Risk assessment for venous thromboembolism in chemotherapy treated ambulatory cancer patients: a machine learning approach[1]

Running title Machine learning for VTE risk prediction

[1]*Patrizia Ferroni, M.D., [2]*Fabio Massimo Zanzotto, Ph.D., [1]Noemi Scarpato, Ph.D., [3,4]Silvia Riondino, M.D., [5]Umberto Nanni, Ph.D., [4]Mario Roselli, M.D., [1,3]Fiorella Guadagni, M.D.

*First authors for equal contribution

[1]San Raffaele Roma Open University, Via di Val Cannuta,247 Rome, Italy
[2]Department of Enterprise Engineering, University of Rome "Tor Vergata", Rome, Italy
[3]BioBIM (InterInstitutional Multidisciplinary Biobank, IRCCS San Raffaele Pisana, Rome, Italy
[4]Department of Systems Medicine, Medical Oncology, University of Rome "Tor Vergata", Rome, Italy
[5]Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University, Rome, Italy

Contact information

Patrizia Ferroni	patrizia.ferroni@sanraffaele.it
Fabio Massimo Zanzotto	fabio.massimo.zanzotto@uniroma2.it
Noemi Scarpato	noemi.scarpato@unisanraffaele.gov.it
Silvia Riondino	silvia.riondino@sanraffaele.it
Umberto Nanni	umberto.nanni@dis.uniroma1.it
Mario Roselli	mario.roselli@uniroma2.it
Fiorella Guadagni	fiorella.guadagni@sanraffaele.it

Corresponding author

Prof. Fiorella Guadagni, San Raffaele Roma Open University, Interinstitutional Multidisciplinary Biobank (BioBIM), SR Research Center, IRCCS San Raffaele Pisana, Via di Val Cannuta, 247, 00166 Rome - Italy
Tel: +39 06 52253733; e-mail: fiorella.guadagni@sanraffaele.it; alternate e-mail: guadagnifiorella@gmail.com

Keywords: Clinical decision support systems, machine-learning, random optimization, venous thromboembolism, cancer.

Word count: 2955

[1]Financial support for this study was provided in part by research funding from the European Social Fund, under the Italian Ministry of Education, University and Research PON03PE_00146_1/10 BIBIOFAR (CUP B88F12000730005). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

ABSTRACT

Objective: To design a precision medicine approach aimed at exploiting significant patterns in data, in order to produce venous thromboembolism (VTE) risk predictors for cancer outpatients that might be of advantage over the currently recommended model (Khorana score).

Design: Multiple kernel learning (MKL) based on support vector machines (SVM) and random optimization (RO) models were used to produce VTE risk predictors [referred as Machine-Learning (ML)-RO] yielding the best classification performance over a training (3-fold cross validation) and testing set.

Results: Attributes of the patient dataset (n=1179) were clustered into 9 groups according to clinical significance. Our analysis produced 6 ML-RO models in the training set, which yielded better likelihood ratios (LRs) than baseline models. Of interest, the most significant LRs were observed in two ML-RO approaches not including the Khorana score (ML-RO-2: +LR=1.68, -LR=0.24; ML-RO-3: +LR=1.64, -LR=0.37). The enhanced performance of ML-RO approaches over the Khorana score was further confirmed by the analysis of the areas under the Precision-Recall curve (AUCPR), which were superior in the ML-RO approaches (best performances: ML-RO-2: AUCPR=0.212; ML-RO-3-K: AUCPR=0.146) compared to the Khorana score (AUCPR=0.096). Of interest, the best fitting model was ML-RO-2, in which blood lipids and body mass index/performance status retained the strongest weights, with a weaker association with tumor site/stage and drugs.

Conclusions: Although the monocentric validation of the presented predictors might represent a limitation, these results demonstrate that a model based on MKL and RO may represent a novel methodological approach to derive VTE risk classifiers. Moreover, this study highlights the advantages of optimizing the relative importance of groups of clinical attributes in the selection of VTE risk predictors.

INTRODUCTION

In recent years, the approach to medicine has been pressured by a growing availability of electronic health records (EHR) and by the consequent demand to provide *precision medicine*, especially in oncology, where the development of targeted therapies might improve treatment delivery and clinical outcome.

