An empirical investigation of factors driving lactate levels in sepsis patients using machine learning methods

Saraswathi Sathees1^, Shamsudheen Marakkar2, Ramya Vunikili3, Shabeeb Kannattikuni3, Kipp W Johnson4, Benjamin S Glicksberg6, Joseph Scarpa7, Anoop Vadakkan Cherian8, Oommen K Mathew9, Lakshminarayanan Subramanian3, Joel T Dudley4 and Khader Shameer10*

1University of Connecticut, Storrs, CT; 2Molecular Robotics, Kerala, India; 3Courant Institute, New York University, NY; 4Georgetown University, Washington, DC; 5Institute for Next Generation Healthcare, New York, NY; 6Bakar Computational Health Sciences Institute, UCSF, CA; 7Cornell University, New York, NY; 8Max Planck Institute for Heart and Lung Research, Germany; 9Indian Institute of Information Technology, Kottayam; 10Northwell Health, New York, NY

*Corresponding author; ^ Presenting author

Introduction: Sepsis is a life-threatening disease caused as a result of the body's response to infection [1]. According to the World Health Organization, it is estimated that more than 30 million people are affected by sepsis worldwide every year, potentially leading to 6 million deaths. Early diagnoses of sepsis are critical to reducing clinical acuity of sepsis. Recent studies have shown that an increase in lactate level rate as a crucial biomarker to diagnose sepsis. However, it is unclear how lactate level increase with other routinely measured clinical biomarkers or survival rate in sepsis population. We hypothesize that lactate dynamics will have an impact on other clinical biomarkers, and changes in lactate levels could be detected using a combinatorial, digital biomarker composed of routinely measured laboratory tests.

Methods: Briefly, we compiled a training cohort and testing cohort with patients diagnosed with sepsis and associated laboratory data from MIMIC databases. A validation cohort was compiled using phenomic data in UK Biobank for validation. Collectively, essential variables affecting the lactate levels are identified. A library of features was evaluated for correct classification type (lactate increase, lactate decrease, lactate increase with sepsis test definite and lactate increase with sepsis test negative) and adjusted if necessary. In the Model stage, after partitioning the dataset, a GridSearchCV on at least three different ML algorithms are run to identify the best parameters to use for a model. Three model interpretability plots, such as partial dependence plot, variable importance plot, and scatterplots from the best fitting model, were compiled to represent the variable importance. Lastly, in the Analyze stage- Error metrics table (i.e., RMSE, MAE, standard deviation, MAPE, bias), the variable importance plot of different algorithms is compared. The best model was chosen from the models run based on comparing the model accuracy and error metrics. Figure 1 shows the workflow diagram of the analytic approaches used in the study and can be extended to other data-driven analytics investigations.

Results: It is imperative to treat sepsis in its early stages due to its fatality. Analyzing the phenotype trajectories and applying machine learning algorithms, we can study how the lactate level rate affects the sepsis patients. Additionally, this could result in better healthcare delivery systems for the population-scale risk stratification of sepsis and improve patient outcomes.

References: