Risk Adjustment and Risk Stratification in Quality Measurement

This document provides information about risk adjustment and risk stratification. Risk adjustment promotes fair and accurate comparison of health care outcomes across measured entities. Risk stratification allows for comparison of health care performance within peer groups of measured entities rather than across all measured entities. This document offers insight into risk adjustment and risk stratification models and supplements the information found in the *Measure Specification* content on the *Measures Management System (MMS) Hub*.

1. Background

The purpose of risk adjustment and risk stratification is to decompose the measured entity-level variation into factors that are and are not correlated with (meaning, are independent of) the quality construct. Risk adjustment refers to the inclusion of risk factors associated with a measure score in a statistical model of measured entity performance captured at the person, facility, community, or other levels. Risk stratification groups patients or resource services with similar characteristics and then calculates multiple performance scores for the measured entity. Measure developers often risk adjust and/or stratify outcome measures and cost and resource use measures, however not all outcome measures need risk adjustment or risk stratification.

Risk adjustment at the person level, also referred to as case-mix adjustment, aims to answer the question: "How would the performance of various units compare if hypothetically they had the same mix of patients?" (National Quality Forum [NQF], 2014, p. v). Thus, risk adjustment increases the likelihood of fair comparison of measured entity performance, which is to compare "apples with apples." It involves controlling for confounding factors -- meaning systematic differences within the population of interest -- in the modeling of measured entity performance. Confounding factors may be clinical (e.g., types, number, or severity of conditions), demographic (e.g., age, gender), functional (e.g., dementia) and/or social (e.g., income, education, geography) in nature.

Taking confounding factors into account could prevent the model from incorrect specification or the estimates of performance scores from bias. The variation in measured entity-level (e.g., clinician or facility) performance may be due to variation in quality or variation in factors independent of quality

(e.g., factors like the age or severity of illness of patients). Independent of quality means the clinician treats the patients exactly the same way, but patients who have the factor (older or sicker) have worse outcomes than patients who do not (younger or less sick). In such a circumstance, selecting one clinician over another based on a quality measure not accounting for these independent factors would not result in improved outcomes for the population. Risk adjustment attempts to solve that problem and increase the likelihood of selecting a clinician or facility based on performance results in improved outcomes for the population.

Considering confounding factors becomes even more important with use of performance scores as a basis for calculating the amount of incentives or penalties for value-based purchasing and many Alternative Payment Models (APMs).

2. Ways to Account for Risk

2.1 When to use Risk Adjustment and/or Stratification

Risk stratification refers to reporting outcomes separately for different groups, unadjusted by the risk factor associated with the grouping. For example, if we use a variable representing the social risk factor of federal poverty level (FPL), and this variable has four levels (e.g., under 100% of the FPL, 100% to under 200% of the FPL, 200% to under 300% of the FPL, and 300% of FPL or above), then we would conduct risk stratification by running the statistical model of measured entity performance without the FPL variable on each of the four levels of FPL. There are two circumstances when risk adjustment in combination with stratification might be the most appropriate risk adjustment strategy (i.e., when the goal of the strategy is fair comparisons). The first circumstance is when patient factors are not independent of the quality construct. The second circumstance is when there is treatment heterogeneity, which is another case when the patient factors and quality construct are not independent, but for legitimate clinical reasons. As part of a risk adjustment strategy, a consensus-based entity (CBE) suggests using risk models in conjunction with risk stratification when use of a risk model alone would result in obscuring important health care disparities. Recent guidance, funded by CMS, suggests the minimum standard is stratification of risk-adjusted measures by social and functional risk factors to improve the ability to measure disparities and differential outcomes even if not adjusting for social or functional characteristics (NQF, 2022). If stratification is not possible and the patient factors are non-observable, reliability adjustment can eliminate the inherent bias introduced by low case volume (Dimik, et al., 2012). Risk adjustment is appropriate when the patient factors are correlated with the outcome and not correlated with the quality construct. Table 1 provides a high-level framework for risk adjustment strategies.

Table 1. Framework for Risk Adjustment Strategies

Relationship of Patient Factors and Quality Construct	Measurement of Patient Factors	
	Observable	Non-observable
Correlated (not independent)	Patient group stratification	Peer group stratification
Uncorrelated (independent)	Risk adjustment	Reliability adjustment

The Assistant Secretary for Planning and Evaluation (ASPE) recommendation for when to adjust for social risk is based on the type of measure and the program (ASPE, 2020, p. 34). See Table 2. No indicates a recommendation not to adjust for social risk factors and yes indicates a recommendation to adjust for social risk factors.

Table 2. ASPE Recommendations for Social Risk Factor Adjustment

Measure Type	Whether to Adjust for Social Risk Factors	
	Quality Reporting Programs	Value-Based Purchasing Programs
Process Measures	No	No
Outcome Measures	No	No
Patient Experience Measures	Yes	Yes
Resource Use Measures	No	Yes
Program Performance Scores	No	No

Measure developers should always consider what method would be most appropriate for accounting for social risk factors (e.g., risk adjustment, stratification by groups within a quality measure, stratification at the quality measure level). Stratification at the quality measure level may be similar to peer group stratification, in which patient factors are unobserved and correlated with observed measured entity characteristics. The recommendation for measure developers developing quality measures using the Blueprint content guidance is to explore use of a risk adjustment strategy, i.e., use of a statistical risk adjustment model, and, if necessary, risk stratification for selected populations. For CMS to accept an outcome measure and the CMS CBE to endorse it, the measure developer must demonstrate appropriate use of a risk adjustment and/or stratification strategy. Measure developers should provide rationale and strong evidence if an outcome measure is not risk adjusted or risk stratified.