A major challenge that oncologists are facing is the risk assessment of chemotherapy-associated venous thromboembolism (VTE), which may result in treatment delays with detrimental effects on disease outcome.[1] Nonetheless, all current consensus guidelines do not recommend routine prophylaxis for the primary prevention of VTE in chemotherapy-treated cancer out-patients[2,3] although "*it may be considered for selected high-risk patients*".[3] Thus, selecting patients for prophylactic anticoagulation is perceived as a growing necessity in cancer patient management, fostering the demand for risk assessment models.

However, predicting VTE risk for cancer patients is a compelling challenge where precision medicine can play a crucial role, as VTE risk differs not only among patients, but even in the same patient over the course of cancer natural history.[4,5,6] In 2008, Khorana and colleagues proposed a VTE risk assessment model that uses a combination of routinely available variables.[7] To date, the Khorana score (KS) is the sole model available for VTE prediction in chemotherapy-treated cancer outpatients. Hence, it has been proposed in recent guidance statements.[8] Nonetheless, although validated by some independent groups,[9,10] others did not,[11,12,13] as the KS fails to classify >50% of patients (intermediate risk), in whom clinical decision making remains challenging. Expanded risk scoring models, including biomarkers,[9] or anti-cancer drug,[14] were proposed to implement KS, but VTE risk prediction for chemotherapy-treated cancer outpatients remains sub-optimal.

A solid base on which to build a precision medicine tool in oncology is represented by Machine-Learning (ML),[15-19] which can derive patterns in clinical and biochemical knowledge[20] and has been previously applied to learn VTE risk predictors in the general population.[21]

Therefore, aim of the present study was to analyze the performance of a multiple kernel machine learning (MKL) model that combines support vector machines (SVM),[22,23] and *random optimization* (RO)[24] to produce

VTE risk predictors in a population of consecutive ambulatory cancer patients representative of a general practice cohort. These predictors exploit significant patterns in data – connoting causality between individual features and VTE – and can be used in the development of a clinical decision support system for VTE risk stratification prior to chemotherapy start.

METHODS

Learning VTE Risk Predictors within a Precision Medicine Approach

To deal with heterogeneity of clinical attributes, our methodology is based on a MKL model[25,26] that combines SVM[22] to learn classifiers and *RO*[24] to devise relative importance of different groups of clinical attributes in final predictions.

$$f(x) = sgn\left(\sum_{i=1}^{N} \alpha_i \langle \overrightarrow{w_i}, \overrightarrow{g_i} \rangle + b\right) \quad (1)$$

Based on MKL our VTE risk predictors are binary classifiers that have to determine whether patients x will have an high-risk to develop a VTE event in the future (*y=1*) or not (*y=-1*). In Equation 1, $sgn(z)$ is the sign function that is 1 if $z > 0$ and -1 if $z < 0$, patients x are represented with their clinical attributes $\vec{x} = (x_1, \dots, x_n)$ divided in groups $\vec{x} = [\vec{g}_1, \vec{g}_2, \dots, \vec{g}_N]$, $\overrightarrow{w_i}$ are the decision hyperplanes for each group of attributes, and α_i are the relative weights of the groups of attributes $\vec{g}_i$. Using α_i, these VTE risk predictors take into account the heterogeneity of clinical attributes.

VTE risk predictors are learned with an n-fold cross validation on a *training set* that allows to derive parameters α_i with RO by optimizing the F-measure of classifiers $f(x)$ whose decision hyperplanes $\overrightarrow{w_i}$ are learned with SVM. F-measure is defined as:

$$F\text{-}measure = \frac{2PR}{P + R}$$

that is a harmonic mean of *positive predictive value* (*PPV)* and *sensitivity*, which are called Precision (P) and Recall (R) in ML. As RO depends on the initial seed, we run the learning multiple times. Predictors are then sorted according to their decreasing F-measure on the training set.

To assess their validity, learned VTE risk predictors are evaluated on a separated *testing set*.

Our method to find the best VTE risk predictors has two major benefits: first, it selects the best predictors on training data; second, it determines relative weights α_i among groups of clinical attributes. These weights give useful insights on how predictors take their decisions.

