

Beyond the d-dimer – Machine-learning assisted pre-test probability evaluation in patients with suspected pulmonary embolism and elevated d-dimers

Joshua Gawlitza, Sebastian Ziegelmayr, Heinrike Wilkens, Philippe Jagoda, Paul Raczeck, Arno Buecker, Jonas Stroeder



PII: S0049-3848(21)00394-7

DOI: <https://doi.org/10.1016/j.thromres.2021.07.001>

Reference: TR 8210

To appear in: *Thrombosis Research*

Received date: 21 April 2021

Revised date: 14 June 2021

Accepted date: 1 July 2021

Please cite this article as: J. Gawlitza, S. Ziegelmayr, H. Wilkens, et al., Beyond the d-dimer – Machine-learning assisted pre-test probability evaluation in patients with suspected pulmonary embolism and elevated d-dimers, *Thrombosis Research* (2018), <https://doi.org/10.1016/j.thromres.2021.07.001>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

## Beyond the d-dimer – machine-learning assisted pre-test probability evaluation in patients with suspected pulmonary embolism and elevated d-dimers

Joshua Gawlitza, MD<sup>1</sup>; Sebastian Ziegelmayr, MD<sup>1</sup>; Heinrike Wilkens, MD<sup>2</sup>; Philippe Jagoda, MD<sup>3</sup>;  
Paul Raczeck, MD<sup>3</sup>; Arno Buecker, MD<sup>3</sup>; Jonas Stroeder, MD<sup>3</sup>

<sup>1</sup> Clinic of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, Technical University Munich, Ismaninger Straße 22, 81675 Munich, Germany

<sup>2</sup> Cardiology, Angiology, Pulmonary and Intensive Care, Saarland University Medical Center, Kirrberger Strasse 100, 66424 Homburg, Germany

<sup>3</sup> Clinic for Diagnostic and Interventional Radiology, Saarland University Medical Center, Kirrberger Strasse 100 (Building 41), 66424 Homburg, Germany

## Introduction

Acute pulmonary embolism (PE) is a leading cardiovascular cause of death, resembling a common indication for emergency computed tomography (CT). Nonetheless, in clinical routine most CTs performed for suspicion of PE excluded the suspected diagnosis. As patients with low to intermediate risk for PE are triaged according to the d-dimer, its relatively low specificity and widespread elevation among elderly might be an underlying issue.

Aim of this study was to find potential predictors based on initial emergency blood tests in patients with elevated d-dimers and suspected PE to further increase pre-test probability.

## Methods

In this retrospective study all patients at the local university hospital's emergency room from 2009 to 2019 with suspected PE, emergency blood testing and CT were included. Cluster analysis was performed to separate groups with distinct laboratory parameter profiles and PE frequencies were compared. Machine learning algorithms were trained on the groups to predict individual PE probability based on emergency laboratory parameters.

## Results

Overall, PE frequency among the 2045 analyzed patients was 41%. Three clusters with significant differences ( $p < 0.05$ ) in PE frequency were identified: C1 showed a PE frequency of 43%, C2 40% and C3 33%. Laboratory parameter profiles (e.g. creatinine) differed significantly between clusters ( $p < 0.0001$ ). Both logistic regression and support-vector machines were able to predict clusters with an accuracy of over 90%.

## Discussion

Initial blood parameters seem to enable further differentiation of patients with suspected PE and elevated d-dimers to raise pre-test probability of PE. Machine-learning-based prediction models might help to further narrow down CT indications in the future.

## Introduction

Acute pulmonary embolism is the third leading cardiovascular cause of death and resembles one of the most common indications for emergency chest computed tomography (CT). [1, 2] The incidence of clinically suspected pulmonary embolism is about 0.6/1000 per year and valid diagnosis is important because the mortality is 1 in 4 if left untreated. [3, 4] Nonetheless, in clinical routine most CT scans performed for suspicion of pulmonary embolism are negative. This empiric observation was substantiated through a meta-analysis by Quiroz et al. as only 15 to 38% of the clinically suspected pulmonary embolisms were verified in the consecutive CT. [5] One of the reasons for this relatively low prevalence might be the diagnostical approach to suspected pulmonary embolisms. According to the German S2k guidelines, patients with low to intermediate risk for pulmonary embolism should be triaged according to the d-dimer. [6] In case of elevated d-dimers, a chest CT should be performed to rule out pulmonary embolism. Thereby, although the actual quality of its negative predictive value is currently under further investigation, d-dimer has become the leading laboratory test in initial probability evaluation of pulmonary embolism in emergency care. [7, 8] As an elevated d-dimer can have several causes beyond pulmonary embolism, it is leading to a reduced pre-test specificity. [9, 10] Given the low positive predictive value of clinical assessment including d-dimer, further investigation of the pre-diagnosical emergency tests (including blood tests) in patients with suspected pulmonary embolism and elevated d-dimers might be helpful.