It is the measure developer's responsibility to determine whether to account for variation in factors intrinsic to the patient before comparing outcomes and to determine how to best apply these factors in the quality measure specifications. Vogel and Chen (2018) noted "failure to address risk adjustment in an adequate manner can lead to biased conclusions that may adversely impact decision-making in both research and policy contexts" (p.1).

Clinical Quality Language (CQL) has the capability to model risk adjustment and stratification. Historically, for risk-adjusted eCQMs, measure developers have been

- describing risk adjustment methodology in the metadata
- using quality measure data post hoc to risk adjust
- representing variables for risk adjustment as required supplemental data
- including the logic or algorithm in the risk adjustment section of the Health Quality Measure Format (HQMF)

In the 2023 CMS Inpatient Prospective Payment System final rule, CMS adopted a risk-adjusted and stratified eCQM beginning with calendar year 2023 reporting. The risk adjustment methodology report for the Severe Obstetric Complications eCQM, is available on the Electronic Clinical Quality Improvement (eCQI) Resource Center (Yale New Haven Health Services Corporation -Center for Outcomes Research and Evaluation, 2021). CMS anticipates adopting more risk-adjusted eCQMs.

Measure developers may model stratification in CQL. Per the Measure Authoring Tool User Guide, the measure developer may include only one stratification in a single measure package grouping. However, the measure developer may add multiple strata to a single stratification.

2.2 Features of Risk Adjustment Models

The measure developer must evaluate the need for a risk adjustment strategy (i.e., risk adjustment, stratification, or both) for all potential outcome measures and statistically assess the adequacy of any strategies used. In general, a risk adjustment model possesses features such as those listed in Table 3, which was partially derived from a description of preferred features of models used for publicly reported outcomes (Krumholz et al., 2006). While some of the descriptions of the features in Table 3 are targeting risk adjustment models, the features are not exclusive to risk adjusted measures or risk adjustment models. Table 3 provides descriptions of each feature with more detail provided in subsections 2.2.1-2.2.6.

Table 3. Features of Risk Adjustment Models

Feature	Description
Sample definition	Clearly define the sample(s), clinically appropriate for the quality measure's risk adjustment, and large enough for sufficient statistical power and precision.
Appropriate time intervals	Clearly define the time intervals for model variables, ensure they are sufficiently long to observe an outcome and recent enough to retain clinical credibility.
High data quality	Data should be reliable, valid, complete, comprehensive, and rely on as few proxy measures as possible.
Appropriate variable selection	Selected adjustment or stratification variables should be clinically meaningful.
Appropriate analytic approach	The analytic approach must be scientifically rigorous and defensible, and consider multilevel or clustered organization of data (if necessary).
Complete documentation	Fully document risk adjustment and/or stratification details and the model's performance and disclose all known issues.

2.2.1 Define the Appropriate Measure Development Sample

Distributions of characteristics and their interactions within a measure development sample (i.e., the population used to develop the model) should be representative of the overall population on which the measure developer is applying the risk model. Measure developers should clearly and explicitly define the sample(s) as well as all inclusion, exclusion, and exception criteria they used to select the sample. Risk adjustment models generalize well (i.e., fit the parent population) to the extent the samples used to develop, calibrate, and validate them appropriately represent the parent population. Measure developers need to explain their rationale for using selected samples and offer justification of the sample's appropriateness.

2.2.2 Appropriate Time Intervals

The time interval is the time frame the measure developer uses to determine cases for inclusion in the population of interest and outcome of interest and includes an index event and a period of time. The measure developer should clearly state the criteria used to formulate decisions regarding the selection of the time interval and explain these criteria in the quality measure documentation. The time interval criteria used to identify risk factors for the stated outcomes should be clinically appropriate and clearly stated (e.g., the risk factor occurs within 24 hours of admission). Risk factors should be present at the start of care to avoid mistakenly adjusting for factors arising due to deficiencies in measured care unless there is use of person-time adjustments. Outcomes should occur soon enough after care to establish

they are the result of that care. For example, measure developers may use renal failure as one of the comorbidities for risk adjustment of a hospital mortality measure. If poor care received at the hospital caused the patient to develop renal failure after admission, it would be inappropriate to adjust for renal failure for that patient.

If not using person-time adjustments, the evaluation of outcomes must also be based on a standardized period of assessment. If there is no standardization of the periods of the outcome assessments, such as the assessment of events during hospitalization, there may be bias in the evaluation because measured entities have different practice patterns (e.g., varying lengths of stay).

2.2.3 High Data Quality

The measure developer must ensure the data they use for risk adjustment are of the highest quality possible. Considerations in determining the quality of data include

- There is reliable data collection. The method of collection must be reproducible with minimal variation between one collection and another if the same population is the source.
- Data are sufficiently valid for their purpose. Validation ultimately rests on the strength of the logical connection between the construct of interest and the results of operationalizing their measurement, recording, storage, and retrieval.
- Data are sufficiently comprehensive to limit the number of proxy measures required for the model. Obtaining the actual information is sometimes impossible, so using proxy measures might be necessary for certain projects.
- Data are as recent as possible. If the measure developer used 1990 data in a model designed for use tomorrow, many people would argue the health care system has changed so much since 1990, the model may not be relevant.
- Data are as complete as possible. Data should contain as few missing values as possible. Missing values are difficult to interpret and lower the validity ⓘ of the model.
- Documentation and full disclosure of data sources, including the dates of data collection, manner of data cleaning, data manipulation techniques (if applicable), and the data's assumed quality.