Patient dataset for VTE risk assessment

Patient dataset was attained by joint efforts between the PTV Bio.Ca.Re. (Policlinico Tor Vergata Biospecimen Cancer Repository) and the BioBIM (InterInstitutional Multidisciplinary Biobank, IRCCS San Raffaele Pisana), and consisted of 1179 consecutive ambulatory cancer patients with primary or relapsing/recurrent solid cancers, who were prospectively followed under the Institutional ethics approval in accordance with the principles embodied in the Declaration of Helsinki. All patients were required to be at the start of a new chemotherapy regimen and no patient received thromboprophylaxis. Eligibility criteria are detailed in Supplementary Table 1. Clinical characteristics and laboratory attributes of patients are summarized in Supplementary Table 2.

All patients received specific anti-cancer treatment, with or without supportive care agents, according to guidelines for cancer treatment by site. All patients were regularly seen at the Medical Oncology ward of the Department of Systems Medicine, PTV, at time of scheduled chemotherapy visits, or at the occurrence of clinically suspected VTE. Deep venous thrombosis (DVT) or pulmonary embolism (PE) were diagnosed as previously reported.[6] During a 1-year median follow-up, VTE occurred in 8% (29 PE and 65 DVT) of patients (median time-to-event 3 months). Thirty-four (2.9%) patients had a previous history of VTE, and 5 (0.4%) had concurrent DVT on the first week of treatment. Forty-one of 94 events were incidentally diagnosed (16 PE and 25 DVT) at time of restaging. Competing mortality at 6 months was approximately 2%, and 25 patients without VTE died of their disease during this timeframe.

Experimental settings

To test our methodology and default methods, the patient dataset was used as follows: 1) clinical attributes were clustered in 9 groups; 2) the patient dataset was randomly divided in training and testing set, 3) values x

of continuous clinical attributes c were rescaled with functions $f(x) = -0.5 + (x - m_c)/(M_c - m_c)$ where m_c and M_c are the minimal and the maximal values of c in the training set; 4) missing clinical attribute values were treated according to (Predictive) Value Imputation (PVI) method by replacing missing values with the average of the attribute observed in the training set.[27]

Group clustering was performed according to the clinical significance of the attributes included in the patient dataset. In particular, demographic variables and tumor site/stage were individually considered given their importance as risk factors for VTE.[5,28] Hematological attributes, including blood cell counts,[8,29] and neutrophil and platelet to lymphocytes ratios,[30] were grouped together. Similarly, individual attributes concerning fasting blood lipids, glycemic indexes and liver and kidney function were clustered within three individual groups. Body mass index (BMI) and Eastern Cooperative Oncology Group Performance Status (ECOG-PS) were considered within the same group. Supportive and anti-cancer drugs were collectively considered under the definition of "drugs". Details on groups of clinical attributes are reported in Figure 1.

To learn our VTE risk predictors, the patient dataset was randomly divided in two needed sets:

1) training set: 70% of the cases were used to learn risk predictors with SVM and to optimize the parameters α_i with RO with a 3-fold cross validation (see Equation 1)
2) testing set: 30% of the cases were used to test the learned risk predictors

We performed 5 different learning sessions on the *training set* with 5 different RO initializations (Table 2). The final performance was then evaluated on the separated *testing set* (see Table 3). Experiments were performed including or not the KS.

Statistical analysis

Machine Learning used for the primary analysis was run on KELP.[31] Bayesian analysis was performed, and positive (+LR) and negative (-LR) likelihood ratios were used to estimate the probability of having or not VTE, using a free web-based application (http://statpages.org/). Time-to-event was calculated from the enrolment date until VTE or the most recent follow-up visit. VTE-free survival curves were calculated by the Kaplan–Meier

method and the significance level was assessed by log-rank test using a computer software package (Statistica 8.0, StatSoft Inc., Tulsa, OK). For administrative censoring VTE was considered to be an event if occurring during chemotherapy administration, but not subsequent follow-up.

This study had no external funding source.