Aim of this study was to find potential predictors based on the initial emergency blood tests in patients with elevated d-dimers and suspected pulmonary embolism to further increase the pre-test probability prior to CT and to identify a possible predictor which is advantageous over the standard course of action .

## Methods

### Study design

In this retrospective study patient data and CTs were included from January 2009 to April 2020. Inclusion criteria were: 1. patient entered the hospital via the emergency room, 2. patient received standard emergency blood testing in the emergency room including d-dimers, 3. patient received a CT for evaluation of pulmonary embolism. Patients entering the hospital through the emergency room are in general triaged into the surgical and non-surgical department (inclusion criteria 1). All patients included were seen by a specialist of internal medicine, making the suspected diagnosis of pulmonary embolism after clinical examination and anamnesis. Further, emergency blood testing was performed, including d-dimers, based on the suspected diagnosis (inclusion criteria 2). After positive d-dimers, the CT indication was set and CT pulmonary angiography performed (inclusion criteria 3). The study design was in accordance with the Declaration of Helsinki and was approved by the local ethics committee (blinded for review) of our institution (blinded for review).

### Data analysis

#### Profound data

As found in the database query, emergency blood testing consists of 38 different laboratory parameters: alpha-amylase, alkaline phosphatase, aspartate aminotransferase (ASAT), bilirubine, calcium (Ca), cholinesterase (CHE), chloride, creatinine, creatine kinase (CK), creatine kinase myocardial band (CK-MB), c-reactive protein (CRP), D-Dimer, fibrinogen, protein, glomerular filtration rate (GFR), gamma-glutamyltransferase (GGT), glucose, hemoglobin (HB) hematocrit (Hct), international normalized ratio (INR), potassium (K), lactate dehydrogenase (LDH), leukocytes, lipase, mean corpuscular hemoglobin concentration (MCHC), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV), sodium (Na), platelets, mean platelet volume (MPV), partial

thromboplastin time (PTT), Quick, erythrocytes, red blood cell distribution width (RDW), troponin, thyroid-stimulating hormone (TSH) thrombin time and urea.

The performed CT was used as the gold standard for evaluation of pulmonary embolism. In the current evaluation there was no differentiation between central and peripheral embolism. All CTs were evaluated using the four-eye principle by a radiological resident and specialist (> 10 years of experience). All data was rendered anonymous prior to collection in a database, including blood testing and CT results.

#### Statistical evaluation

Data analysis was performed using R, python and JMP 14 (SAS, Cary, USA). [11, 12] At first, differences in laboratory parameters between patients with and without pulmonary embolism were calculated. First all parameters were tested for normal distribution. As a preselected group was examined, almost no normal distribution was found, resulting in the choice of Kruskal-Wallis-test for further testing. In a next step, the data was curated for the cluster analysis. Only parameters with significant ( $p < 0.05$ ) differences between patients with and without pulmonary embolism were included into the cluster analysis after incrementally testing cluster performance with more variables. As a ten-year time frame was covered in this study, not all parameters have been standards in the emergency blood testing yet (CK-MB) or were determined over the years (CHE). Therefore, only parameters measured by 75% of the cohort were included into the cluster analysis. After parameter selection, data was pre-processed by removing all patients with missing laboratory parameters to reach 0% missing data for cluster building. Data was normalized prior to k-means clustering. K-means++ was used for initialization and evaluated cluster range was between 2 and 8. [13] The silhouette coefficient was used as a measurement to determine the best cluster count for optimal differentiation. [14] Individual clusters were compared in regard to the defining parameters using Kruskal-Wallis-Test and post-hoc Dunn test after testing for unequal variance by Levene test.

After cluster evaluation, three supervised learning algorithms were used for cluster prediction evaluation: logistic regression, AdaBoost, random forest and a support-vector machine (SVM). Parameters were adopted to predictive performance and ended up as following. AdaBoost was using 100 estimators at a learning rate of 1 and a fixed seed for the random generator. SAMME.R was used as classifier and a square loss function. The random forest used 13 trees with 5 attributes considered at each split. The tree depth was not limited but no smaller subsets than 5 were allowed prior to splitting. The SVM used a cost of 1 and an epsilon of 0.1. A sigmoid (tanh) kernel function was used. The iteration limit was 100 at a numerical tolerance of 0.001. Leave-one-out cross-validation was used for model evaluation.

## Results

### Patient data

After applying the inclusion criteria, the initial study cohort consisted of 2045 patients. 832 of these (40.7%) actually showed a pulmonary embolism in the correlating CT. As mentioned above, the long timespan and consecutive changes in the routine laboratory as well as the large cohort with unavoidable data gaps lead to overall 8.5% missing laboratory parameters.

