Detailed Business Requirements

Supporting documentation on episode risk adjustment

State of Ohio

July 29, 2014

1. CONTEXT AND PURPOSE OF RISK ADJUSTMENT

Principal Accountable Providers (PAPs) participating in episode-based payment models are compared based on their performance on quality metrics and based on the average spend for episodes treated by each PAP. The credibility and effectiveness of an episode-based payment model therefore rests on the comparability and fairness of the episode spend measure used in the comparisons. Risk adjustment is one of several mechanisms that episode-based payment models may use to achieve comparability in episode spend across PAPs.

Risk adjustment specifically captures the impact on episode spend of documented clinical risk factors that typically require additional care during an episode and are outside the control of the PAP. The goal of risk adjustment is to account for different levels of medical risk across patient panels and, by doing so, reduce incentives for tactical selection of patients (i.e., avoiding riskier and more costly patients) when payments are tied to episode spend performance.

Importantly, risk adjustment is applied after other mechanisms that aim to create fair episode spend comparisons across PAPs. The other mechanisms include:

- Targeted inclusion of spend in the episode: Only spend for diagnoses, procedures, and medications related to the episode is included in the calculation of episode spend.
- Exclusion of episodes for clinical reasons: Episodes with substantively different clinical pathways (e.g., due to comorbidities, the age of patient, left against medical advice, death) are excluded from the payment model.
- Exclusion of episodes for business reasons: Episodes where payment or eligibility rules (e.g., third party liability, inconsistent enrollment, dual eligibility), provider characteristics (e.g., out of state, certain provider types), patient characteristics (e.g., long-term care residents), or missing or exceptional claims information (e.g., long hospitalizations, incomplete claims) indicate fundamental differences in the available information to calculate episode spend are excluded from the payment model.

- Exclusion of high outlier episodes: Episodes with very high episode spend which may indicate a unique or highly uncommon event are excluded from the payment model.

Five principles ensure that the risk adjustment process for episode-based payment models is effective, practical, acceptable, sound, and meaningful:

- **Equitability:** In order to be effective, risk adjustment must fairly reflect the medical risk presented by a given patient in a given type of episode (e.g., an asthma episode or a perinatal episode). For example, for episode-based payment models, the overall medical risk of a patient which correlates with their total cost of care is less relevant than the specific medical risk that correlates with spend for a given type of episode. Therefore, an episode-specific approach to risk adjustment is applied which reflects the effect of episode-specific risk factors of a given patient on episode spend for a given type of episode.
- **Reproducibility:** In order to be practical, the risk adjustment process must offer a consistent methodology that works for different types of episodes, different populations, and different payers. While the methodology is consistent, the resulting clinically meaningful and statistically significant risk factors and risk coefficients may differ depending on the context where an episode is implemented. For example, differences in payment practices may lead to different risk coefficients for the same risk factor in the insured populations of two different payers. Or, differences in the clinical characteristics of patients and in treatment patters may lead to a risk factor being clinically meaningful and statistically significant in the insured population of one payer, but not in the insured population of another payer.
- **Transparency:** In order to be acceptable, the risk adjustment process must be well documented and the documentation must be accessible to all stakeholders. For example, the descriptions provided in this document aim to create the needed transparency.
- **Statistical validity:** In order to be sound, the risk adjustment process must be based on validated statistical techniques applied in an appropriate and mathematically rigorous manner. For example, the process described in this document was developed in collaboration with experts in health economics, actuarial science, and statistics to ensure the required rigor. In addition, the process uses publicly available software packages that have been described

and reviewed in the academic literature as the basis to perform the statistical analyses of the risk adjustment process.

- **Clinical validity:** In order to be meaningful, all inputs and outputs of the risk adjustment process must be subject to clinical review. The clinical review ensures that the risk factors and coefficient identified have logical and causal relationships with episode spend. For example, clinicians may identify potential risk factors such as diagnoses or procedures that are likely to impact patient care. In addition, clinicians may identify conditions that are likely to be complications of patient care and therefore should not be considered as risk factors.

The risk adjustment process generates a set of clinically meaningful and statistically significant risk factors and risk coefficients. The risk coefficients are used to calculate a risk score for each episode given the risk factors that are present for the episode. The risk score represents the ratio of the expected episode spend when no risk factors are present to the expected episode spend given the set of risk factors present for the episode. Multiplying the observed episode spend by the risk score results in the risk-adjusted episode spend. Risk-adjusted episode spend represents how much spend would have been incurred during the episode had there been no risk factors present, all other things being equal. By minimizing the effect of clinically documented medical risk that is outside the control of the PAP on episode spend, risk-adjustment contributes – along with the other measures mentioned above – to the fairness of the episode spend comparisons that underlie episode-based payment models.

