# Systemization of Logistic Regression Analysis for Pharmacometric Applications

J Passarell, C Passarell, D Hitchcock, J Fiedler-Kelly, TH Grasela

Cognigen Corporation, a subsidiary of Simulations Plus, Inc.; Buffalo, NY
cognigencorp.com

W-49

## ABSTRACT

**Objectives:** Efficacy and safety data are oftentimes collected as binary (yes/no) data in clinical trials during drug development. The implementation and growing use of CDISC standards in data collection and categorization of adverse event data using the MedDRA dictionary has facilitated the standard format of clinical trial data collected across the pharmaceutical industry. A commonly used statistical methodology for analyzing binary data is logistic regression (LR) analysis. The objective of this work was to develop a system to standardize analysis dataset creation, exploratory data review, and LR analysis procedures for exposure-response analyses of binary endpoint data.
**Methods:** SAS® software was used to develop a code library to transform source clinical trial data into an analysis-ready dataset for use in exposure-response analyses. A library of SAS® code for the creation of standard exploratory graphs and tables was also developed. A systematic approach to statistical analysis using SAS® PROC LOGISTIC and NONMEM was developed, based on standard methods for model building and discrimination, to facilitate the calculation of standard statistics and production of typical diagnostic plots for model building and evaluation.[1]
**Results:** The standardized process for dataset creation, exploratory data analysis, and LR was tested on 10 compounds and refined as new variations and additional data checks were identified. This refined process and systematic approach resulted in a greater than 70% decrease in analyst time required for evaluation of exposure-response relationships for binary endpoints. Other positive benefits of system implementation include a reduction in training time for new pharmacometricians and improved quality and consistency of reporting for LR exposure-response analyses.
**Conclusions:** Standardization of analysis-ready dataset creation, exploratory graphical evaluation, and the LR analysis process for binary endpoints has proven instrumental in generating timely understanding of exposure-response relationships to facilitate model-based decision making under tight timelines and allows for the evaluation of additional endpoints and synthesis of findings across endpoints.

## METHODS

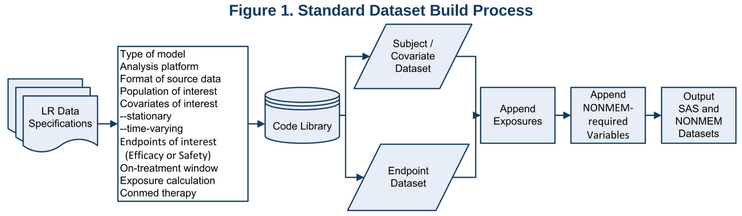
### Data Considerations

- When pooling multiple studies
  - Will individual study datasets be used? Use of individual study datasets requires pooling and can involve complicated data standardization routines.
  - Is there an integrated efficacy or safety analysis dataset available? This is the simpler coding case.
  - Determine if the efficacy and safety endpoints were collected and defined the same across studies.

### Analysis Dataset Construction

- A dataset requirements form was developed specifically for the evaluation of exposure-response relationships using LR.
  - Used to communicate the structure and content of the dataset required for the analysis.
  - Requires the input of key pieces of information for the analysis from source data (for example, endpoints, time, covariates, etc.) and facilitates the selection of programming templates from the code library used for assembly of the analysis dataset.
  - Provides the list and order of data deletions.
- A standard dataset build process was developed (**Figure 1**):
  - Step 1 - Covariate dataset is built based on the requested stationary and time-varying covariates specified in the requirements, for those patients in the population of interest.
  - Step 2 - The endpoint data are processed from the derived (ADaM) or source datasets. Records are added for patients with “no event” if such data were not collected (for example, adverse events that did not occur).
  - Step 3 - The required covariates are appended onto the endpoint records.
  - Step 4 - Individual exposure measures are appended onto the merged endpoint and covariate data.
  - Step 5 - NONMEM®-required variables (for example, MDV) are appended to the merged endpoint, covariate, and exposure data.
- Standard rules for data checking and handling of data anomalies are built into the code templates (for example, management of missing dates and imputation of missing covariates).
- Existing program templates can be used to incorporate concomitant medications and/or generation of drug exposures, as needed.
- Template programs have been developed and quality controlled (QCd) to accommodate these and other variations of the data assembly process.
  - Template programs provide a solid starting point and can be used as is or customized if the data, study design, or requirements do not align with the templates.
  - These new or updated templates are then added to the code library for use by the entire data programming department to facilitate and reinforce the use of standardized, QCd code on future projects and programs.