2.2.4 Appropriate Variable Selection

The risk adjustment model variables should be clinically meaningful or related to clinically meaningful variables. When developing a risk-adjusted model, the clinical relevance of included variables should be apparent to subject matter experts (SMEs). When variables are clearly clinically relevant, they serve two purposes: the clinical relevance contributes to the face validity of the model and the likelihood the model explains variation identified by health care professionals, and/or the literature as important to the outcome. Parsimonious models and their outcome are likely to have the highest face validity and be optimal for use in a model. The measure developer must determine which risk factors to retain in the risk adjustment model, for example retained variables are clinically relevant and statistically significantly associated with the outcome(s).

Occasionally, the measure developer may consider including proxy variables in the risk adjustment model based on prior research. This situation may arise when direct assessment of a relevant variable is not possible, and there is a requirement for use of a substitute or proxy variable. However, the relevance of these substitute variables should be empirically appropriate for the clinical topic of interest. For example, medications taken might be useful as a proxy for illness severity or progression of a chronic illness, provided practice guidelines or prior studies clearly link the medication patterns to the illness severity or trajectory. Similarly, the measure developer should consider inclusion of variables previously

shown to moderate the relationship between a risk adjustor and the quality measure. Moderating variables are interaction terms sometimes included in a model to understand complex information structures among variables (e.g., a prior mental health diagnosis may be only weakly associated with a measured outcome, but it may interact with another variable to strongly predict the outcome). Moderating variables and interaction terms, when needed, require specialized data coding and interpretation.

2.2.5 Appropriate Analytic Approach

An appropriate statistical model is determined by many factors. Measure developers may use logistic regression or hierarchical logistic regression when the outcome is dichotomous. However, in certain instances, they may use the same data to develop a linear regression model provided that doing so does not violate key statistical assumptions.[1] Selecting the correct statistical model is imperative because an incorrect model can lead to erroneous or misleading results. The analytic approach should also consider any multilevel and/or clustered organization of data, which is typically present when assessing institutions such as hospitals from widespread geographic areas.

Risk factors retained in the model should account for substantive and significant variation in the outcome. Overall differences between adjusted and unadjusted outcomes should also be pragmatically and clinically meaningful. Moreover, risk factors should not be related to stratification factors. A statistician can guide the measure development team and recommend the most useful variable formats and appropriate models.

2.2.6 Complete Documentation

Transparency is one of the key design principles in the Blueprint content on the *CMS MMS Hub*. When measure developers do not disclose the steps they used to create a risk adjustment model, others cannot understand or fully evaluate the model. The *Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement* on the endorsement of proprietary measures promotes the full disclosure of all aspects of a risk adjustment model used in measure development.

For the sake of transparency, measure developers should fully describe the risk adjustment method used; performance of the risk adjustment model, its components, and its algorithms; sources of the data and methods used to clean or manipulate the data; and the code (e.g., SAS) and documentation for how to run the calculation code. Documentation should be sufficient to enable others to reproduce the findings. The expectation is that the quality measure documentation will incorporate statistical and methodological recommendations from a knowledgeable statistician to explain the model and justify its selection.

To promote a clear understanding of the model and rationale for decisions made, the risk adjustment methodology documentation should

- Confirm the timeframes used in the model are
  - an important attribute of the model
  - clearly stated and appropriate
  - appropriate for implementation in the selected program
- Discuss the variables included in the model. If using race, sex, or social factors as risk adjustment variables, ensure they do not obscure health care disparities. CMS has a continued interest in

[1] There is no intention to suggest logistic regression is appropriate to model continuous manifest variables (i.e., available data). Nonetheless, measure developers use various forms of logistic regression to model latent traits (i.e., inferred variables modeled through related observations) assumed to be continuous, except where the available data are dichotomous, such as the probability of receiving a specified health care service.

identifying and mitigating disparities in clinical care areas/outcomes across patient demographics.

- Justify the analytic approach/types of models used.
  - Confirm the appropriateness of techniques used to assess the model.
  - Review the predictive validity, discriminant validity, and overall fit of the model.

CMS posts methodology reports for most risk-adjusted measures on the CMS *Measure Methodology* page or the *eCQI Resource Center*.

2.3 Risk Adjustment/Stratification Procedure

Measure developers should follow several steps when developing a risk adjustment or risk stratification model:

- Choose and define an outcome
- Define the conceptual model
- Identify potential data sources and variables
- Identify risk factors and timing
- Model and empirically test the data
- Assess the model
- Document the model

Some models may not lend themselves appropriately to all of these steps, e.g., empirical testing. An experienced statistician and clinical expert together can help determine the need for each step.

2.3.1 Choose and Define an Outcome

When selecting outcomes appropriate for risk adjustment and/or stratification, the time interval for the outcome must be meaningful, the definition of the outcome must clearly define what to count and not to count, and one must be able to collect the outcome data reliably. An appropriate outcome has clinical or policy relevance. It should occur with sufficient frequency to enable statistical analysis unless the outcome is a preventable and serious health care error that should never happen. Measure developers should evaluate outcome measures for both validity and reliability, as described in the Blueprint *Measure Testing* content on the *CMS MMS Hub*. Whenever possible, measure developers should consult with clinical experts, such as those participating in the technical expert panel (TEP), to help define appropriate and meaningful outcomes. Finally, as discussed in the supplemental material, *Person and Family Engagement in Quality Measurement*, patients should be involved in choosing which outcomes are appropriate for quality measurement. They are the ultimate experts on what is meaningful to their experience and what they value.

