PhUSE EU Connect 2018

Paper AR05

Implementing Advanced Analytics and Machine Learning in Risk Based Strategies

Rajneesh Patil, IQVIA, Raleigh-Durham, USA

ABSTRACT

An integral part of any future strategy deploying RBM is the use of advanced data sciences. The speaker shall share how new age contemporary analytics and machine learning methodologies are driving a significant transformation in clinical trial monitoring.

Shall discuss the evolution of advanced analytics and how we could apply contemporary statistical science to advance the approach to risk identification and mitigation. The paper will share learnings around practical models that helped reduce white noise/false positives in the alert systems deployed for RBM. Shall also share certain specific examples of machine learning models used in certain therapeutic areas to address protocol specific risks based on patient population chosen per study design.

INTRODUCTION

The presentation shall review concepts of machine learning and artificial intelligence in the context of risk-based monitoring (RBM) implementation and how clinical operations organizations can build foundational elements to enable such capabilities.

ADVANCED ANALYTICS

Most common approaches to key risk indicator (KRI) management rely upon traditional statistical methods. This implies a reliance on approaches where mean, median, mode and standard deviation are used to identify outlier sites. Usually these methods follow the traditional method of setting thresholds and then utilize a standard deviation measure to define outlier behavior. While these methods could be useful, in our experience we found these to yield significantly high false positive results (white noise), thereby reducing the value of monitoring effectiveness.

To solve the problem of increased false positive results in KRI management, more contemporary statistical methods could be used. Also, the concept of thresholds is not required because these methodologies can allow for 'relative risk assessment' between clinical sites as against comparing the sites to study level threshold.

For example: Instead of looking at each KRI metric as an individual parameter; we can combine several KRI metrics together using methods like poisson-gamma distribution and/or beta-binomial distribution.

Examples of KRI's: Screen failure rate, Adverse Event rate, SAE rate, Protocol Deviation rate, Query rate

Model-based trial monitoring analytics (Poisson-gamma model)

Assume a trial with N clinical sites. Suppose that in center i, the events are happened according to Poisson Process with rate λ_i, where λ_i is viewed as independent gamma distribution with parameter Ga (α, β).

Basically, this is a hierarchical (multilevel) model, meaning the coefficient is itself given a model. When we have N sites with number events and exposure time or site exposure time for the events, the event rate for site i follow a distribution with parameters λ_i (first model with coefficient $Poisson$ (λ_i)). Then, λ_i itself has a model of independent gamma distribution with parameters Ga (α, β). So in this case, the number of events with corresponding exposure time follows a multilevel model, which is Poisson-gamma model with parameters Ga (α, β).

Estimate parameter
In statistics, after we assume the data follows a model, the next step would be calculating the parameters for the model. Using statistical algorithms (such as MLE), we may easily get parameters α, β for Poisson-gamma model.
Consideration: Events are occurring on an ongoing basis in the given time frame for each of the sites in a clinical trial.

Following the successful application of poisson-gamma distribution and/or beta-binomial distribution; the 'p value' scores are secured for each site. These scores are then subject to an assessment through Mahalanobis distance.

Mahalanobis Distance is useful in detecting outliers in multivariate data, meaning that by looking at individual KRIs' p-values, a data point might not be the outliers, but considering them all using Mahalanobis Distance, the data point might be a outlier. It is useful in detecting multivariate outliers. A point can be a multivariate outlier even if it is not a univariate outlier for any variable, making Mahalanobis distance a more sensitive measure than checking dimensions individually.

By employing these methodologies of 'composite assessment of site KRI's; the false positives are reduced significantly because these observations take into account factors like site open duration, patient volume at a site and patient exposure to study drug.

MACHINE LEARNING
Data in clinical trials and medical research are collected from various sources and formats with increasing volume and complexity. Labeling (e.g., signaling a site with quality issue, identifying anomalies in the lab test results, diagnosing diseases, and recruiting eligible subjects) is a critical process of transfer raw data to knowledge and then to decision making. However, generating labels can be very time-consuming and labor-intensive particularly in large complex data. For instance, over 500 charts reviews and 100 months are needed to recruit 10 patients for a non-alcoholic steatohepatitis (NASH) study. Unlabeled data also impede misconduct detection and monitoring of patient safety and site health.

Unsupervised machine learning (ML) helps close this gap using label-free statistical computation to detect risks and recognize complex patterns in large data from multiple sources. Whereas supervised ML relies on labeled data to understand associations and make predictions, unsupervised ML leverages data-driven techniques (e.g., hierarchical clustering, PCA, and isolation trees) without a priori training or too much parameter tuning. Unsupervised ML can even be utilized to create labels, detect targeted population, and rank risk levels, making it an essential step before supervised learning and advanced predictive analytics.

This paper deals with examples of implementing novel unsupervised ML algorithms to tackle the lack-of-label issues in the clinical and medical data. The applications of these algorithms have been shown to effectively detect anomalies in the lab tests and vital signs, which in turn enables near real-time automation in the centralized monitoring. By combining data mining techniques, unsupervised ML algorithms can promisingly improve the efficiency in patient recruitment and misconduct identification, and also allow early detection or noninvasive diagnosis of early-stage asymptomatic yet dangerous diseases (e.g., dementia and NASH).

The presentation component of this paper will represent examples around machine learning methods used in outlier patient detection and predictive models to identify predictors of delayed or under reporting of adverse events across sites.

Types of models implemented and compared:
- Random Forest (RF)
- Gradient Boosting Trees (GBM)
- Deep Neural Network (DNN)
- Logistic Model with Regularization (GLM)

Tune important hyper parameters (e.g., GBM model):
- Number of trees

- Learning rate
- Learn rate annealing
- Max depth of tree
- Min improvement at each split

In one case study, a stacked model approach was used to improve overall accuracy, precision and recall to over 90%. Therefore increasing the reliance of predictions significantly as compared to individual models.

CONCLUSION

1] Use of advanced and contemporary statistical models can provide composite assessment of site risk score by combining several key risk indicators into one comprehensive risk score. This allows for significant improvement in specifically identifying true risk at clinical sites as it applies to operational KRI's.
2] Use of machine learning models are effective in subject outlier detection and predictive modelling to identify sites at risk in factors like delayed or underreporting

REFERENCES

https://en.wikipedia.org/wiki/Mahalanobis_distance
https://en.wikipedia.org/wiki/Negative_binomial_distribution

ACKNOWLEDGMENTS

Data scientist team at IQVIA: Jian Huang and Wei Xue

RECOMMENDED READING

This section is not required.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Rajneesh Patil
Company: IQVIA
Address: 4820 Emperor Boulevard
City / Postcode: Durham 27703
Work Phone: + 1 984.219.4115
Email: rajneesh.patil@iqvia.com
Web: www.iqvia.com