to ensure that correct data collection methodology is used.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Year of data validated | Sensitivity (%) | Specificity (%) | Positive Predictive Value (%) | Negative Predictive Value (%) |
| New Hampshire | 2014-2015 | 80.6 | 98.9 | 96.2 | 93.7 |
| Texas | 2015 | 83.2 | 99.1 | 95.9 | 95.8 |
| Vermont | 2015 | 100 | 100 | 100 | 100 |
| Alabama | 2017 | 96.7 | 99.7 | 98.9 | 99.2 |
| Georgia | 2017 | 80.3 | 98.7 | 94.2 | 94.9 |

**External validation of NHSN CLABSI data has been conducted by at least 5 states since 2015 (NH, TX, VT, AL, GA), using different sampling methods. These validations indicated a pooled mean sensitivity of 87.5% (range: 80.3%-100%), specificity of 99.3% (range: 98.7% - 100%), positive predictive value of 96.9% (range: 94.2% - 100%) and negative predictive value of 96.9% (range: 93.7% - 100%). External validation across the 5 states consisted of 2,594 chart reviews and of these 81 cases were incorrectly classified, yielding an overall classification error rate of 3.1%.**

**Chart reviews were conducted by trained auditors across the 5 state health departments. These audits identified 520 CLABSI events that should have been reported and among those 463 events were correctly reported by healthcare facilities (57 missed events). Three states (GA, NH, TX) identified incorrect secondary BSI attribution as a reason for failure to report CLABSI events - the cases did not met criteria for a primary source of infection so the BSI should have been reported as a primary CLABSI. This resulted in at least 5 underreported events. One state (NH) noted that one CLABSI was missed due to a symptom being missed by the facility during chart review. One state (AL) identified that underreporting occurred as a result of positive test results not being reported for CLABSI surveillance. One state (GA) identified a failure to report CLABSI events that occurred outside the repeat infection timeframe from a prior CLABSI event.**

**Among the 2,074 charts that were identified as not meeting the NHSN CLABSI definition, 2,050 charts were correctly called as “CLABSI negative” by the healthcare facilities, thereby leading to 24 over reported CLABSIs. Two states (NH, TX) identified incorrect secondary BSI attribution as a reason for overcalling CLABSI events - the BSI should not have been reported because the cases met criteria for a primary source of infection to which the BSI should have been attributed as secondary. This resulted in at least one over reported CLABSI event. States also reported reasons for errors in reporting both under- and over reported CLABSI events that included misidentifying the date of event or location of attribution (GA), inexperienced infection preventionist (AL), and “misinterpreting rules” of surveillance (AL).**

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

Validity of exclusionary criteria, mainly MBI, was established using Delphi process of SME panel with 100% consensus on the MBI definition and criteria.

Validity of inclusionary criteria, namely lab confirmed BSI was not retested for validity purposes.

**Validations of critical data elements performed by external validators indicated a pooled mean sensitivity of 87.5% (range: 80.3%-100%), specificity of 99.3% (range: 98.7% - 100%), positive predictive value of 96.9% (range: 94.2% - 100%) and negative predictive value of 96.9% (range: 93.7% - 100%).**

**2b1.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

**The results of the Delphi process and consensus of exclusionary criteria, via expert review reaffirm the CLABSI measure is valid for use as a quality measurement. The SIR is based on the standardized mortality ratio, which is a widely accepted method for summarizing the predicted to observed ratio. The CLABSI SIR can distinguish good from poor quality. In some places where large scale CLABSI prevention programs have been implemented over the past several years, significant reductions in the CLABSI SIR have been seen. However, there are still facilities with significantly high CLABSI SIRs, indicating that they have not made progress in reducing CLABSI (high SIRs indicating poor quality) while other facilities have made progress in reducing CLABSI and see significantly low CLABSI SIRs (low SIRs indicate better quality). The CLABSI SIR is used by many state health departments in public reporting of HAI data, and the Centers for Medicare and Medicaid Services (CMS) has included the CLABSI SIR in its Hospital Inpatient Quality Reporting Program and Hospital Value Based Purchasing Program.**

**The CLABSI SIR is only calculated when sufficient denominator data has been reported, i.e. when the number of predicted CLABSIs is greater than 1. In order to allow for an assessment of CLABSI experience in facilities with lower exposure to urinary catheters, the ARM is used. The ARM uses statistical techniques to adjust for lower exposure to urinary catheters, in addition to other risk factors, and produces a measure that is interpreted similarly to the SIR.**

Norms have not been established however we have a series of studies that show agreement. Very High specificity, PPV and NPV and high sensitivity.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**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*)

**Not applicable.**

**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*)  
**Not applicable.**

**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*)  
**Not applicable.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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*** [***2b5***](#section2b5)***.***

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

**No risk adjustment or stratification**

**Statistical risk model with**  **risk factors relevant to the facility type**

**Stratification by**  **risk categories**

**Other,** Click here to enter description

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

**The risk modeling was conducted using negative binomial regression, in which risk factors were evaluated by both univariate and multivariate modeling steps. Univariate models were fist constructed to evaluate the relationship between each risk factor and the CLABSI incidence rate.**

**For detailed specifications of the risk model please refer to the SIR Guide (p. 5) at:**

[**https://www.cdc.gov/nhsn/pdfs/ps-analysis-resources/nhsn-sir-guide.pdf**](https://www.cdc.gov/nhsn/pdfs/ps-analysis-resources/nhsn-sir-guide.pdf)