### Differences between patients with and without pulmonary embolism

In relation to the individual parameters, 16 of the 38 parameters showed significant differences between patients with and without pulmonary embolism. (Table 1) Highly significant differences ( $p = <0.001$ ) were found for D-Dimer, GFR, INR, Creatinin, Quick and urea. In contrast to our anticipation, no significant differences were found for leukocytes and troponin.

Table 1: differences in parameters between patients with and without pulmonary embolism

Parameter	n with parameter	Normal values	no embolism	embolism	p-value
alpha-Amylase	1150	13-53 U/l	62 ± 100	56 ± 35	0.1358
AP	1975	40-129 U/l	97 ± 102	94 ± 68	0.3952
ASAT	1901	10-50 U/l	47 ± 108	39 ± 90	0.0835
Bilirubin	1986	<1.2 mg/dl	0.66 ± 0.66	0.61 ± 0.56	0.1092
Calcium	1995	2.2-2.6 mmol/l	2.3 ± 0.2	2.3 ± 0.2	0.0532
Cholinesterase	1154	5.32-12.92 kU/l	6.6 ± 2	7 ± 2	0.0203*
CK	2032	0-190 U/l	154 ± 400	121 ± 220	0.0177*
Chlorid	1116	96-110 mmol/l	101 ± 11	101 ± 7	0.1443
CK-MB	1613	0-24 U/l	26 ± 38	27 ± 94	0.6194
CRP	2018	<5 mg/l	46 ± 72	39 ± 57	0.0145*
D-Dimer	1810	<0.5 mg/l	3.8 ± 4.8	5.4 ± 6	<0.0001*
Fibrinogen	1506	180-350 mg/dl	427 ± 181	400 ± 161	0.0032*
Protein	1827	66-87 g/l	70 ± 9	70 ± 7	0.1768
GFR	1736	>60 ml/min/l	67 ± 30	72 ± 27	0.0001*
gamma-GT	1994	<60 U/l	82 ± 160	73 ± 116	0.1262
Glucose	1633	<100 mg/dl	139 ± 57	138 ± 58	0.7338
HB	2024	14-18 g/dl	12.9 ± 2	13.3 ± 2	0.0019*
Hämatokrit	2023	41-50 %	37.9 ± 7	38.7 ± 6	0.0099*
INR	2033	0.85-1.15	1.1 ± 0.4	1 ± 0.3	<0.0001*
Potassium	2034	3.8-5.2 mmol/l	4.1 ± 0.7	4.2 ± 0.6	0.5789
Creatinin	2030	0.7-1.2 mg/dl	1.1 ± 0.6	1 ± 0.5	0.0009*
LDH	1998	0-242 U/l	199 ± 44	174 ± 154	0.081
Lipase	1969	13-50 U/l	38.9 ± 58	37.6 ± 74	0.6593
MCHC	2027	32-37 g/dl	33.7 ± 3.5	34 ± 3.2	0.068
MCH	2027	27-33 pg	29.7 ± 3.8	29.7 ± 3.5	0.8795
MCV	2027	80-99 fl	87.2 ± 11	86.9 ± 9	0.4267
MPV	1981	7.8-11 fl	10.1 ± 0.9	10 ± 0.9	0.0452*
Sodium	2014	135-145 mmol/l	137 ± 16	139 ± 9	0.0031*
Thrombocytes	2027	140-400 *10 <sup>3</sup> /μl	234 ± 120	239 ± 91	0.4718
PTT	1963	21-34 s	39 ± 34	38 ± 34	0.388
Quick	1974	70-130 %	85 ± 20	90 ± 18	<0.0001*
Erythrocytes	2027	4.5 – 5.9 *10 <sup>7</sup> /μl	4.3 ± 0.8	4.4 ± 0.7	0.0012*
RDW	2027	11.5-14.5 %	14.4 ± 2	14 ± 2	0.0056*
Troponin	1816	<14 pg/ml	42 ± 119	36 ± 81	0.1908
TSH	1439	0.27-4.2 mIU/ml	2 ± 1.9	2 ± 2.6	0.5565
Thrombin time	982	14-21 s	18.8 ± 7	18.4 ± 6	0.4439
Urea	2002	17-48 mg/dl	44 ± 27	39.7 ± 20	<0.0001*
Leukocytes	2027	3.9-10.2 *10 <sup>3</sup> /μl	10.1 ± 6	10.1 ± 5	0.9794

---

AP: alkaline phosphatase; ASAT: aspartate aminotranspharase; CK: creatin kinase; CK-MB: creatin kinase myocardial-band; CRP: c-reactive protein; GFR: glumeral filtration rate; HB: hemoglobin; LDH: lactate dehydrogenase; MCHC: mean corpuscular hemoglobin concentration; MCV: mean corpuscular hemoglobin; MCV: mean corpuscular volume; MPV: mean platelet volume; PTT: partial thrombin time; RDW: red blood cell distribution; TSH: thyroid-stimulating hormone  
 The p-value is given for the individual Kruskal-Wallis-test. Significant values ( $p = <0.05$ ) are marked (\*).

