**Machine Learning Prediction of Mortality in Venous Thromboembolism Patients: The BBC-VTE Cohort**

Wahbi K. El-Bouri, Alexander Sanders, Gregory Y.H. Lip, on behalf of the BBC-VTE Investigators.

**Affiliations**/**Institutions**:

**Conflicts of Interest:**

**Introduction**

Venous thromboembolism (VTE), including deep vein thrombosis (DVT) and pulmonary embolism (PE), is one of the main causes of preventable death in hospitals in the UK. Current clinical risk scores to predict mortality of patients with VTE are the pulmonary embolism severity index (PESI) and the simplified PESI (sPESI) which have similar predictive power1,2.

**Purpose**

To evaluate the ability of machine learning algorithms to predict mortality in patients admitted with VTE and to compare their predictive capability with the sPESI score for 30-day mortality.

**Methods**

The BBC-VTE was a retrospective multicentre patient cohort established to determine clinical features and novel aspects of risk prediction for VTE (and VTE-related complications) in a contemporary cohort.

We include a cohort of 1554 patients (mean age 65.6 years; 53% female) who represent all consecutive admissions with a final diagnosis of VTE to one of 3 regional hospitals in the West Midlands, UK during the years 2012-2014. The dataset was split into training (70%) and validation (30%) cohorts. Using machine-learning, we trained two tree-based models, Random Forests (RF) and XGBoost (XG), using 5-fold cross-validation on the training cohort to predict patient mortality. Clinical variables included age, gender, laboratory blood analysis on admission including c-reactive protein (crp), platelets (plt), neutrophils (neut), white blood cell count (wbc), monocyte count (mono), haemoglobin (hb) and creatinine (cr), discharge oral anticoagulation (OAC), previous malignancies, ethnicity, history of heart failure and chronic lung disease, amongst others. This was validated using the held-out validation cohort and compared to a simple logistic regression model.

To provide a comparison with the sPESI score, we extracted a sub-group of patients (n=652) who had values for oxygen saturation, systolic blood pressure, heart rate, history of cancer, history of cardiopulmonary disease, and age. We used RF to determine the mortality prediction using: i) only the sPESI variables listed and; ii) all the clinical variables available to us. This was then compared against the standard sPESI prediction for this cohort. C-indices (AUC) were used for comparison.

**Results**

The c-indices for RF and XG using the full patient cohort were 0.85 [95% CI: 0.80 – 0.90] and 0.82 [95% CI: 0.77 - 0.87], with the logistic regression c-index being 0.83 [95% CI: 0.78 – 0.88]. The reported sPESI c-index was significantly smaller (p < 0.05) than the RF c-index (0.75 [95% CI: 0.69-0.80])2. The most important features for prediction of mortality indicated by the RF algorithm are age, admission blood levels, discharge OAC, and previous malignancy (Fig. 2).

The sPESI score c-index for the subgroup of patients was found to be 0.72. In comparison, using RF with the same variables gives a significantly larger (p < 0.05) c-index of 0.78 [95% CI: 0.73 – 0.83]. When using all clinical variables available the c-index increased to 0.85 [95% CI: 0.80 – 0.90].

Chart, line chart, scatter chart

Description automatically generatedChart

Description automatically generated

**Fig.1** a) ROC curve for random forest model predicting mortality using all variablesin dataset. b) Comparison of two RF models on a patient subgroup, one using all variables (blue), the other using just sPESI variables (orange). The final green ROC curve is the sPESI score ROC curve for this patient sub-group.

Chart, funnel chart

Description automatically generated

**Fig.2** The top 10 Feature importances from the RF model using all variables

**Conclusion**

Application of machine learning using simple clinical variables in hospital settings canimprove prediction of mortality post-VTE event above-and-beyond the current simplified PESI risk score. Prospective study is warranted to validate the algorithm on external datasets and to construct individualised risk predictions, incorporating into hospital decision aids and mobile health technologies.

**References**

1. Aujesky, D. *et al.* Derivation and validation of a prognostic model for pulmonary embolism. *Am. J. Respir. Crit. Care Med.* **172**, 1041–1046 (2005).

2. Jiménez, D. *et al.* Simplification of the pulmonary embolism severity index for prognostication in patients with acute symptomatic pulmonary embolism. *Arch. Intern. Med.* **170**, 1383–1389 (2010).