**Summary of Cohort Selection, Data Mining Methods and Algorithms**

The paper by Chicco and Jurmal (2020) presents a statistical study on *sepsis*, which can be a life-threatening condition caused by an exaggerated reaction of the body’s reaction to infection. It can lead to organ failures and eventual death in a short span of time. While there are laboratory tests available to positively test for sepsis, those can be time consuming and hence may not be fruitful for critical patients. In their study, the authors have shown using statistical analysis and machine learning algorithms that three quick-to-determine factors – age, sex, sepsis episode number – alone can provide a prediction with a high degree of confidence.

**Cohort Selection**

Chicco and Jurman (2020) analyzed a dataset made of 110,204 admissions of 84,811 hospitalized subjects between 2011 and 2012 in Norway in their paper for their *primary* and *study cohort* selection. These patients were diagnosed with infections, systemic inflammatory response syndrome (SIRS), sepsis by causative microbes, or septic shock. The data came from the Norwegian Patient Registry and the Statistics Norway agency, the details of which are provided in the reference numbers 92 and 93 of their paper.

The authors considered all of the 110,204 admissions as the *primary cohort*. The authors then used a set of selection criterion as detailed in their paper to create a *study cohort* of 19,051 admissions.

For *validation cohort*, the authors used dataset of South Korean critically ill patients whose medical records were collected between January 2007 and December 2015 and publicly released. The detail of that dataset is provided by reference number 96 of the paper. From their original dataset, the authors selected the data of 137 patients who had already 1 or 2 septic episodes

**Data Mining Methods and Algorithms**

The authors used traditional univariate biostatistics tests and machine learning algorithms to analyze and data and make predictions.

In terms of traditional statistical analysis, the authors specifically used Mann-Whitney tests and chi-squared tests to determine three prediction features for forecasting survivals.

From the perspective of machine learning, the authors used five different techniques. These are:

1. Linear SVM
2. Radial SVM
3. Naïve Bayes
4. Linear Regression
5. Gradient Boosting

The authors nine scores based on various confusion matrix thresholds to compare the performance of the different machine learning algorithms.

**References**

Chicco, D. & Jurman, G. (2020). Survival prediction of patients with sepsis from age, sex, and

septic episode number alone. *Sci Rep****10****, 17156. [https://doi.org/10.1038/s41598-020-](https://doi.org/10.1038/s41598-020-73558-3)*

*[73558-3](https://doi.org/10.1038/s41598-020-73558-3)*