RESULTS

The weights α_i of groups of clinical attributes for the ROs models are reported in Table 1. Tables 2 and 3 summarize the results achieved using the risk predictors selected on the training and testing sets out of 5 runs obtained with RO using (*ML-RO-1-K through ML-RO-5-K*), or not (*ML-RO-1 through ML-RO-5*) the KS, and 4 different baseline models: 1) *Khorana k≥3*: pure KS with cutoff at 3;[8] 2) *Khorana-ML*: a SVM VTE event predictor trained with a polynomial kernel of degree 2 that uses only the KS as feature; 3) *Basic-ML-K*; 4) *Basic-ML.* The two latter predictors are SVM VTE predictors where each group of clinical attributes has the same weight: *Basic-ML-K* uses KS and *Basic-ML* does not use it.

As shown in Table 2, a ML approach with RO was capable of improving VTE risk prediction compared to *Khorana k≥3* or Khorana-ML as demonstrated by a substantial improvement of the f-measure, translating in comparable precision (or positive predictive value – PPV) and considerably higher recall (or sensitivity) values.

To better characterize the performance of the proposed method, +LR and -LR were calculated for all ML-RO model in comparison with *Khorana k≥3* or Khorana-ML. As shown in Table 2, the LRs achieved using the KS (with or without a ML approach) were not significant in terms of VTE risk prediction. Conversely, all ML-RO models including the KS resulted in an overall improvement of the LRs for VTE risk prediction, whereas the ML-RO approaches, not including the KS, yielded significant results in ML-RO-1 ($p<0.0001$) and ML-RO-3 ($p=0.015$), and ML-RO-4 ($p=0.007$), but not in the other ML-ROs (Table 2).

When the algorithm was applied to the testing set, among all ML models including the KS the best fitting model was represented by *ML-RO-3-K* (+LR=1.52, -LR=0.55; $p=0.017$) (Table 3). On the other hand, the ML approach not including the KS yielded the best results in *ML-RO-2* (ML-RO-2: +LR=1.68, -LR=0.24; $p<0.0001$).

Table 1: Weights α_i of groups of clinical attributes for the different models

Method	Sex	Age	Tumor site & stage	BMI & ECOG	Hematology	Liver & kidney function	Glycemic asset	Blood lipid pattern	Drugs	Khorana Score
Khorana-ML	0	0	0	0	0	0	0	0	0	1
Basic-ML-K	1	1	1	1	1	1	1	1	1	1
ML-RO-1-K	0.0963	0.0604	0.2218	0.9787	0.1161	0.0117	0.2334	0.0543	0.6735	0.0267
ML-RO-2-K	0.0205	0.0304	0.8914	0.0577	0.0684	0.0256	0.0136	0.6652	0.1003	0.0000
ML-RO-3-K	0.0581	0.0190	0.2437	1.2319	0.2636	0.2253	0.1265	0.3052	0.0523	0.0596
ML-RO-4-K	0.1905	0.0048	0.9769	0.0116	0.4515	0.5668	0.4381	0.0619	0.0081	0.0906
ML-RO-5-K	0.0962	0.1390	0.5158	0.0326	0.0148	0.4563	0.1417	0.0823	0.4154	0.0401
Basic-ML	1	1	1	1	1	1	1	1	1	0
ML-RO-1	0.0170	0.0035	0.1157	0.0538	0.0025	0.2511	0.7096	0.0046	0.1891	0
ML-RO-2	0.1241	0.1144	0.3129	0.7672	0.0973	0.1420	0.0488	1.0548	0.2636	0
ML-RO-3	0.1253	0.7654	0.2521	0.1808	0.0149	0.0616	0.0000	0.6499	0.3054	0
ML-RO-4	0.4300	0.0023	0.0924	0.5372	0.1270	0.3742	0.3055	0.3222	0.3853	0
ML-RO-5	0.5643	0.1205	0.1957	0.4045	0.0539	2.62E-06	0.9197	0.0057	0.1174	0

Table 2: Results of basic predictors and predictors based on machine-learning with random optimization – Results on training set – ML-RO models are ranked according to F-Measure.