**Figure 1. Standard Dataset Build Process**




### Exploratory Data Analysis

- A library of SAS® code for the creation of standard exploratory graphs and tables specifically for binary endpoint data was developed to facilitate rapid evaluation and assessment of exposure-response relationships. Exploratory plots of raw data provide enhanced understanding of the informational content of the data relative to the models to be tested and evaluations to be performed.[2]
- Standard exploratory data analysis library includes:
  - scatterplot matrices of individual exposure measures and calculation of correlation measures for each pair of exposures;
  - scatterplot matrices of covariates of interest and calculation of correlation measures for each pair of covariates;
  - frequency distributions of the individual exposure measures, overall and stratified by binary endpoint (response);
  - boxplots of exposure measures by binary endpoint;
  - exposure-response quantile plots of the estimated probability of response in quantiles (bins) of the exposure distributions plotted against the mean or median exposure in that bin;
  - summary statistics for each of the patient covariates and exposure measures, overall and stratified by binary endpoint;
  - tabulation of the observed binary endpoint for each level of the various categorical patient covariates (that is, contingency tables);
  - boxplots of continuous patient covariates, stratified by binary endpoint; and
  - empirical logit plots versus continuous covariates and drug exposure measures. Simple binning method was used to group the continuous covariates and drug exposure measures.[3]

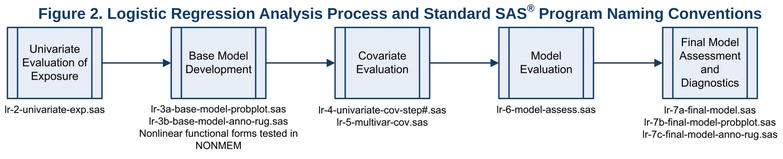
$$Empirical\ Logit_i = Ln\left(\frac{y_i + 0.5}{M_i - y_i + 0.5}\right)$$

Where:
$y_i$ is the number of patients with the occurrence of the event of interest in the $i$th group of the continuous patient covariate or exposure measure (predictor); and
$M_i$ is the total number of patients in the $i$th group of the continuous patient covariate or exposure measure.

### Logistic Regression Analysis

- A systematic approach to statistical analysis of binary endpoint data using SAS® PROC LOGISTIC and/or NONMEM® was developed.[4,5]
- Standard LR methods for model building and discrimination are used to facilitate the calculation of standard statistics and production of typical diagnostic plots for model building and evaluation.[6]
- Standard process of LR analysis is shown in **Figure 2**.

**Figure 2. Logistic Regression Analysis Process and Standard SAS® Program Naming Conventions**

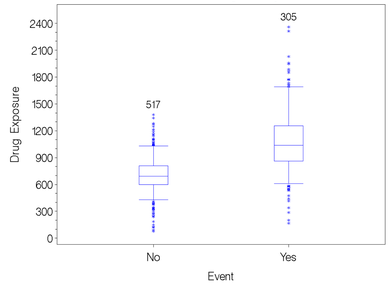



## RESULTS

### Exploratory Data Analysis

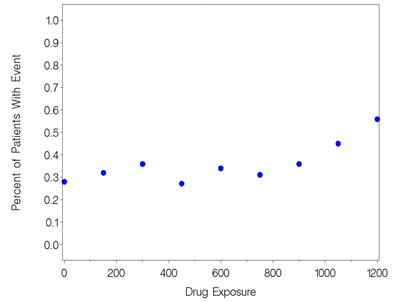
- Example boxplot of drug exposure versus binary endpoint illustrates the typical relationship between exposure and response (**Figure 3**).
- Example plot of the percent of patients with the event of interest versus grouped drug exposure (**Figure 4**) shows whether the percent of patients with the event increases or decreases with increasing drug exposure.
- Example plot of empirical logit versus grouped drug exposure with smoothing spline (**Figure 5**) helps determine the functional form of the relationship to be formally modeled.

**Figure 3. Boxplot of Drug Exposure Versus the Occurrence of Binary Endpoint**



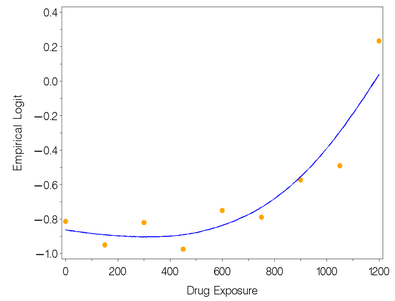

Boxes are 25th, 50th, and 75th percentiles; whiskers are 5th to 95th percentiles. Asterisks show data points outside this range. The number of subjects is above each box.

**Figure 4. Percent of Patients With Event Versus Drug Exposure**




The circles represent the median exposure and associated observed probabilities.

