Using *R* for Data Mining in Vaccine Manufacturing: *Finding Needles in Biological Haystacks*

Nelson Lee Afanador

Center for Mathematical Sciences
Merck Manufacturing Division
Merck & Co., Inc.
WP28-100
P.O. Box 4
West Point, PA 19486
215-652-0067
215-993-1177
nelson.afanador@merck.com

Presenter: Nelson Lee Afanador

Key Words: Data Mining, Random Forest, Recursive Partitioning, Vaccine Manufacturing

Purpose: To illustrate the application of data mining tools available in *R* in helping drive at root causes for changes in either cellular growth or viral propagation. Two vaccine manufacturing case studies will be presented.

Abstract

Vaccine manufacturing is the latest field in which advanced data mining methods are beginning to gain widespread acceptance among biologists, engineers, and data analysts. Vaccine manufacturing itself is a complex biological process composed of hundreds of steps carried out over several months. During the manufacture of a single vaccine lot hundreds of processing variables and raw materials are monitored. The challenging aspects with respect to mining biological manufacturing process data begins with obtaining a suitable dataframe which to analyze, proceeds to inherent variation in raw material composition and processing conditions, and ends with high inherent variation in the measurement systems. As such, identifying the root cause candidates for changes in cellular growth or viral propagation is extremely challenging due to the high number of candidate variables and variable processing conditions.

Given the large numbers of available candidate variables the traditional methods of univariate statistical process control charting, analysis of variance, and least squares regression leave many questions unanswered. Random Forest (**randomForest**) and single-tree recursive partitioning (**rpart**), coupled with cumulative sum charts (**qcc**), have proven to be important methods in helping drive at potential root causes for observed changes. Leading candidate variables identified via the overall analysis can then be further investigated using more traditional statistical methods, including designed experiments.

These data mining methods are setting a new standard for vaccine root cause investigations and have proven valuable at helping solve complex biological problems. Their effectiveness, and implementation using *R*, will be illustrated with two case studies.