Method	Precision (PPV)	Recall (Sensitivity)	F-measure	+LR (95%CI)	-LR (95%CI)
Khorana (k>=3)*	0.122	0.075	0.093	1.59 (0.56-3.99)	0.97 (0.87-1.02)
Khorana-ML	0.122	0.075	0.093	1.57 (0.55-3.95)	0.97 (0.88-1.02)
Basic-ML-K	0.096	0.642	0.167	1.21 (1.00-1.39)	0.68 (0.43-1.00)
ML-RO-1-K	0.126	0.761	0.217	1.57 (1.27-1.84)	0.54 (0.35-0.77)
ML-RO-2-K	0.119	0.791	0.207	1.45 (1.22-1.62)	0.46 (0.27-0.73)
ML-RO-3-K	0.115	0.687	0.197	1.40 (1.10-1.69)	0.69 (0.48-0.92)
ML-RO-4-K	0.110	0.776	0.192	1.33 (1.05-1.59)	0.71 (0.49-0.96)
ML-RO-5-K	0.107	0.776	0.189	1.25 (1.03-1.44)	0.65 (0.41-0.96)
Basic-ML	0.091	0.537	0.155	1.24 (0.99-1.48)	0.74 (0.51-1.02)
ML-RO-1	0.117	0.716	0.202	1.58 (1.30-1.80)	0.48 (0.30-0.72)
ML-RO-2	0.115	0.731	0.198	1.15 (0.89-1.41)	0.85 (0.60-1.12)
ML-RO-3	0.115	0.702	0.197	1.33 (1.05-1.58)	0.69 (0.48-0.95)
ML-RO-4	0.114	0.627	0.193	1.50 (1.19-1.77)	0.61 (0.42-0.85)
ML-RO-5	0.111	0.672	0.191	1.09 (0.84-1.33)	0.91 (0.65-1.19)

*Patients with brain cancer (n=5) were excluded from the analysis (Khorana score not applicable)

+LR: positive likelihood ratio

-LR: negative likelihood ratio

Table 3: Results of basic predictors and predictors based on machine-learning with random optimization – Results on testing set – ML-RO models are ranked according to F-measure on the *training* set (see Table 2)

Method	Precision (PPV)	Recall (Sensitivity)	F-measure	+LR (95%CI)	-LR (95%CI)
Khorana (k>=3)*	0.136	0.111	0.122	1.90 (0.46-6.05)	0.94 (0.76-1.04)
Khorana-ML	0.136	0.111	0.122	1.91 (0.46-6.08)	0.94 (0.76-1.04)
Basic-ML-K	0.099	0.852	0.177	1.33 (1.01-1.50)	0.41 (0.13-0.99)
ML-RO-1-K	0.105	0.741	0.184	1.43 (1.01-1.73)	0.54 (0.24-0.99)
ML-RO-2-K	0.100	0.778	0.177	1.35 (0.98-1.60)	0.53 (0.22-1.03)
ML-RO-3-K	0.112	0.704	0.193	1.52 (1.05-1.90)	0.55 (0.27-0.95)
ML-RO-4-K	0.100	0.667	0.174	1.35 (0.91-1.71)	0.66 (0.34-1.09)
ML-RO-5-K	0.096	0.778	0.171	1.29 (0.94-1.53)	0.56 (0.23-1.10)
Basic-ML	0.078	0.593	0.137	1.02 (0.66-1.35)	0.97 (0.44-1.50)
ML-RO-1	0.082	0.556	0.143	1.08 (0.68-1.47)	0.91 (0.53-1.36)
ML-RO-2	0.122	0.889	0.214	1.68 (1.29-1.86)	0.24 (0.06-0.65)
ML-RO-3	0.119	0.815	0.208	1.64 (1.21-1.90)	0.37 (0.14-0.78)
ML-RO-4	0.092	0.593	0.159	1.23 (0.79-1.63)	0.79 (0.44-1.21)
ML-RO-5	0.108	0.741	0.188	1.46 (1.03-1.77)	0.53 (0.24-0.96)

*Patients with brain cancer (n=2) were excluded from the analysis (Khorana score not applicable)

+LR: positive likelihood ratio

-LR: negative likelihood ratio

Finally, the improvement of the ML approach with RO was confirmed by plotting recall vs. precision for the different systems on the test. Figure 2 reports the recall vs. precision curves for the basic and the two best fitting models. As shown, ML-RO-2 was the best predictor with an area under the precision-recall curve (AUCPR) of 0.212).