2.3.2 Define the Conceptual Model

Measure developers should develop a priori, a clinical hypothesis or conceptual model about how potential risk factors relate to the outcome. There is no one best way to model risk adjustment. The conceptual model serves as a map for development of a risk adjustment or stratification model. It defines the understanding of the relationships among the variables and, as such, helps identify which risk factors, patients, and outcomes are important, and which to exclude. Because the cost of developing a risk adjustment and/or stratification model may be prohibitive if the measure developer included every potential risk factor, the conceptual model also enables the measure developer to prioritize among risk factors and evaluate the cost and benefit of data collection. An in-depth literature review can greatly enhance this process. Alternatively, the existence of large databases and computational approaches such as machine learning allow for statistical analyses to explore the data for

relationships between outcomes and potential, not yet clinically identified adjustment factors, but exist empirically. Measure developers should be aware of the potential for spurious relationships.

The first step in developing or selecting the conceptual model is identifying relationships among variables. This process should include

- Conducting a review of the literature and canvassing expert opinion to establish variable relationships and identify measurable patient factors related to the outcome and are either independent of the quality construct (i.e., for risk adjustment) or not independent (i.e., for stratification).
- Obtaining expert opinion from measured entities and subject matter experts with relevant specialties, experienced statisticians and research methodologists, and relevant interested parties ⓘ such as patient advocates. Measure developers should use a TEP if seeking diverse input. The supplemental material, *Technical Expert Panels*, covers the standardized process for convening a TEP.
- Conducting empirical analyses to further support variable selection (when appropriate data are available) to identify potential factors for consideration by SMEs.

2.3.3 Identify the Risk Factors and Timing

Use a conceptual model and expertise promoting selection of risk factors

- clinically relevant in the case of clinical risk factors
- reliably collected
- validly operationalized
- sufficiently comprehensive
- associated with the outcome
- clearly defined
- identified using appropriate time frames

In addition to these attributes, risk factors should also align with CMS CBE policies for endorsed quality measures if seeking CBE endorsement.

2.3.4 Acquire Data (Sample, if Necessary)

Health care data can come from many sources, but the three most frequently used are claims data, patient record data, and survey data. Of these, the most common source of data for developing risk adjustment models is claims data reported by the measured entity. However, identifying sources of social risk factors may be more difficult as measured entities historically have not collected or reported these data accurately or consistently. Once the measure developer identifies data sources and secures permission to use the data, they may need to link relevant databases and perform data preparation tasks, including assessment of data reliability and validity, if not previously confirmed. If using samples, the measure developer should draw the samples using predefined criteria and methodologically sound sampling techniques. The measure developer needs to use data from diverse measured entities and diverse patients. Testing to determine the suitability of data sources and testing for differences across data sources may also be necessary.

The alpha and beta testing discussion in the Measure Testing content on the *CMS MMS Hub*, provides details of testing processes.

2.3.5 Analyze the Data

In addition to clinical judgment used to define the conceptual model and select candidate variables, the measure developer should conduct empirical analysis to help determine which risk factors to include or exclude. While primarily used to assess addition of a biomarker to a risk model, measure developers may want to consider using net reclassification improvement (or index) techniques to compare risk models (e.g., Kerr et al., 2014; McKearnan et al., 2018). Measure developers should consider the four factors in Subsections 2.3.5.1-2.3.5.4 when developing an appropriate risk adjustment model.

2.3.5.1 Sufficient Data

When creating a risk adjustment model, there should be enough data available to ensure a stable model. Different statistical rules apply to different types of models. For example, a model with a more common outcome may require more than 30 cases per patient factor to consistently return the same model statistics across samples. If the outcome is uncommon, then the number of cases required could be much larger (Iezzoni, 2013). Other factors may also affect the needed sample size, such as a lack of variability among risk factors with small sample resulting in partial collinearity among risk factors and a corresponding decrease in the stability of the parameter estimates. A statistician can provide guidance to determine the appropriate sample sizes based on the characteristics of the sample(s) and the requirements of the types of analyses in use.

2.3.5.2 Methods to Retain or Remove Risk Adjustors

Whenever possible, it is preferable to fit a model with as few variables as possible to explain any variance as completely as possible. This is often known as model simplicity or model parsimony. The principle of parsimony captures the balance between errors of underfitting and overfitting inherent in risk adjustment model development. For example, developing a model with many predictors can result in variables that primarily explain incremental variance unique to a data source or available samples (i.e., overfitting) and can also result in reduced stability of parameters due to increased multicollinearity (collinearity among more than two variables). In contrast, a model with fewer predictors may reduce the amount of explained variance possible for the quality measure (i.e., underfitting).

When evaluating these models, determination of the preferred model may depend on the availability of other samples to validate findings and detect overfitting and the degree of multicollinearity among predictors. However, in general, the simpler model may provide a more robust explanation since it uses fewer variables to explain nearly the same observed variability. In addition, simpler models are likely to reduce the cost of model development because the simpler models involve collecting fewer variables and may be less likely to show signs of model overfitting. Measure developers may achieve parsimonious models by omitting statistically significant predictors offering minimal improvement in predictive validity or overall model fit and by combining clinically similar conditions to improve performance of the model across time and populations. However, in situations with high visibility or potentially widespread fiscal repercussions, CMS has employed some of the most sophisticated models available, such as Hierarchical Generalized Linear Models as described in *Statistical Issues in Assessing Hospital Performance*.