---

## Clustering

After preprocessing the data, a total of 1146 patients were included for the cluster analysis. Following parameters were included for clustering, according to the above-mentioned criteria: CK, CRP, Creatinin, D-Dimer, erythrocytes, GFR, Glucose, HB, Hct, INR, MPV, RDW, Sodium, Quick and urea. With a silhouette score of 0.179 three clusters were chosen according to k-means. Cluster 1 consisted of 649 patients, cluster 2 of 325 and cluster 3 of 172 patients. Cluster 3 had with 33% significantly less patients with pulmonary embolism when compared to Cluster 1 (43%) and Cluster 2 (40%; Chi-squared p-value =  $<0.05$ ).

When looking at the individual parameters, Kruskal-Wallis-Test showed highly significant differences for all parameters between the clusters. (Table 2) Except for CRP and Sodium between Cluster 2 and 3 as well as CK between Cluster 1 and 3 post-hoc testing showed highly significant differences for all parameters between the individual clusters. Patients of cluster 3, the group with the lowest percentage of pulmonary embolism, showed higher creatinine, d-dimer, glucose, INR, CRP and urea when compared to the patients in the other clusters. Meanwhile, GFR and Quick were the lowest of all groups in cluster 3, suggesting in synopsis with the elevated parameters a hepatorenal or inflammatory issue in these patients. In contrast, the cluster with the highest percentage of pulmonary embolism (cluster 1), showed the highest GFR and lowest CRP of all clusters. Quick value was significantly higher in Cluster 1 and 2 when compared to cluster 3, suggesting a pro-coagulatory environment in cluster 2 and 3. The parameter-based separation of the different clusters is further visualized in Figure 1.

---

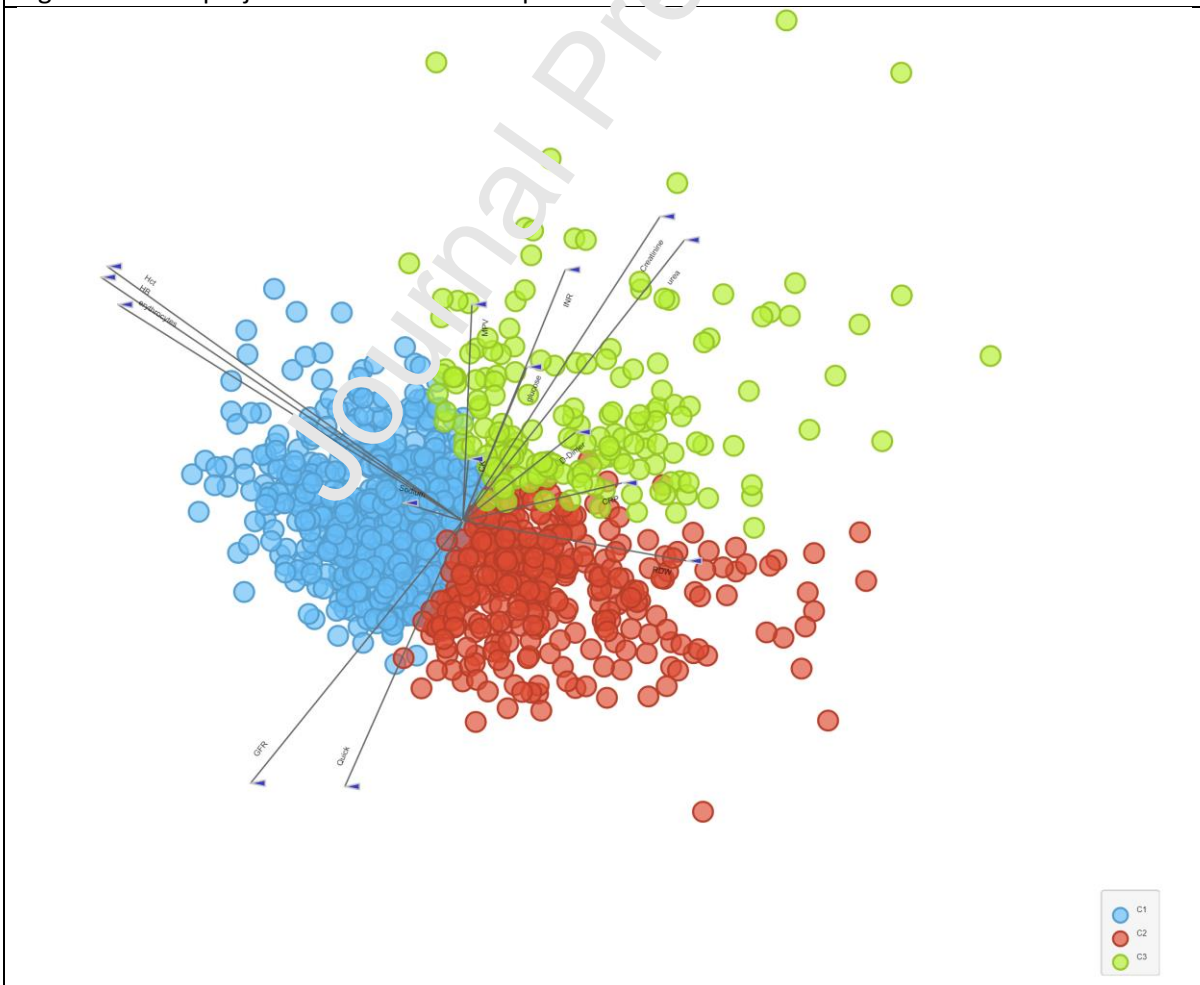
Table 2: parameter differences between clusters

