

Support for GLM and Complex Models Used for Individual, Small Group & Association Group

Scope: Applies to all Montana ACA Individual and Small Group Major Medical rate filings that use GLMs or other complex statistical / machine-learning models in developing rating factors or projecting claims.

Provide an answer and/or support for the following items; if not applicable, state “N/A” with a short explanation.

I. GLM Model Support

(A) Identify Each GLM Model Used in the Rate Filing.

Describe each GLM model used in the rate filing addressing at least the following issues:

- 1) Describe the model’s role in the rating process and provide the reasons why the model is an appropriate choice for that role.
- 2) Indicate if multiple GLMs are used in the model (i.e. Separate GLMs for pharmacy, medical, other service categories); Provide a chart indicating how GLMs are related in the process. Indicate if Logistic Regression is being employed.
- 3) Disclose reliance on data and models supplied by others, and state reliance on the outside actuary or expert who is responsible for developing or maintaining the model.
- 4) Identify the software used to build and fit the model to the data.
- 5) Provide the qualified actuary’s statement addressing reliance on external data/models and ASOP 56.
- 6) If the filing includes first-time use of a GLM model, an update to a prior model or replacement with a different model, then identify and explain the changes in calculations from the prior filing. Measure and describe the rating impacts on renewing policies due to the changes, and describe the process used by management to mitigate or get comfortable with those impacts to members.

(B) Describe the Data Used to Develop the GLM Model.

- 1) Identify all data sources used to develop the model, list the data elements in each source, and identify the data elements selected for use in the model.
- 2) Provide a summary of the experience used to develop the model, indicate the experience periods of the data used, indicate the lines of business, geographic regions that were used, and summarize the demographics of the underlying experience that was used (age / gender distributions, plan designs).
- 3) Provide an ASOP 23 summary: data sources, key limitations/concerns, and how you evaluated reasonableness and consistency.

Support for GLM and Complex Models Used for Individual, Small Group & Association Group

- 4) List third party data sources (including vendors) used in the models and key variables taken from each.
- 5) State approximate missing data percentages for key rating/input variables and how missing or “unknown” values are treated.

(C) Provide a Summary of the Experience Input Into the Model for the Plan Being Rated.

- 1) Summarize the plan’s claim, premium, membership (member months), demographic data and experience that is input into the GLM model for rating. Claims should be summarized by service categories (Inpatient, Outpatient, Physician, Pharmacy, Other), split out by utilizations and unit costs.
- 2) Explain procedures used to merge data from different sources, and creation of the datasets used in the model inputs.
- 3) Document the process for reviewing the appropriateness, reasonableness, consistency, and comprehensiveness of the data. Disclose material findings from the data review and identify any potential material limitations, defects, or unresolved concerns found or believed to exist in the data.
- 4) Document any audits that have been performed on the databases used in the model to verify that the data reconciles to the claims, premium, membership, financial data systems of the company.

(D) Describe the Model Variables and Development of Model Weights.

- 1) Provide a data dictionary containing the following: (i) variable name or abbreviation; (ii) detailed definition/description; (iii) data type (i.e., nominal, ordinal, discrete, or continuous); (iv) any modifications applied to data (e.g., transformation, capping, trending, scaling, etc.); (v) treatment in model (e.g., modeled, control, offset).
- 2) Provide a description of the Condition Identification Algorithms utilized in the model to identify health conditions and diseases.
- 3) Indicate any Risk Groupers that are being utilized by the model, such as: Episode Treatment Groupers (ETGs) based on episodes of care, Hierarchical Condition Categories (HCCs), Diagnostic Cost Groups (DCG), Diagnosis Related Groupers (DRGs), NDC J Codes, LOINC, or other groupers that are used.
- 4) Define each predictor variable used in the model.
- 5) Describe any procedures or filters that had the effect of excluding or adjusting any of the observed data prior to the model fitting process. Adjustments or modifications made to the data would include trending, development, exclusion of large claims, capping individual claims, deletion of anomalous observations, techniques to deal with missing data, etc. Provide reasons for each such adjustments and modifications.
- 6) Provide all lists of weights that were developed. This includes weights by disease/condition category that were developed and are applied to determine member’s risk scores.

Support for GLM and Complex Models Used for Individual, Small Group & Association Group

- 7) Provide the formula relationship between the data inputs and the model outputs, with a definition of each model input and output. If the model includes offset variables, identify and explain. Indicate whether the model estimates expected future claims, loss ratios, pure premiums, frequency/severity, or something else.
- 8) Describe what steps were taken to identify high pairwise correlations amongst predictors and multicollinearity within the model. Provide the following: (i) describe any identified issues and explain how they were resolved; (ii) state what the largest pairwise correlation is in the data set; (iii) correlation matrix.

(E) Describe the Use of Model Outputs in the Rating System.

- 1) Indicate how the model outputs are used within the rating system.
- 2) Indicate what calculations, judgments and adjustments, if any, were made before using the model output in the rating system.
- 3) Indicate how Model Outputs used in rating the plan are being applied alongside other rating variables so as to not double count member's future expected claims. Describe the process that was used, either within the GLM model, or in the process of applying outputs and rating factors together, to ensure the outputs and rating factors have been combined appropriately for all rating cells.

(F) Describe How the Model Was Tested for Significance.

- 1) Describe the methods used to assess the statistical significance/goodness of the fit of the model, such as lift charts, statistical tests, and holdout samples. This assessment should include model projection results compared to historical actual results to verify that modeled results bear a reasonable relationship to actual results.
- 2) Provide all statistical measures to indicate the model predictive value and goodness- of-fit (R, R-squared, Adjusted R-squared, Standard Error, Model F-Ratio and its significance).
- 3) Provide statistical summaries to indicate that each variable in the model is statistically significant and is predictive for future expected claims (F-test, t- test, p-values, other).
- 4) If applicable, indicate the Sensitivity of the model – the % of members with a condition who are identified; and indicate the Specificity of the Model – the % of members identified with a condition who do not have the condition.
- 5) Provide the output coefficient table showing the coefficients for each variable used in the models. This should typically include for each variable:
 - a) Unstandardized coefficient indicating how much the dependent variable varies with the independent variable when all other independent variables are held constant;
 - b) Standardized coefficient beta;
 - c) Confidence interval for each coefficient;

Support for GLM and Complex Models Used for Individual, Small Group & Association Group

II. Support for Other Complex Models

(A) Identify all other complex models used, which may include:

- 1) Machine Learning models, including but not limited to:
 - i. Neural networks;
 - ii. Gradient boosting trees;
 - iii. Random forest method.
- 2) Time Series models
- 3) Decision Trees
- 4) Generalized Bayesian Models
- 5) Non-Linear Regression Models
- 6) All other Complex Models.

Describe each model and identify its role in the rating process and provide the reasons why that type of model is an appropriate choice for that role.

Provide full support similar, where applicable, to what was identified in the Sections I.A to I.F above to be provided for GLM models