2. RISK ADJUSTMENT PROCESS

Risk factors are identified in an iterative process informed by medical best practice, expert opinion, and statistical testing. The use of clinical input and statistical testing throughout the risk adjustment process aims to generate risk factors that represent genuine reasons for differences in care pathways (and therefore episode spend) and are not artifacts of the data or indirect measures of other conditions.

The risk adjustment process involves three steps: identification of potential risk factors, model selection, and estimation of risk coefficients.

- **Identification of potential risk factors:** To identify a set of clinically validated potential risk factors that are likely to impact episode spend, a series of statistical and clinical inputs are considered:
  - First, in order to support the clinical discussion and to avoid missing any risk factors that are obscure but strongly impact episode spend, a basic data mining technique is applied to determine which potential risk factors are correlated with episode spend. The output of the data mining step is used for discussion purposes only, and potential risk factors identified using data mining are only included in subsequent steps if clinical review provides a clear clinical rationale for their impact on episode spend. Due to complex interaction effects and limited data compared to the number of potential risk factors being tested, the output of the data mining process is not guaranteed to identify all clinically meaningful risk factors. It is also likely to identify some risk factors that appear to matter in-sample, but have little predictive value out-of-sample. However, the model will capture most of the potential risk factors that have a large impact on episode spend. As such, the outputs from the data mining model are used as a safeguard to ensure that all potential risk factors that have a statistically significant effect on episode spend are considered by the clinical advisory group.

    The method used for data mining is as follows:

    - The primary source for potential risk factors is the publically available Clinical Classification Software (CCS) created by the United States Agency for Healthcare Research and Quality (AHRQ). Developed for clustering patient diagnoses and procedures into clinically meaningful categories, CCS is a frequently used, validated, and transparent method of identifying patient conditions using historical claims data (http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp). The CCS assigns all ICD-9 diagnosis codes to one of 624 categories that represent four hierarchical levels of granularity. The CCS categories form the base universe of potential risk factors.
    - To identify the CCS categories that are used as potential risk factors in the data mining step, each episode is flagged for whether the patient had a diagnosis code in a given CCS category in the year prior to the episode or during the episode. To ensure that the potential risk factors are measureable only potential risk factors that affect 50 or more episodes (or, for episodes with low volume, 25 or more) are

taken into consideration. To ensure that the potential risk factors are mutually exclusive and collectively exhaustive, each family of CCS categories (e.g., 1, 1.1, 1.2, 1.3, 1.1.1, 1.1.2, etc.) is rolled up from level 4 to 3 to 2 (not 1) until all CCS categories at a given level are present for 50 (or 25) or more episodes.

- Using the thus identified potential risk factors as independent variables and the natural log of normalized episode spend as the dependent variable, three backward ordinary least squares (OLS) based variable selection algorithms are run using ten-fold cross validation on all valid episodes from the most recent year of complete claims data. (Implemented using the statistical package "RMS" – http://cran.r-project.org/web/packages/rms/rms.pdf). The three algorithms each use a different parameter for variable selection in order to generate more and less conservative lists of potential risk factors. The parameters for variable selection are, from most to least conservative, the Akaike information criterion (AIC), a p-value threshold of .05 and a p-value threshold of .20. The CCS categories that are identified as statistically significant by these three methods are set aside for further clinical evaluation.

– Second, the list of potential risk factors generated by data mining is supplemented by the risk factors used in the Health Care Incentives Improvement Institute's (HCI3) Prometheus payment model and, where available, risk factors mentioned in the literature.

– Third, a clinical advisory group composed of physicians, medical directors, and other clinical experts determines the set of clinically validated potential risk factors that should undergo further statistical testing. To select the set of potential risk factors for further testing, the clinical advisory group initially proposes a set of conditions that are likely to impact episode spend based on clinical experience. The potential risk factors may be identified through diagnoses, procedures, APR-DRGs, demographic data, or through a combination thereof, depending on the clinical advisory group's input. In addition, the clinical advisory group reviews the list of potential risk factors generated by data mining and the risk factors used by Prometheus and others to supplement their initial list. The clinical advisory group then narrows down the initial list to only those potential risk factors that are supported by a strong clinical rationale for their impact on episode spend. The list

developed by the clinical advisory group includes the medical codes (e.g., CCS category, diagnoses, procedures) that identify each clinically validated potential risk factor and the time window during which the medical codes need to be present (e.g., prior to the episode, during the trigger window).