When developing a risk adjustment model, the choice of variables for inclusion often depends on estimated parameters in the sample rather than the true value of the parameter in the population. Consequently, when selecting variables to retain or exclude from a model, the idiosyncrasies of the sample, as well as factors such as the number of candidate variables and correlations among the candidate variables, may determine the final risk adjustors retained in a model. Improper model selection or not accounting for the number of, or correlation among, the candidate variables may lead

to risk adjustment models that include suboptimal parameters or overestimated parameters—making them too extreme or inappropriate for application to future datasets. This is model overfitting, as the model is more complicated than needed and describes random error instead of an underlying relationship.

The measure developer can use statistical model fitting methods such as stepwise regression, or an adaptation thereof, to identify the simplest combination of variables providing high predictive value without overfitting. Another step to consider to minimize model overfitting is selection of model variables based on resampling methods and assessment of the model in multiple/diverse samples (refer also to subsection 0, Generalizability). There is a strong recommendation to consult with clinical experts, ideally during candidate variable selection, when examining the performance of candidate variables in risk adjustment models. This expertise may help inform relationships among model parameters and may help justify decisions to retain or remove variables.

2.3.5.3 Generalizability

Measure developers should take steps to ensure generalizability of their findings to target/initial populations when developing their models. Measure developers should begin with parameter estimation and test with diverse measured entities and persons. Researchers often use two datasets in building risk adjustment models: a development (i.e., calibration) dataset and a validation dataset. Measure developers should use the development/calibration dataset to develop the model or calibrate coefficients, and use the validation dataset to determine the appropriate extent of the application of the model to parent populations. When assessing generalizability to the population of the development dataset, the measure developer may collect the two datasets independently—which can be costly—or the measure developer may split one dataset using random selection.

Either of these methods enables evaluation of the model's generalizability to the population and helps avoid any model features arising from idiosyncrasies in the development sample. Additional validation using samples from different time periods may also be desirable to examine stability of the model over time.

2.3.5.4 Multilevel (Hierarchical) Data

The potential for observations to be "nested" within larger random groupings or levels frequently occurs in health care measurement (e.g., there is nesting of patients under physician groups, which may in turn be nested under hospitals). The risk adjustment model should account for these multilevel relationships when present. In developing risk adjustment models, measure developers should investigate theoretical and empirical evidence for potential patterns of correlation in multi-level data. For example, the risk adjustment model should address patients in the same Inpatient Rehabilitation Facility (IRF) with tendencies to have similar outcomes based on a variety of factors.

The measure developer should examine these multilevel relationships by building models designed to account for relationships between observations within larger groups. Terms for these types of models include multilevel model, hierarchical model, random-effects model, random coefficient model, and mixed model. These terms all refer to models explicitly modeling the "random" and "fixed" variables at each level of the data. In this terminology, the assumption is to measure a "fixed" variable without error, where the value or measured characteristic is the same across samples (e.g., male vs. female, nonprofit vs. for-profit facility) and studies. In contrast, the assumption is "random" variables are to be values drawn from a larger population of values (e.g., a sample of IRFs), where the value of the random variable represents a random sample of all possible values of that variable.

Traditional statistical methods (i.e., linear regression and logistic regression) require observations (e.g., patients) in the same grouping to be independent. When observations co-vary based on the organization of larger groupings, these methods fail to account for the hierarchical structure and there are violations of the assumptions of independence among the observations. This situation may ultimately lead to underestimated standard errors and incorrect inferences. Attempts to compensate for this problem by treating the grouping units as fixed variables within a traditional regression framework is undesirable because it reduces the generalizability of the findings.

Multilevel models overcome these issues by explicitly modeling the grouping structure and by assuming the groups reflect random variables (usually with a normal distribution) sampled from a larger population. They consider variation at different grouping levels and allow modeling of hypothesized factors at these different levels. For example, a multilevel model may allow modeling patient-level risk factors along with the facility-level factors. If the measure developer has reason to suspect hierarchical structure in the measurement data, they should examine these models. The measure developer should apply the models within common frameworks used for risk adjustment (e.g., ordinary least squares regression for continuous outcomes, logistic regression for binary outcomes), as well as less common longitudinal frameworks such as growth (i.e., change) modeling.

Developments in statistics are enabling researchers to improve both the accuracy and the precision of nested models using computer-intensive programs. These models include estimation of clustering effects independent of the main effects of the model to better evaluate the outcome of interest. For example, there are indications use of precision-weighted empirical Bayesian estimation produces more accurately generalizable coefficients across populations than methods relying on the normal curve for estimation (e.g., linear regression). There is also use of hierarchical factor analysis and structural equation modeling.

2.3.6 Assess the Model

There is a requirement to assess the model for a newly developed risk adjustment model and when using an "off-the-shelf" adjustment model because an existing risk adjustment model may perform differently in the new quality measure context. When multiple data sources are available (e.g., claims and chart-based data), the recommendation is to assess the model performance for each data source to allow judgment regarding the adequacy and comparability of the model across the data sources.

Measure developers should assess all models they develop to ensure the models do not violate underlying assumptions (e.g., independence of observations or assumptions about underlying distributions) beyond the robustness established in the literature for those assumptions. Measure developers should assess models to determine predictive ability, discriminant ability, and overall fit of the model.

Some examples of common statistics used in assessing risk adjustment models include the R^2 statistic, receiver-operating characteristic (ROC) curve, the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), and Hosmer-Lemeshow test (HL test). However, several other statistical techniques can enable measure developers to assess different aspects of model fit for different subpopulations as well as for the overall population. The engagement of an experienced statistician is critical to ensure the selection of the most appropriate methods during model development and testing.