---

	Cluster 1	Cluster 2	Cluster 3	p-value
% patients with embolism	43	40	33	
CRP	27 ± 44	54 ± 70	75 ± 96	<0.0001
CK	134 ± 241	108 ± 204	182 ± 556	<0.0001
Creatinine	0.9 ± 0.2	1 ± 0.3	1.7 ± 1	<0.0001
D-Dimer	3.6 ± 4.7	4.6 ± 5	7.4 ± 7.2	<0.0001
Erythrocytes	4.8 ± 0.4	3.8 ± 0.6	4.2 ± 0.6	<0.0001
Glucose	130 ± 42	127 ± 44	188 ± 88	<0.0001
GFR	77 ± 24	70 ± 30	38 ± 19	<0.0001
HB	15 ± 1	11 ± 1	13 ± 2	<0.0001
Hct	42 ± 3	33 ± 4	37 ± 5	<0.0001
INR	1 ± 0.1	1 ± 0.2	1.3 ± 0.5	<0.0001
MPV	10 ± 0.8	10 ± 0.9	11 ± 1	<0.0001
Quick	93 ± 15	89 ± 16	74 ± 24	<0.0001
RDW	14 ± 1	15 ± 2	15 ± 2	<0.0001
Sodium	140 ± 4	136 ± 19	138 ± 4	<0.0001
Urea	33 ± 12	37 ± 15	73 ± 32	<0.0001

Shown are the parameters' mean values for each cluster with standard deviation as well as the result of the Kruskal-Wallis-Test