**Figure 5. Empirical Logit Plot Versus Drug Exposure With Smoothing Spline**




### Logistic Regression Analysis

- Evaluation of drug exposure in the base model
  - Example table of univariate fits of each drug exposure measure (**Table 1**) can be produced automatically using either the standard SAS® program or the KIWI™ application[7] if NONMEM® is used as the analysis tool.

**Table 1. Summary of Drug Exposure Evaluation for Exposure-Response Analysis of Binary Endpoint**

| Exposure Measurement | Functional Form | Change in VOF[a] | Degrees of Freedom | *P* value[b] |
|---|---|---|---|---|
| Reference model: minimum VOF = 1084.235 | | | | |
| $AUC_{(0\text{-}24)}$ | Linear | -9.809 | 1 | 0.0017 |
| $C_{max}$ | Linear | -9.416 | 1 | 0.0022 |
| $AUC_{(0\text{-}24)}$ | Power | -3.794 | 1 | 0.0514 |
| $C_{max}$ | Power | -3.517 | 1 | 0.0608 |
| $C_{min}$ | Power | -1.536 | 1 | 0.2153 |
| $C_{min}$ | Linear | -1.228 | 1 | 0.2678 |

Abbreviations: $AUC_{(0\text{-}24)}$, area under the concentration-time curve from time 0 to 24 hours; $C_{max}$, maximum observed drug concentration; $C_{min}$, minimum observed drug concentration; VOF, value of the objective function.
[a] Change in the value of the objective function relative to the reference model.
[b] Statistical significance (α = 0.05).

- Covariate evaluation - example table illustrating forward selection results is provided in **Table 2**. This table is automatically generated using the KIWI™ application[7] and can be exported as a QCd, formatted Word® table for direct import into technical reports or presentations.

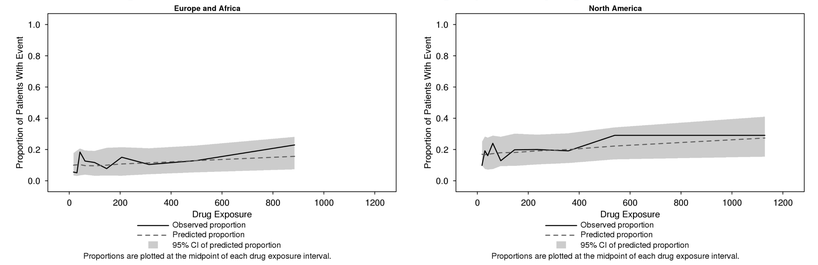
**Table 2. Summary of Forward Selection of Covariates for the Exposure-Response Analysis of Binary Endpoint**

| Step | Covariate Added | Functional Form | Change in VOF[a] | Degrees of Freedom | *P* value[b] |
|---|---|---|---|---|---|
| Reference Model Minimum VOF = 1084.23 | | | | | |
| 1 | Region | Additive | 67.2170 | 4 | <0.0001 |
| 1 | Age (y) | Linear | 20.5211 | 1 | <0.0001 |
| 1 | Baseline Weight (kg) | Linear | 12.3168 | 1 | 0.0004 |
| 1 | Sex | Additive | 10.7533 | 1 | 0.0010 |
| 1 | Race | Additive | 14.4607 | 5 | 0.0129 |
| Reference Model Minimum VOF = 1017.018 | | | | | |
| 2 | Baseline Weight (kg) | Linear | 3.09802 | 1 | 0.0784 |
| 2 | Sex | Additive | 1.88990 | 1 | 0.1692 |
| 2 | Race | Additive | 7.03245 | 5 | 0.2182 |
| 2 | Age (y) | Linear | 1.27319 | 1 | 0.2592 |
| No covariates significant at α = 0.05 | | | | | |

Abbreviations: VOF, value of the objective function.
[a] Change in the value of the objective function relative to the reference model.
[b] Statistical significance (α = 0.05).