2.3.6.1 R2 Statistic

A comparison of the R^2 statistic with and without selected risk adjustment is frequently used to assess the degree to which specific risk-adjusted models predict, explain, or reduce variation in outcomes unrelated to an outcome of interest. Measure developers may use the statistic to assess the predictive

power of risk-adjusted models overall. In that case, values for R^2 describe how well the model predicts the outcome based on values of the included risk factors.

The R^2 value for a model can vary, and no firm standard exists for the optimal expected value. Experience or previously developed models may inform which R^2 value is most reasonable. In general, the larger the R^2 value, the better the model. However, the measure developer may also need clinical expertise to help assess whether remaining variation is primarily related to differences in the measured quality. Extremely high R^2 values can indicate something is wrong with the model.

2.3.6.2 ROC Curve, Area Under the Curve (AUC), and C-statistic

Measure developers often use the ROC curve to assess models predicting a binary outcome (e.g., a logistic regression model), where there are two categories of responses. Measure developers may plot the ROC curve as the proportion of target outcomes correctly predicted (i.e., a true positive) against the proportion of outcomes incorrectly predicted (i.e., a false positive). The curve depicts the tradeoff between the model's sensitivity ⓘ and specificity ⓘ.

Figure 1 shows an example of ROC curves. Curves approaching the 45-degree diagonal of the graph represent less desirable models (Curve A) when compared to curves falling to the left of this diagonal indicating higher overall accuracy of the model (Curves B and C). A test with nearly perfect discrimination will show a ROC curve passing through the upper-left corner of the graph, where sensitivity equals 1, and 1 minus specificity equals zero (Curve D).

The ROC AUC often quantifies the power of a model to correctly classify outcomes into two categories (i.e., discriminate). The AUC, sometimes referred to as the c-statistic ⓘ, is a value varying from 0.5 (i.e., discriminating power not better than chance) to 1.0 (i.e., perfect discriminating power). The interpretation of the AUC can represent the percent of all possible pairs of observed outcomes in which the model assigns a higher probability to a correctly classified observation than to an incorrect observation. Most statistical software packages compute the probability of observing the model AUC found in the sample when the population AUC equals 0.5 (i.e., the null hypothesis). Both non-parametric ⓘ and parametric methods ⓘ exist for calculating the AUC, and this varies by statistical software.

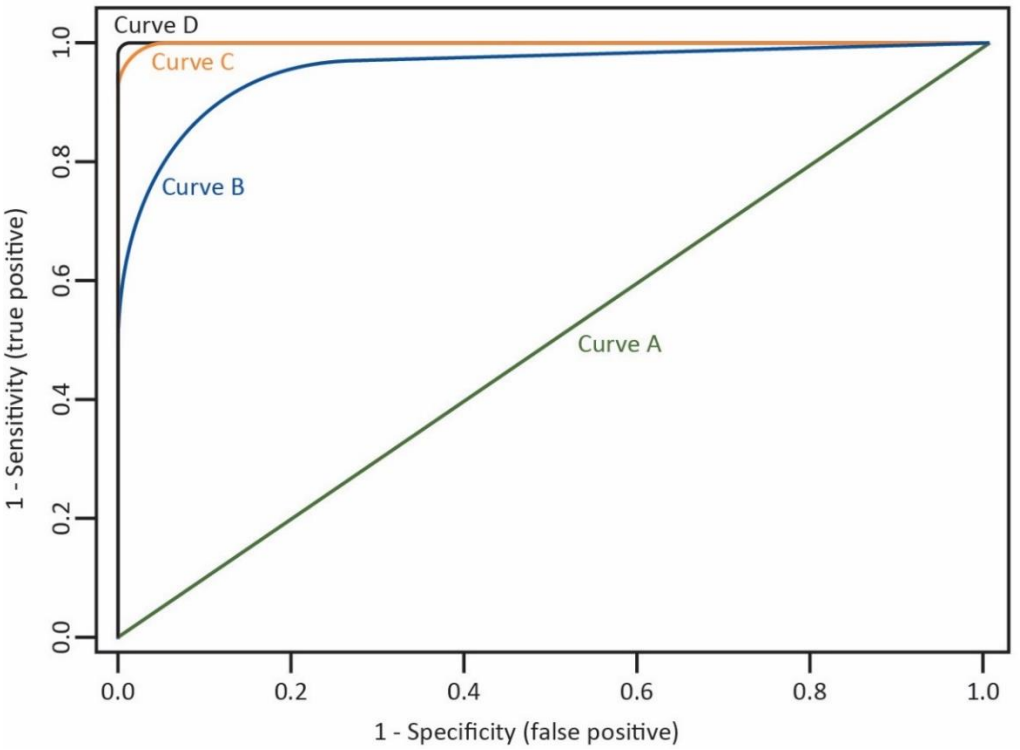

Figure 1. Example of ROC Curves

2.3.6.3 The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)

The measure developer can use AIC and BIC to compare different logistic or linear regression models created from the same set of observations. Calculation of both the AIC and BIC is from the likelihood, a measure of the fit of the model. With the introduction of additional variables, the likelihood of a model can only increase, but both the AIC and BIC contain a penalty for the number of variables in the model. This means an additional variable must explain enough extra variation to "make up for" the additional complexity in the model. The measure developer calculates AIC and BIC as (where k is the number of variables in the model and n is the number of observations in the model):

AIC = -2*log (maximum likelihood) + 2*k

BIC = -2*log(likelihood) + k*ln(n)

The magnitude of the BIC penalty depends on the total number of observations in the dataset (the penalty is higher for a larger dataset), but the penalty for the AIC is the same regardless of the dataset size. A higher likelihood represents a better fit, because the measure developer calculates the AIC and BIC from -2*log(likelihood), lower values of AIC and BIC represent better models (Akaike, 1974; Schwarz, 1978).