Figure 1: linear projection of clusters over parameters

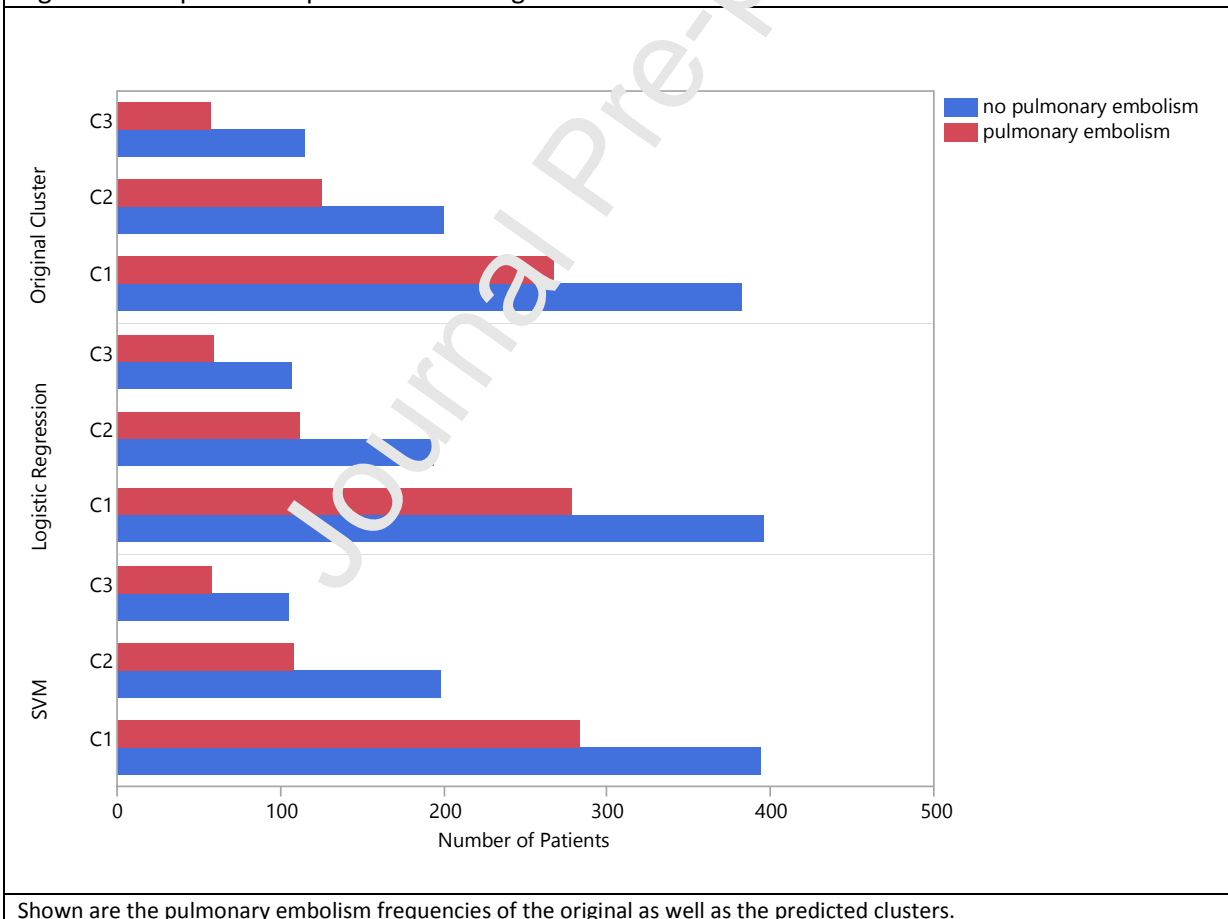


As shown in the linear projection graph, the three clusters show a distinct distribution in line with the profound laboratory parameters. The length of the arrows thereby correlates with the explained variance of the parameter.

## Cluster prediction

When looking at the prediction models, logistic regression and SVM showed best performances with AUCs of 0.974 and 0.945, respectively. The random forest was able to reach an AUC of 0.947 while the AUC of AdaBoost fell below 0.9. In regard to precision accuracy only logistic regression and SVM reached values above 90%. Figure 2 shows the predicted clusters and the correlating pulmonary embolism frequencies of the predicted SVM and logistic regression models in comparison to the original clusters. As shown, there were no significant differences in pulmonary embolism frequencies when comparing the original as well as the predicted clusters

Figure 2: comparison of predicted and original clusters



## Discussion

Aim of this study was to improve the pre-test probability in patients with suspected pulmonary embolism based on the initial, emergency blood testing and find potential predictors for pulmonary embolism. As initially shown, there are significant differences between patients with and without pulmonary embolism in regard to the laboratory parameters. Beyond the coagulation parameters especially retention parameters (creatinine, urea, GFR) showed highly significant differences, suggesting a differentiation between pulmonary embolism and metabolic/renal disease. Based on these findings, we were able to identify three clusters with significant differences in pulmonary embolism frequencies. Cluster 3 showed 30% fewer pulmonary embolism cases compared to cluster 1 and 23% when compared to the initial study cohort average. This finding could be especially interesting in the initial probability evaluation of patients with suspected pulmonary embolism. Similar to the initial comparison of patients with and without pulmonary embolism, patients of cluster 3 showed significantly higher levels of creatinine and urea. This suggests a higher frequency of hepatorenal pathologies when compared to the other clusters. Interestingly, this cluster showed significantly higher d-dimer levels as well, although the fewest cases of pulmonary embolism were found for the group. But as reported before, especially in patients with diabetic nephropathy, d-dimer levels can rise significantly. [15] This is in contrast to previous findings, which suggest the level of d-dimer elevation correlates with the probability and severity of pulmonary embolism. [16] On the other hand, Masotti et al. showed, that especially in elderly (>80 years) the clinical relevance of the d-dimer is decreasing significantly as other reasons for the elevation become more dominant. [17] When looking at the elevated retention parameters as well as the significantly higher glucose level in cluster 3 patients, this could be the reason for the higher d-dimers despite fewer embolism cases. A disrupted water balance might mimic the pulmonary symptoms. These indirect signs of metabolic disease might be a reason for the overall higher INR and lower Quick in cluster 3, respectively. Either patients had a hepatic disease, limiting synthesis of coagulation factors or they have been on anticoagulatory drugs in the first place, limiting the probability for pulmonary embolism in contrast to the other two clusters.