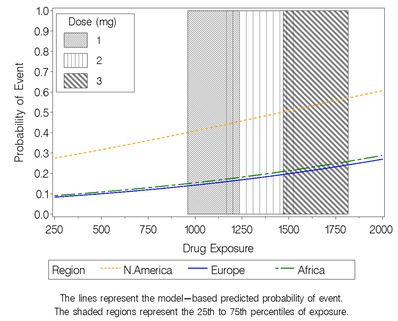
- Model evaluation using Hosmer-Lemeshow goodness-of-fit $\chi^2$ and area under the receiver operating characteristic curve.
  - These standard statistical tests are automatically output as part of the model evaluation process.[1,8]
  - Visual predictive check plots[9] can be produced using the KIWI™ application[7] to assess the predictive ability of the model (**Figure 6**).
- Final model
  - Standard code reads in the final model output to produce a plot of the final model predictions versus drug exposure, stratified by significant covariate(s) with shaded regions representing the 25th to 75th percentiles of drug exposure for each dose level (**Figure 7**).
  - Standard code is used to plot the observed and predicted probability of the event versus drug exposure, stratified by significant covariate(s) (**Figure 8**). The hash marks near the x-axis represent the individual exposure values in the patients who experienced the event. The symbols represent the observed proportions of patients who experienced the event in bins of exposure, plotted at the median exposure for the bin. The simple binning method was used to construct a set of empirical probabilities that represent the data and were compared to the model-predicted probability of event.[3]
  - These figures provide a better understanding of the concordance between the model-predicted probabilities of event and the observed occurrence over the range of drug exposure.

**Figure 6. Visual Predictive Check Versus Drug Exposure, by Significant Covariate**



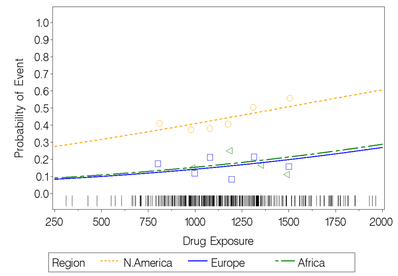

Proportions are plotted at the midpoint of each drug exposure interval.

**Figure 7. Final Model-Predicted Probability of Event Versus Drug Exposure, by Significant Covariates**




The lines represent the model–based predicted probability of event.
The shaded regions represent the 25th to 75th percentiles of exposure.

**Figure 8. Final Observed and Model-Predicted Probability of Event Versus Drug Exposure, by Significant Covariate**




The lines represent the model–based predicted probability of event.
The symbols represent the median drug exposure and associated observed probabilities.
The hash marks near the x–axis represent the individual drug exposure for patients with an event.

## CONCLUSIONS

Standardization of analysis-ready dataset creation, exploratory graphical evaluation, and the LR analysis process for binary endpoints has proven instrumental in generating timely understanding of exposure-response relationships to facilitate model-based decision making under tight timelines and allows for the evaluation of additional endpoints and synthesis of findings across endpoints.

## REFERENCES

1. Hosmer DW, Jr, Lemeshow S. *Applied Logistic Regression*. 2nd ed. New York, NY: John Wiley & Sons; 2000.
2. Owen JS, Fiedler-Kelly J. *Introduction to Population Pharmacokinetic/Pharmacodynamic Analysis with Nonlinear Mixed Effects Models*. Hoboken, NJ: John Wiley & Sons Incorporated; 2014.
3. Pavan Kumar VV, Duffull SB. Evaluation of graphical diagnostics for assessing goodness of fit of logistic regression models. *J Pharmacokinet Pharmacodyn*. 2011;38(2):205-222.
4. SAS [computer program]. Version 9.2. Cary, NC: SAS Institute; 2009.
5. NONMEM [computer program]. Version 7, Level 1.2. Ellicott City, MD: ICON Development Solutions. 2010.
6. Agresti A. *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons, Inc., 1996.
7. KIWI [computer program]. Version 1.3. Buffalo, NY: Cognigen Corporation; 2015.
8. Rana S, Midi H, Sarkar SK. Validation and performance analysis of binary logistic regression model: proceedings of the WSEAS international conference on environment, medicine and health sciences. United States Branch of the World Scientific and Engineering Academy and Society Web site. Page 51-55; http://www.wseas.us/e-library/conferences/2010/Penang/EMEH/EMEH-09.pdf. Accessed March 28, 2012.
9. Bergstrand M, Hooker AC, Karlsson MO. Visual Predictive Checks for Censored and Categorical Data. http://www.page-meeting.org/pdf_assets/7002-Poster_PAGE_VPC_090618_final.pdf.

Passarell J, Passarell C, Hitchcock D, Fiedler-Kelly J, Grasela TH. Systemization of logistic regression analysis for pharmacometric applications. Poster presented at: Sixth American Conference on Pharmacometrics (ACoP6) ; October 4-7, 2015; Crystal City, Virginia.

For additional information, please contact
Julie Passarell
Cognigen Corporation, a subsidiary of Simulations Plus, Inc.
1780 Wehrle Drive, Suite 110, Buffalo, NY 14221 USA
+1-716-633-3463, ext. 266 or julie.passarell@cognigencorp.com