- **Model selection:** To select the final set of risk factors that are used to estimate the risk coefficients, an exhaustive search of all possible risk adjustment models using the clinically validated potential risk factors is conducted. The model selection process involves the following steps:
  - Any clinically validated potential risk factor that is not present in a sufficient number of episodes to calculate a credible estimate of its risk coefficient (typically 50 or 25 episodes, depending on the volume of data available) is treated as a comorbidity episode exclusion. Because there is reason to believe that all the clinically validated potential risk factors impact episode spend, but the impact cannot be accurately measured if few episodes are affected by a risk factor, any episode with such a risk factor cannot be equitably included in episode-based payments and is therefore excluded.
  - For each possible combination of the remaining clinically validated potential risk factors an OLS regression model is estimated using the risk factors as independent variables and the log of normalized episode spend as the dependent variable over all remaining valid episodes during the most recent year of complete claims data. (Implemented using the statistical package "LEAPS" – http://cran.r-project.org/web/packages/leaps/leaps.pdf). Every possible model is generated, from a single clinically validated potential risk factor on one extreme to all the clinically validated potential risk factors together on the other.
  - From all models with a given number of risk factors the model with the highest R-squared (i.e., the model with the highest percent of variance in episode spend explained by the risk factors) is selected. The resulting set of models is an ordered list of the most predictive models by number of included risk factors (i.e., best model with one risk factor, best model with two risk factors, etc., all the way to best model with all clinically validated potential risk factors). The Bayesian information criterion (BIC), Mallows's C_p, and adjusted R-squared, three model selection parameters most commonly used in statistical literature, are calculated for each of the most predictive models and the best model based on the

selection criteria for each parameter is identified. The selection criteria are: the model with the lowest BIC, the model with the highest number of risk factors where the total number of risk factors is less than one half Mallow's C_p, and the model with the highest adjusted R-squared. Of the three candidate models selected based on each selection criterion, the one with the median number of risk factors is selected as the optimal risk adjustment model.

- **Estimation of risk coefficients**: To estimate the risk coefficients, both a linear OLS and a log-linked generalized linear model (GLM) based on the Poisson distribution are estimated using the risk factors that were selected in the model selection step as the independent variables and normalized episode spend as the dependent variable, over all remaining valid episodes during the most recent year of complete claims data. (Implemented using the "lm" and "glm" functions in R – http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html [lm] and http://stat.ethz.ch/R-manual/R-patched/library/stats/html/glm.html [glm]).

  Before the risk coefficients are finalized, the following checks are performed:

  - The variance inflation factor (VIF) for the risk factors is calculated to ensure that there is no severe multi-colinearity in the data. (Implemented using the statistical package "HH" – http://cran.r-project.org/web/packages/HH/HH.pdf). If any VIF higher than three exists, it is examined and risk factors may be removed or transformed (e.g., risk factors may be rolled up to a higher level of CCS categories, broken out to a lower level of CCS categories, or combined into one risk factor) based on clinical input to reduce multi-colinearity.
  - The average predicted normalized episode spend for each decile of predicted spend, from highest to lowest, is compared to the average observed normalized episode spend to test for systematic over/under prediction of high or low spend episodes. Additionally, the average predicted normalized episode spend for episodes with a given number of risk factors is compared to the average observed normalized episode spend to test for systematic over/under prediction of high or low risk episodes. The regression model (OLS or GLM) that best fits the data without systematic bias is used to generate risk coefficients and the other model discarded. If such over/under prediction exists in both models risk factors may be removed or transformed based on clinical input to

address the systematic over/under prediction. Alternatively, a different regression model may need to be chosen that better conforms to the observed distribution of episode spend.

- The frequency and goodness of the prediction of the "base case" with no risk factors is tested. If too few episodes (e.g., <10%) have no risk factors present or if episodes with no risk factors are systematically over/under predicted, risk factors may be removed or transformed based on clinical input to ensure that the "base case" of no risk factors is well-represented and well-predicted.
- An exception to removing or transforming risk factors due to under/over prediction exists for episodes of highly comorbid patients. Due to the large number of risk factors for episodes of highly comorbid patients, the episode spend is difficult to predict accurately using a regression model. Typically the presence of many risk factors generates episode spend predictions that are either too high (when using a GLM) or too low (when using OLS) due to the compounding effect of each risk factor. If only highly comorbid episodes show systematic over/under prediction then the risk factors are not removed or transformed. Instead episodes with more than the specific number of risk factors are excluded. Highly comorbid episodes are defined as episodes that have more than a specific number of risk factors present, have a systematic bias in the percent difference of predicted normalized episode spend and observed normalized episode spend of more than plus or minus 7.5%, and make up less than 2% of valid episodes. If the exclusion of highly comorbid episodes leads to another risk factor no longer being associated with a sufficient number of valid episodes to generate a credible estimate of its impact then that risk factor becomes an episode exclusion.
- Finally, the risk factors and their risk coefficients are presented to the clinical advisory group for feedback and a final clinical check. If any risk factor appears to have disproportionate or illogical impact then a deeper analysis of co-linearity is performed, the risk factors may be removed from the model, or transformed.

