***Machine Learning Framework for Predicting Vaccine Immunogenicity***

My chosen article is titled “Machine Learning Framework for Predicting Vaccine Immunogenicity” can be found at:

<https://www.informs.org/Impact/O.R.-Analytics-Success-Stories/Machine-Learning-Framework-for-Predicting-Vaccine-Immunogenicity>

**Problem definition**

It is very important to predict how each person will respond to vaccination differently. This can help us to understand what the best vaccination is to protect individual from infection and also improve the design and invention of new vaccine.

**Vaccine procedure**: Most of vaccine design is based on small clinical trials. In the vaccination process each indivisible receive a virus and in most cases T cell and bells will be activate and may secrete antibody and their correspondent RNA will be generated ,eventually these compounds will be recorded. And this procedure will be repeated in many trial and error cases to finalize the result.

Measuring the innate response of immune cell(T cell and B cells) in each individual after receiving vaccine can determine does the vaccine had effect or not.

In most cases the RNA data many days after vaccination is used to evaluating the vaccine efficacy.

**Limitation:** many of the first receivers of these vaccines are the doctors (since many people do not volunteer for these studies

In some cases, like flu vaccine, we need to identify potential virus 6-8 months before season, in a trial and error approach

**Goal:** Since Effective of the vaccine is unknown, prediction of vaccine efficacy and immunity is the main goal for this research.

Researchers from CDC, Georgia Tech, and Emory University have developed machine learning framework, named DAMIP, for finding gene signatures that can predict vaccine immunity and efficacy.

**Question**:

In this project we would like to know *what are the potential machine learning model might be used to create the solution? What are the steps?*

*What kind of data should be used?*

*How we can predict whether individual will benefit from vaccine or not?*

**Proposed analytics models**

**Data collection:**

* *For initial response*: blood level of secreted antibody by T cells response after receiving vaccine
* *For prolong response*: we needmicroarray RNA data which have been produced in each individual many days after receiving vaccine.

These RNA genes signatures that can be used for prediction of T cell response and vaccination efficacy

**Sampling:** In this project we need to do random sampling. In this case each individual is chosen entirely by chance and each member of the population has an equal chance, or probability. we should consider the sample to be representative to the target population, as much as possible, with the least possible error and without substitution or incompleteness.

We also need to controlfor possible interaction of some variables in our analysis. For example, if we have data of most old individual it might introduce bias to our models.

**Model 1:**

**Given/Variables**: data of antibody secreted by T cell in different days after vaccine and different gene signatures of microarray RNA

**Model:** *Correlation analysis*

**To:** see if there is association betweenantibody secreted by T cell and different microarray RNA gene

*Correlation analysis*: using data of antibody secreted by T cell in different days after vaccine and gene signatures to identify genes signatures that are correlated with antibody production

With usage of correlation matrix, the gene signatures associated with antibody secreted by T cell can be identified

*In the next step:*

**Model 2:**

**Given/Variables:** Genemicroarray RNA data and vaccination RNA data

**Model:** Clustering Classifiers

**To:** find most effective group of gene signature RNA data

*clustering Classifiers*: clustering gene signature (RNA data) acquired from previous part and vaccination RNA data output to the three classification of the most effective and moderate effective group and non-effective group

**Model 3:**

**Given/Variables:** most effective and moderate effective group RNA data and secreted antibody level in individuals

**Model:** Logistic Regression

**To:** predict the probability of effectiveness of the gene signatures

*Logistic Regression*: Using the most effective and moderate effective group RNA data and secreted antibody level in individuals to identify the probability of effectiveness of the gene signatures in vaccination.

***variable selection*:** Use variable selection like stepwise or lasso to identify the genes that are mostly have higher effect on our regression output and explain most of the variability in the data. (alpha of 0.05 is used as the cutoff for significance)

**Last step:**

we can test the blind data on our proposed regression model to predict how much will be the vaccine effective on each individual.

Then we can compare (visualized) the blind and real prediction data to see how much the accuracy of our model is.

**Future direction**

We can also re-run the model over longer periods of the times on our individual data to see how much the efficacy the vaccine has been preserved and how long this efficacy is preserved.

And we can redo the later model for based on the more longer-term gene signatures.

We can also retest our model on different age group to see how much the effect is different.

We can look at these analyses on different sample groups like Infants, elderly people to see how much of this efficacy is preserved.