As shown, it is possible to train models on the cluster prediction and thereby quickly categorize patients based on the initial emergency blood testing. This could be a crucial step into improvement of pulmonary embolism pre-test probability evaluation. Nonetheless, our planned prospective study needs to show the real-world practicability and performance of these cluster predictions. Further it should be noted, that the clinical pre-test probability is crucial for specific pulmonary embolism diagnostic. [18] The implementation of clinical parameters might further improve our trained models. Although clustering would most likely have benefitted from demographic as well as physiological parameters, there were two major limitations in this approach. First, we wanted to focus directly on the d-dimer's junction between clinical probability and imaging indication. This decision in patients with low to intermediate risk is currently solely dependent on d-dimer and thereby a laboratory parameter. As laboratory parameters are usually easily accessible through the clinical information system, this would allow an easy implementation of the models in the future or at other sites while staying in the same information regime (laboratory testing) as the current evaluation parameter (d-dimer). The accessibility is another limitation in regard of the demographic/physiological parameters. Only in recent years, the emergency department has implemented an electronical structured recording sheet, making the automated document crawling for e.g. heart rate/blood pressure impossible for a time-span of ten years. Most previous studies have focused on improving the clinical pre-test probability by different models. [19-21] In contrast to the initially named incidence of pulmonary embolism between 15% and 38%, the present cohort showed a slightly higher incidence of 40.1%. [5]

Our study is subject to several limitations. First, our results are based on a large retrospective cohort, consisting of partially incomplete data over a ten-year time-span. A few patients were missing d-dimers for example, due to technical errors in the blood sampling/testing (e.g. hemolytic probe, not enough blood for testing) but received CT nonetheless due to clinical likelihood of pulmonary embolism. Although patients with missing data were excluded, this profound dataset is more challenging for creation of general models and predictions. Nonetheless, we believe that a large

dataset like this contains valuable information and is unique in its kind. Especially, as the current cohort mimics the potential workflow of the models shown, regarding all patients came into the ER with suspected PE. Although imputation would be favorable over exclusion of data, the inconsistency of the retrospective data set complicated this as over the years some parameters have been omitted from or added to the initial blood testing in the ED - e.g. cholinesterase was dropped around mid-way of the observational period. Thereby, not only small gaps but half the cohort size would have been imputed data, making imputation impractical. [22] Another limitation are the clusters found and their separation score. Although all three clusters show significant and pathophysiological discrete properties, the silhouette score was relatively low, suggesting blurry borders between the clusters with corner cases, fitting into two separate clusters. A third limitation lies within the trained models and the probability of overfitting. Although, we used cross-validation and reduced the number of initial parameters for cluster evaluation and consecutive model training, we cannot rule out a slight overfit of our models. Again, prospective evaluation, possibly at more than one centre is needed to evaluate their performance. Although the models are currently not influencing medical decision in patients with suspected PE, the prediction models are running, still within a study setting, aside the clinical routine, evaluating their real-world performance prospectively. In case of positive results in the prospective analysis, a second evaluation step in patients with elevated d-dimer suspected PE and low to intermediate risk would be imaginable for further triaging in the clinical workflow - e.g. patients with high pre-test probability according to the models receive immediate imaging (urgent CT indication) while patients with low pre-test probability have only a "time-sensitive" CT indication. Another limitation of the current models is the preselection due to exclusion of patients with missing d-dimers. Following the current guidelines, patients with hemodynamic instability or high clinical probability for PE should directly undergo imaging. [6] Keeping this in mind, the current models are aimed towards patients with low to intermediate probability for PE, exactly like the d-dimer itself. Nonetheless, the currently running prospective evaluation should identify the model performance in patients with high clinical probability of PE as well.

In conclusion, our results provide important evidence, that initial evaluation based on the emergency blood testing alone can significantly improve the precision of pre-test probability for patients with suspected pulmonary embolism and elevated d-dimers.