2.3.6.4 The Hosmer-Lemeshow (HL) Test

Although the AUC/c-statistic values provide a method to assess a model's discrimination, the measure developer can assess the quality of a model by how closely the predicted probabilities of the model agree with the actual outcome (i.e., whether predicted probabilities are too high or too low relative to true population values). This is known as calibration of a model. The HL test is a commonly used procedure for assessing the calibration of a model and the goodness of fit in logistic regression in the

evaluation of risk adjustment models. This test assesses the extent to which the observed values/occurrences match expected event rates in subgroups of the model population. However, as with any statistical test, the power increases with sample size. This can be undesirable for goodness of fit tests because in very large data sets, there is significance in small departures from the proposed model. Meaning, statistically, the measure developer may deem the goodness of fit poor, even though the clinical significance is small to none. An alternative to consider is the dependence of power on the number of groups (e.g., deciles) used in the HL test. Using this approach, the measure developer can standardize the power across different sample sizes in a wide range of models. This allows an "apples-to-apples" comparison of goodness of fit between quality measures with large and small denominators (Paul, et al., 2012).

Generally, in a well calibrated model, the expected and observed values agree for any reasonable grouping of the observations. Yet, high-risk and low-frequency situations pose special problems for these types of comparison methodologies; therefore, an experienced statistician should address such cases.

The expectation is the measure development team will include a statistician to accurately assess the appropriateness of a risk-adjusted model. Determining the best risk-adjusted model may involve multiple statistical tests more complex than those cited here. For example, a risk adjustment model may discriminate very well based on the c-statistic, but poorly calibrated. Such a model may predict well at low ranges of outcome risk for patients with a certain set of characteristics (e.g., the model produces an outcome risk of 0.2 when roughly 20% of the patients with these characteristics exhibit the outcome in population), but predict poorly at higher ranges of risk (e.g., the model produces an outcome risk of 0.9 for patients with a different pattern of characteristics when only 55% of patients with these characteristics show the outcome in population). In this case, the measure developer should consult one or more goodness-of-fit indices to identify a superior model. Careful analysis of different subgroups in the sample may help to further refine the model. This may require additional steps to correct for bias in estimators, improve confidence intervals, and assess any violation of model assumptions. Moreover, differences across groups for non-risk-adjusted measures may be clinically inconsequential when compared to risk-adjusted outcomes. It may be useful to consult clinical experts in the subject matter to provide an assessment of both the risk adjustors and utility of the outcomes.

2.3.7 Document the Model

The documentation ensures relevant information about the development and limitations of the risk adjustment model are available for review by consumers, purchasers, and measured entities. The documentation also enables these parties to access information about the factors incorporated into the model, the method of model development, and the significance of the factors used in the model. Typically, the documentation contains

- Identification or review of the need for risk adjustment of the quality measure(s).
- A description of the sample(s) used to develop the model, including criteria used to select the sample and/or number of sites/groups, if applicable.
- A description of the methodologies and steps used in the development of the model or a description of the selection of an off-the-shelf model.
- A listing of all variables considered and retained for the model, the contribution of each retained variable to the model's explanatory power, and a description of how the measure developer collected each variable (e.g., data source, time frames for collection).
- A description of the model's performance, including any statistical techniques used to evaluate performance and a summary of model discrimination and calibration in one or more samples.

- Delineation of important limitations such as the probable frequency and influence of misclassification when the model is used (e.g., classifying a high-outcome measured entity as a low one or the reverse) (Austin, 2008).
- Enough summary information about the comparison between unadjusted and adjusted outcomes to evaluate whether the model's influence is clinically significant.
- A section discussing a recalibration schedule for the model to accommodate changes in medicine and in populations; the first assignment of such a schedule is normally based on the experience of clinicians and the literature's results and then later updated as needed.

The measure developer should fully disclose all quality measure specifications, including the risk adjustment methodology. The Measure Information Form and Instructions, Measure Justification Form and Instructions, and Measure Evaluation Report and Instructions, all found in Templates, provide guidance for documenting the risk adjustment information.

 If not modeled in the CQL, the measure developer should provide the risk adjustment instructions in the HQMF and human-readable HyperText Markup Language identifying where the user may obtain the complete risk adjustment methodology.

Documentation should comply with the open-source requirements in the Conditions for Consideration in *Solicitation of Measures and Measure Concepts* (2018), and include all applicable programming code. If calculation requires frequently changing, database-dependent coefficients, the measure developer should disclose the existence of these coefficients and the general frequency with which they change, but they do not need to disclose the precise numerical values assigned, as they vary over time.

3. Key Points

Risk adjustment is a method measure developers can use to account for confounding factors when calculating performance scores. Risk adjustment is especially important in the context of CMS programs using performance scores as a basis for calculating the amount of incentives or penalties for value-based purchasing and many APMs. As such, measure developers should evaluate the need for risk adjustment, stratification, or both, for all potential outcome measures and statistically assess the adequacy of any strategies they use. Measure developers should also consider the appropriateness of adjusting for social risk (i.e., socioeconomic factors) for these quality measures. Measure developers must determine on a case-by-case basis whether they should adjust a quality measure for social risk.