If any risk factors are removed or transformed at any point in the above process then the model selection step is repeated with any removed risk factors removed and any transformed risk factors updated. Any highly comorbid episodes are not be excluded when repeating the model selection

step, because the risk factors that make the episodes highly comorbid may no longer be included in the final model.

When the model selection step generates a set of risk factors that passes the checks described above without the removal or transformation of any risk factors then those risk factors are the final risk factors for the episode and the risk coefficients generated by the selected regression model are the final risk coefficients associated with them.

3. STATISTICAL AND TECHNICAL CONSIDERATIONS

The risk adjustment process described above takes the following statistical and technical considerations into account:

- With a limited number of episodes available to generate risk coefficients, the risk adjustment process should balance the desire to include all of a potentially large set of risk factors that may affect episode spend with the need for each risk factor to be sufficiently independent of other risk factors and have sufficient volume to reliably estimate its impact on episode spend. For example, a payer may identify only 1,000 episodes of a given type in their patient population. While the episodes may be affected by many risk factors, only a subset of the potential risk factors occurs frequently enough to allow for a reliable measurement of their effect on episode spend. Therefore, a process was chosen that first defines a large set of potential risk factors that may affect an episode and then narrows down the large set of potential risk factors to a set of clinically validated, statistically reliable, and practically measurable risk factors.
- The process of risk factor selection should balance the risk of including unimportant risk factors with the risk of not including important risk factors. Repeated clinical input on and evaluation of the risk factors included in the risk adjustment model is used to ensure that every risk factor has a strong clinical justification for its inclusion. Requiring any risk factor in the model to have a strong prior probability of having meaningful impact on episode spend reduces the likelihood of including a risk factor that appears statistically significant in the available data sample due to chance but has no true impact on episode spend. To minimize the risk of not including important risk factors, the clinical evaluation has access to an exhaustive statistical analysis of the impact of a broad set of potential risk factors on episode spend. This approach decreases the likelihood that a

statistically important risk factor is not considered or overlooked by the clinical advisory group, because all potential risk factors that are correlated with episode spend are brought to the group's attention.

- The risk adjustment process should capture the impact of risk factors apart from any impact of unit price differences and avoid capturing risk factors that correlate with the use of high/low cost facilities but do not impact the actual care pathway for the patient. Since inpatient spend typically constitutes a large share of episode spend and may show considerable unit price variation, the episode spend used for risk adjustment is calculated using DRG base-rate-normalized inpatient spend.
- The effect of outlier episodes with very high episode spend should be limited to ensure that the estimates are not unduly biased by a small number of anomalous episodes. Therefore, for the purpose of risk adjustment, spend of outlier episodes is capped at 3 standard deviations above the mean episode spend of valid episodes. Episodes with very high episode spend are capped instead of excluded in order to generate more accurate and equitable predictions.
- The risk adjustment process should take into account that healthcare claims data often has a "long tail" of high spend claims. Regression models where the effects of the risk factors interact multiplicatively have the potential to better fit healthcare claims data that has a "long tail". Such models often (though do not always) predict better than a linear OLS model. Therefore, in the initial stages of the risk-adjustment process, episode spend is used either with a log-transformation or within a log-linked regression model. If the exponential models fail to fit the distribution of spend observed in the data then a linear model is tested.
- The model used to estimate the final risk coefficients should predict episode spend without systematic bias. Therefore, either an OLS or a log-linked GLM based on the Poisson distribution is used to estimate the final risk coefficients, depending on the model that best fits the observed distribution of episode spend. The primary alternative to GLM, using an OLS model on log-transformed episode spend, produces similar risk coefficients but has the downside of requiring its predictions be retransformed back into non-log-transformed episode spend. The retransformation may lead to a "retransformation bias," where the retransformed episode spend no longer is an accurate estimate of the observed episode spend. Methods are available to correct the bias, but

because the GLM is fit on the same scale as episode spend, it generally has lower prediction error than the retransformed log prediction even after bias correction.

While the log-linked GLM regression is the preferred method for estimating the final risk coefficients, as it tends to fit observed medical spend data, calculating a GLM regression is much more computationally intensive than OLS. Given this technical constraint, GLM is not practical for analyses that require generating many hundreds (or thousands, or even millions) of regression models, as some of the intermediary steps in the risk adjustment process require. Therefore, the initial data mining to inform the clinical advisory group as well as the exhaustive search for potential models use an OLS regression on log-transformed episode spend. Since evaluation of the outputs from these steps does not require re-transformation to non-log-transformed episode spend, the use of an OLS model on log-transformed episode spend does not bias model selection or evaluation.