As the probability of VTE occurrence during chemotherapy is also a function of time (being maximal during the first 6 months of treatment)[5,6] we finally performed a survival analysis by the Kaplan–Meier method with log-rank test. Figure 3 reports the Kaplan–Meier curves for patients in the testing set stratified on the basis of *Khorana k≥3* and the two best fitting ML-RO models. As shown, despite a high precision, the KS used at a cut-off ≥3 points, as currently recommended,[8] resulted in a 6-month VTE-free survival rate not significantly different from that of low-risk patients (Figure 2A). On the other hand, optimizing the relative weight of groups of clinical attributes resulted in a substantial improvement of VTE risk prediction. In particular, patients classified at-risk with ML-RO-2 (Figure 3C) had a significantly lower 6-month VTE-free survival compared to patients classified as low-risk.

DISCUSSION

The present study was designed to investigate the performance of ML as a novel methodological approach to derive a VTE risk classifier in chemotherapy-treated cancer outpatients. In the algorithm here presented, we applied a combined approach of kernel machines and RO of performance of binary classifiers, hypothesizing that this method would have found combination of attributes yielding the best classification performance of our predictors over a testing set. The predictive value of our learned models was also compared with the Khorana's risk assessment tool.

The results obtained demonstrated, for the first time to our knowledge, that this approach can be advantageous in VTE risk assessment and allowed us to draw some interesting considerations.

First, the analysis of clinical/biochemical variables identified several risk factors, not previously included in VTE risk models (i.e. blood lipids or ECOG-PS), as evidenced by attributes' weights (Table 2). Moreover, ML models using all clinical attributes (Basic-ML-K, Basic-ML and ML-ROs) showed better F-measures and LRs than generic models (pure KS and Khorana-ML), as verified on the training and, more importantly, on the testing set. Using additional clinical attributes is thus promising.

Second, ML-ROs, which optimize the relative importance of groups of clinical attributes, appeared extremely useful in selecting better VTE risk predictors. It is obvious that on the training set f-measures of ML-ROs were better than Basic-ML as RO was performed on the training set. It is less obvious that ML-ROs generally outperformed Basic-MLs on the testing set in terms of f-measure.

Most importantly, best scoring models in terms of both f-measure and LRs were also clinically plausible, as demonstrated by the finding that blood lipids and BMI and ECOG-PS retained the strongest weight both in ML-RO-3-K and in ML-RO-2 (Table 2). This is consistent with data showing that HDL-cholesterol[32] and ECOG-PS[28,30] might be good predictors of increased VTE risk in chemotherapy-treated cancer patients. Moreover, the ML-RO-2 model showed a weak association with tumor site and stage, and with drugs, which is not surprising, since these variables have been previously related with increased VTE risk.[7,11] Undeniably, advanced cancer, either locally (regional) or distant,[33] has been considered as a risk factor for VTE, and anti-cancer drugs may act as thrombotic triggers.[6,34] One major criticism raised to KS is that it does not consider

treatment-related risk of VTE, at a point that certain anti-cancer agents have been proposed to be used to implement KS.[11]

Of course, we must acknowledge the low performance of both KS and ML-RO predictors in our model, either in terms of PPV or f-measures. This could be explained by the fact that this kind of dataset is extremely unbalanced. Indeed, VTE occurred only in 8% of the cases (in line with literature), which renders the application of ML models extremely difficult, consistently with Larrañaga et al.[19] Previous studies in general population showed better predictive performance,[21] but the test set used generally consisted of VTE cases paired to non-VTE controls. Our study cohort, instead, consisted of out-patients consecutively enrolled, in whom all VTE events were prospectively recorded during chemotherapy. Moreover, while in hospitalized patients cancer is connoted as one of the risk factors for VTE, in an out-patient population, as our, the attribute "cancer" is expanded into several clinical attributes (i.e., site and stage or anti-cancer/supportive drugs) that portend different degrees of risk, and might "weight" differently in the context of a ML algorithm.