**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**.

**Not applicable.**

**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?

**In the interest of minimizing reporting burden, denominator data are aggregate data at the patient care level. As a result, the candidate risk factor data available are descriptive characteristics for patient care locations and healthcare facility. To risk adjust the CLABSI SIR, national NHSN data is analyzed to assess for differences in rates between different patient care locations (ICU, ward, different specialty types, etc) within the data. Additional facility level characteristics (bedsize, affiliation with a medical school, etc) are included in the analysis.**

**In the risk adjustment for the CLABSI ARM, national NHSN data is used to produce a negative binomial risk model that includes patient care location type, medical school affiliation, facility bedsize, and central line device utilization.**

**Model selection was used with variables added if p <0.05. Order of variables were included in the model was based on a combination of deviance, log likelihood and Akaike information criterion statistics.**

**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)**

**No social risk factors applied in the modeling.**

**Due to the paucity of evidence to support social risk factors and data burden data collection for risk adjustment purposes, social risk factors are not collected in NHSN for all patients in the patient population; therefore, these variables are not available in NHSN to be used for risk adjustment modeling.**

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

**Variables were eligible for entering the model at p-value=0.25 and retaining in the model at p-value=0.05 significance level. Factors were entered into a multivariate model using forward selection, based on the lowest Wald Chi-square value. Goodness of fit was assessed at each modeling step using the Akaike Information Criterion (AIC) statistics. The final model resulting from forward selection was confirmed via backwards elimination, in which each variable was sequentially removed based on the highest p-value.**

**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.**

**Did not include social risk factors.**

**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*)  
**Model validation steps:**

1. **For each negative binomial regression model to be validated, produce a table of the regression parameters from the final model**
2. **Generate at least 100 new replicate samples using “sampling with replacement” from the original dataset so that each replicate sample contains the same number of observations as the original dataset**
3. **Fit the final model to each of those new replicate samples and store the regression parameters**
4. **This will produce a set of regression parameters as defined in the original final model for each model fit to each new replicate sample**
5. **Produce a distribution of each regression parameter across all the results from the at least 100 new replicate samples**
6. **Construct an empirical “percentile-based” confidence range using the 2.5 and 97.5 percentile for each parameter**
7. **Consider the model validated among all parameters if the respective confidence range does not include null value**

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**See 2b3.7

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Negative binomial model discrimination and calibration were performed using a combination of deviance, log likelihood and Akaike information criterion statistics. Markov chain Monte Carlo sampling methods inherently rely on large scale simulation to produce posterior parameter estimates evaluated using trace plots and highest probability density intervals. In addition, Markov chain convergence, sampling, and stationarity were assessed using Geweke, Raferty-Lewis and Heidelberger-Welch diagnostics, respectively.

Negative Binomial model calibration was further assessed by calculating the root mean squared error (RMSE) between the observed and model predicted values for the final versus null model across 1000 bootstrap samples. The average RMSE for the final model was 1.432 compared to 1.504 for the null model and demonstrates a 5% improvement.

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
**See 2b3.7**

**2b3.9. Results of Risk Stratification Analysis**:

See 2b3.4a

**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*)  
Both risk adjustment methodologies (stratification based on patient care location type and facility-level factors for the CLABSI SIR and risk modeling using similar factors for the ARM) allow for adequate controlling of factors that can lead to differences in CLABSI risk for patients in acute care hospitals.

**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*)

Not Applicable

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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)*

**NHSN uses the mid p exact test for determining statistically significant differences in performance measurement. This test is applied to facility-specific performance summary statistics. CDC calculates these summary statistics i.e. the CLABSI SIR and ARM, to identify variation from a predicted occurrence of CLABSI based on the experience of a standard population, as well as an assessment of the magnitude of that variation (for example, an SIR of 2.0 indicates a level of occurrence two times higher than what would be predicted). The measures are produced with a confidence interval that can be used to assess the likelihood that the SIR or ARM occurs within a specified range. The confidence interval can be used to assess the SIR or ARM compared to its nominal value of 1.0 (where the number of observed equals the number of predicted CLABSIs).**

**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*)

**Published data from the CDC national and state 2016 HAI progress report shows that 2,345 ACH reported sufficient data to generate a CLABSI SIR in 2016. Approximately 349 healthcare facilities (~14.88%) had SIRs that were statistically significantly less than 1.0, indicating that the facility reported fewer CLABSIIs than predicted. Approximately 169 healthcare facilities (~7.21%) had SIRs that were statistically significantly greater than 1.0, indicating that the facility reported more CLABSIs than predicted.**

**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?*)  
The SIR and the ARM can both be used to assess differences in performance. Facilities that have SIRs significantly lower than 1 are succeeding in preventing CLABSI. Facilities with SIRs that are significantly higher than 1 may not have implemented CLABSI prevention efforts and are potential targets for interventions to improve prevention practices.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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*)

**Not applicable.**

**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*)

**Not applicable.**

**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*)

**Not applicable.**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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*)

Healthcare facilities that submit quality measure data to NHSN for CLABSI and other HAIs must submit all data required for measure calculation; otherwise their data cannot be successfully submitted to NHSN. Within the NHSN reporting system, facilities are prompted each month that they have entered infection (numerator) data but no central line count (denominator) data and vice versa to ensure that monthly data submission is complete for each location that is reported.

Facilities are required to verify if no CLABSI events occurred for an inpatient care unit and month.

**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*)

All CLABSI numerator and denominator data submitted to NHSN must be complete or the data submission is not accepted by NHSN. As a result there is no missing data for which distributions or other characteristics can be tested.

**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*)

NHSN does not produce results pertaining to systematic missing data because the system requires that all data submissions include data used to calculate measure results.