When developing a risk adjustment model, measure developers must identify and clearly define a sample to test the model. That sample should consist of high quality (i.e., valid, reliable, and comprehensive) data, with appropriate time intervals for all model variables. Further, measure developers should provide evidence all variables included in the model are clinically meaningful. Measure developers should also consult with an experienced statistician to outline a sound analytic, scientifically rigorous and defensible approach —including information about how to assess the model (e.g., the R^2 statistic, ROC curve, and HL test). Finally, measure developers should document information about the development and limitations of the risk adjustment model to ensure interested parties can appropriately review and interpret quality measure scores.

References

21st Century Cures Act, Pub. L. No. 114-255, 130 Stat. 1033 (2016). https://www.govinfo.gov/content/pkg/PLAW-114publ255/pdf/PLAW-114publ255.pdf

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Ash, A.S., Fienberg, S.E., Louis, T.A., Normand, S.T., Stukel, T.A., & Utts, J. (2011). *Statistical issues in assessing hospital performance*. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf

Assistant Secretary for Planning and Evaluation, Office of Health Policy. (2020, March). *Report to Congress. Social risk factors and performance in Medicare's value-based purchasing program*. https://aspe.hhs.gov/sites/default/files/migrated_legacy_files//195191/Second-IMPACT-SES-Report-to-Congress.pdf

Austin, P.C. (2008). Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals. *BMC Medical Research Methodology, 8*, 30. https://doi.org/10.1186/1471-2288-8-30

Centers for Medicare & Medicaid Services. (n.d.-a). *Hospital readmissions reduction program (HRRP).* Retrieved June 21, 2023, from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program.html Centers for Medicare & Medicaid Services. (n.d.-b). *Measure methodology*. Retrieved June 21, 2023, from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology

Centers for Medicare & Medicaid Services. (2022, November 4). *Measure authoring tool 6.10 QDM and FHIR user guide*. Retrieved June 21, 2023, from https://www.emeasuretool.cms.gov/sites/default/files/2021-09/MAT%20User%20Guide%20v6.10%20FHIR%2009%3A29%20.pdf

Dimick, J. B., Ghaferi, A. A., Osborne, N. H., Ko, C. Y., & Hall, B. L. (2012). Reliability adjustment for reporting hospital outcomes with surgery. *Annals of Surgery, 255*(4), 703-707. https://doi.org/10.1097/SLA.0b013e31824b46ff

Iezzoni, L. (Ed.). (2013). *Risk adjustment for measuring health care outcomes* (4th edition). Foundation of the American College of Health Care Executives.

Kerr, K. F., Wang, Z., Janes, H., McClelland, R. L., Psaty, B. M., & Pepe, M. S. (2014). Net reclassification indices for evaluating risk-prediction instruments: A critical review. *Epidemiology, 25*(1), 114-121. https://doi.org/10.1097/EDE.0000000000000018

Krumholz, H.M., Brindis, R.G., Brush, J.E., Cohen, D.J., Epstein, A.J., Furie, K., Howard, G., Peterson, E. D., Rathore, S. S., Smith Jr, S. C., Spertus, J. A., Wang, Y., & Normand, S.T. (2006). Standards for statistical models used for public reporting of health outcomes. *Circulation, 113*(3), 456-462. https://doi.org/10.1161/CIRCULATIONAHA.105.170769

McKearnan, S. B., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., O'Connor, P. J. (2018). Performance of the net reclassification improvement for nonnested models and a novel percentile-based alternative. *American Journal of Epidemiology, 187*(6), 1327-1335. https://doi.org/10.1093/aje/kwx374

National Quality Forum. (n.d.). *Risk adjustment guidance*. Retrieved June 26, 2023, from http://www.qualityforum.org/Risk_Adjustment_Guidance.aspx

National Quality Forum. (2014). *Risk adjustment for socioeconomic status or other sociodemographic factors*. http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=77474

National Quality Forum. (2018). *Solicitation of measures and measure concepts*. https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=86124

National Quality Forum. (2021). *Measure evaluation criteria and guidance for evaluating measures for endorsement*. http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=88439

National Quality Forum. (2022, December). *Developing and testing risk adjustment models for social and functional status-related risk within healthcare performance measurement*. https://www.qualityforum.org/Projects/n-r/Risk_Adjustment_Guidance/Technical_Guidance_Final_Report_-_Phase_2.aspx

Office of the Assistant Secretary for Planning and Evaluation. (2016, December). *Report to Congress: Social risk factors and performance under Medicare's value-based purchasing programs*. https://aspe.hhs.gov/system/files/pdf/253971/ASPESESRTCfull.pdf

Paul, P., Pennell, M.L., Lemeshow, S. (2013). Standardizing the power of the Hosmer‐Lemeshow goodness of fit test in large data sets. *Statistics in Medicine, 32*, 67-80. https://doi.org/10.1002/sim.5525

Statistics How To. (2019). *Hosmer-Lemeshow test.* Retrieved June 21, 2023, from https://www.statisticshowto.com/hosmer-lemeshow-test/

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), 461–464. https://doi.org/10.1214/aos/1176344136

Vogel, W. B., & Chen, G. J. (2018). An introduction to the why and how of risk adjustment. *Biostatistics & Epidemiology, 4,* 84-97. https://doi.org/10.1080/24709360.2018.1519990

Wadhera, R. K., Yeh, R. W., & Joynt Maddox, K. E. (2019). The Hospital Readmissions Reduction Program – Time for a reboot. *New England Journal of Medicine, 380*, 2289-2291. https://doi.org/10.1056/NEJMp1901225

Yale New Haven Health Services Corporation -Center for Outcomes Research and Evaluation. (2021, October). *Severe obstetric complications electronic clinical quality measure (eCQM) methodology report*. https://www.cms.gov/files/document/measure-methodology-report.pdf