## References

- [1] K. Keller, L. Hobohm, M. Ebner, K.-P. Kresoja, T. Münzel, S.V. Konstantinides, M. Lankeit, Trends in thrombolytic treatment and outcomes of acute pulmonary embolism in Germany, *European Heart Journal* 41(4) (2020) 522-529.
- [2] J. Seidel, M.B. Bissell, S. Vatturi, A. Hartery, Retrospective analysis of emergency computed tomography imaging utilization at an academic centre: an analysis of clinical indications and outcomes, *Canadian Association of Radiologists Journal* 70(1) (2019) 13-22.
- [3] T. Andersson, S. Söderberg, Incidence of acute pulmonary embolism, related comorbidities and survival; analysis of a Swedish national cohort, *BMC cardiovascular disorders* 17(1) (2017) 1-7.
- [4] P. Egermayer, G. Town, The clinical significance of pulmonary embolism: uncertainties and implications for treatment—a debate, *Journal of internal medicine* 241(1) (1997) 5-10.
- [5] R. Quiroz, N. Kucher, K.H. Zou, F. Kipfmueller, P. Loscallo, S.Z. Goldhaber, U.J. Schoepf, Clinical validity of a negative computed tomography scan in patients with suspected pulmonary embolism: a systematic review, *Jama* 293(16) (2005) 2012-2017.
- [6] S.V. Konstantinides, G. Meyer, C. Becattini, H. Bueno, G.-J. Geersing, V.-P. Harjola, M.V. Huisman, M. Humbert, C.S. Jennings, D. Jiménez, 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS) The Task Force for the diagnosis and management of acute pulmonary embolism of the European Society of Cardiology (ESC), *European heart journal* 41(4) (2020) 543-603.
- [7] E. Fuchs, S. Asakly, A. Karban, I. Tzorani, Age-adjusted cutoff d-dimer level to rule out acute pulmonary embolism: a validation cohort study, *The American journal of medicine* 129(8) (2016) 872-878.
- [8] N. Van Es, T. Van der Hulle, M. Püller, F. Klok, M. Huisman, J. Galipienzo, M. Di Nisio, Is stand-alone D-dimer testing safe to rule out acute pulmonary embolism?, *Journal of Thrombosis and Haemostasis* 15(2) (2017) 323-328.
- [9] A. Dompnmartin, F. Ballieux, P. Thibon, A. Lequerrec, C. Hermans, P. Clapuyt, M.-T. Barrellier, F. Hammer, D. Labbé, M. Vikkula, Elevated D-dimer level in the differential diagnosis of venous malformations, *Archives of dermatology* 145(11) (2009) 1239-1244.
- [10] G. Lippi, L. Bonfanti, C. Saccenti, G. Cervellin, Causes of elevated D-dimer in patients admitted to a large urban emergency department, *European journal of internal medicine* 25(1) (2014) 45-48.
- [11] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevár, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, Orange: data mining toolbox in Python, *the Journal of machine Learning research* 14(1) (2013) 2349-2353.
- [12] R.C. Team, R: A language and environment for statistical computing, (2013).
- [13] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, 1027–1035, *Society for Industrial and Applied Mathematics* (2007).
- [14] S. Aranganayagi, K. Thangavel, Clustering categorical data using silhouette coefficient as a relocating measure, *International conference on computational intelligence and multimedia applications (ICCIMA 2007)*, IEEE, 2007, pp. 13-17.
- [15] S. Vahdat, S. Shahidi, D-dimer levels in chronic kidney illness: a comprehensive and systematic literature review, *proceedings of the national academy of sciences, india section b: biological sciences* (2020) 1-18.

- [16] W. Ghanima, M. Abdelnoor, L. Holmen, B. Niessen, S. Ross, P. Sandset, D-dimer level is associated with the extent of pulmonary embolism, *Thrombosis research* 120(2) (2007) 281-288.
- [17] L. Masotti, P. Ray, M. Righini, G. Le Gal, F. Antonelli, G. Landini, R. Cappelli, D. Prisco, P. Rottoli, Pulmonary embolism in the elderly: a review on clinical, instrumental and laboratory presentation, *Vascular health and risk management* 4(3) (2008) 629.
- [18] C. Kearon, K. de Wit, S. Parpia, S. Schulman, M. Afilalo, A. Hirsch, F.A. Spencer, S. Sharma, F. D’Aragon, J.-F. Deshaies, Diagnosis of pulmonary embolism with d-dimer adjusted to clinical probability, *New England Journal of Medicine* 381(22) (2019) 2125-2134.
- [19] G. Le Gal, M. Righini, P.-M. Roy, O. Sanchez, D. Aujesky, H. Bounameaux, A. Perrier, Prediction of pulmonary embolism in the emergency department: the revised Geneva score, *Annals of internal medicine* 144(3) (2006) 165-171.
- [20] D. Aujesky, D.S. Obrosky, R.A. Stone, T.E. Auble, A. Perrier, J. Cornuz, P.-M. Roy, M.J. Fine, A prediction rule to identify low-risk patients with pulmonary embolism, *Archives of Internal Medicine* 166(2) (2006) 169-175.
- [21] I. Chagnon, H. Bounameaux, D. Aujesky, P.-M. Roy, A.-L. Gourdie, J. Cornuz, T. Perneger, A. Perrier, Comparison of two clinical prediction rules and implicit assessment among patients with suspected pulmonary embolism, *The American journal of medicine* 115(4) (2002) 269-275.
- [22] J.C. Jakobsen, C. Gluud, J. Wetterslev, P. Winkel, When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts, *BMC medical research methodology* 17(1) (2017) 1-10.

**Declaration of interests**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof

## Highlights

- Emergency laboratory parameters can predict lower embolism probabilities
- 24% fewer embolism cases in one cluster despite elevated d-dimers
- Trained models allow for increased pre-test probability of pulmonary embolism

Journal Pre-proof