# A logistic regression-based model to predict ICU mortality: problems and solutions

*Alexander Luchinin, MD, PhD*[1]*, Oleg Kolupaev, PhD*[2]*, Alexey Lyanguzov, MD, PhD*[1]

[1] *The Federal State-Financed Scientific Institution Kirov Research Institute of Hematology and Blood Transfusion under the Federal Medical Biological Agency*

[2] *UNC Chapel Hill, Lineberger Comprehensive Cancer Center*

**Background.** Logistic regression is one the most common methods of analyzing data. It is used for describing associations between variables, including those related to medical data. Here, we use logistic regression to estimate mortality rates in ICU patients with blood diseases.

**Methods.** The study included 202 patients in total with a median age of 57 (19-82) years. There were 112 males and 90 females. A total of 4242 variables were processed for both training and test dataset (80/20 split). Following independent variables were included in the analysis: sex, age, fever, respiratory rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, heart rate, need for inotropic support (Yes/No), blood hypoxemia, HGB level, PLT level, WBC level, creatinine level, CRP level, serum total protein level, serum albumin level, blood total bilirubin, blood procalcitonin, bloodstream infection detection (Yes/No) and Glasgow scale (<>15 scores) (overall 21 predictors). We performed several steps according to method assumptions for prognostic binomial model creation (death yes/no) using R version 3.4.2 and "missForest", "glmnet", "caret", "pROC" packages.

**Results.** As a first step, we have used a random forest approach (missForest package in R) to impute missing data. Next, three different methods were tested to select predictors.

1) Hosmer-Lemeshow bivariable selection. Sympson's paradox emerged when the regression coefficient of mean blood pressure became positive in death prediction contrary to systolic blood pressure. This effect can be attributed to collinearity.
2) Stepwise regression.
3) Regularization for feature selection.

The latter applied LASSO method to only a select 7 predictors instead of 21. We also used log transformation (log 2 base) to improve the quality of our model because some of our numeric variables had outliers. The absolute risk of death (mortality rate) was 67 from 202 (33%), odds – 0.496. Our model for "high-risk of death" detection classified patients in the test dataset with AUC 0.816 accuracy (95%CI 0.679-0.912, McNemar's Test p-value = 0.5), with sensitivity 0.82, and specificity 0.80. We developed an accurate and balanced model with 3 predictors: P (probability of death) = 14.13 - 0.28*log2(PLT) -2.37*log2(total protein) + 2.98*1 (if Glasgow score <15 scores) / 0 (if Glasgow score = 15). The odds of death in ICU increase ~20-fold if the patient has a Glasgow score of less than 15. Also, the odds of death increase by 1.3 and 10.7 times of PLT, or serum total protein level decreases by 2 times accordingly while other predictors remain unchanged.

**Conclusion.** Our work demonstrated that a logistic regression method allows to solve both practical and scientific problems under conditions of all statistical assumptions to this method. This method can be used both to determine the significance of predictors and create prognostic models. The information received can be used in clinical practice and in developing clinical decision support systems.