There are, of course, some limitations to acknowledge. First, the model here reported was designed and validated on a dataset, which was not extracted from the EHR of single patients, due to privacy restrictions in reference to identifiable individuals, as the Medical Oncology Unit stores EHRs under data protection legislation. These records, however, are highly customized into structured and non-structured fields including demographics, medical and family history, vital signs, medications, diagnostics and follow-up updating. Thus, all variables necessary for prediction are easily extractable from EHRs, once the model is validated for clinical use, as recently demonstrated by Lustig et al., who implemented KS with EHRs extraction to readily stratify at-risk patients.[36] Although glycemic profile and blood lipid pattern might not be always included in the pre-chemotherapy patient workout, we should take into consideration that these analytes are easy to perform and relatively inexpensive. This facilitates their inclusion in a validated clinical model with a negligible increase in health care costs.

Another limitation might reside in the fact that the study was monocentric. However, primary aim of this study was not to present a new classifier that other Centers can adopt, but rather to propose the

application of ML approach in VTE risk assessment models. Here, we demonstrate that the use of ML algorithms and RO models might be useful in developing local classifiers capable of improving the original KS, while retaining other advantages (e.g., recalculation based on data advance over time) in a perspective of precision medicine. Presently, we are involved in the development of an operator-friendly web interface, whose server component calculates VTE risk based on ML-RO-2 and returns to the client a binary information on risk (yes/no).

CONCLUSIONS

In conclusion, a ML approach might represent a suitable approach to VTE risk prediction by taking into consideration individual biological variability, environmental exposure and lifestyle, in a context of precision medicine. This is particularly appealing in a Big-Data scenario, in which clinical/biochemical attributes, routinely collected in EHRs, may be all used to design new tools for clinical decision making. Indeed, the method we propose to find the optimal VTE predictors has the unquestionable advantages of selecting the best predictors on training data and to determine the relative weights between groups of clinical attributes. Furthermore, it demonstrates that other variables must be considered in VTE risk evaluation, thus strengthening the concept that data should not be considered singularly but in a more general association, as advocated by precision medicine.

This risk stratification approach well fits with others who identified the need of developing new guidelines or of identifying topics deserving further ad hoc clinical trials,[37] and might help in filling the gap left by current guidelines concerning VTE prophylaxis.

Ongoing research involves: 1) the use of other optimization methods such as simulated annealing and genetic algorithms; 2) the development of a web server interface using the proposed algorithm and its external validation by collaborating oncology wards. Nonetheless, the results here reported add further evidence to the rising idea that locally trained models may be of advantage over the classic scoring schemes, which, in time, can lose their prediction value and become less accurate.

Competing Interests

Authors declare no conflict of interest.

REFERENCES

1. Liebman HA, Khorana AA, Kessler CM. Clinical Roundtable Monograph: The Oncologist's Role in the Management of Venous Thromboembolism. Clin Adv Hematol Oncol. 2011;9(1):1–15.
2. Mandalà M, Falanga A, Roila F On behalf of the ESMO Guidelines Working Group. Venous thromboembolism in cancer patients: ESMO Clinical Practice Guidelines for the management. Ann Oncol. 2010;21 Suppl 5:v274–6.
3. Lyman GH, Bohlke K, Khorana AA, Kuderer NM, Lee AY, Arcelus JI, Balaban EP, Clarke JM, Flowers CR, Francis CW, Gates LE, Kakkar AK, Key NS, Levine MN, Liebman HA, Tempero MA, Wong SL, Somerfield MR, Falanga A; American Society of Clinical Oncology. Venous thromboembolism prophylaxis and treatment in patients with cancer: American Society of Clinical Oncology clinical practice guideline update 2014. J Clin Oncol. 2015;33(6):654–6.
4. Sousou T, Khorana AA. New insights into cancer-associated thrombosis. Arterioscler Thromb Vasc Biol. 2009;29(3):316–20.
5. Di Nisio M, Ferrante N, De Tursi M, Iacobelli S, Cuccurullo F, Büller HR, Feragalli B, Porreca E. Incidental venous thromboembolism in ambulatory cancer patients receiving chemotherapy. Thromb Haemost. 2010;104(5):1049–54.
6. Roselli M, Ferroni P, Riondino S, Mariotti S, Laudisi A, Vergati M, Cavaliere F, Palmirotta R, Guadagni F. Impact of chemotherapy on activated protein C-dependent thrombin generation - Association with VTE occurrence. Int J Cancer. 2013;133(5):1253–8.
7. Khorana AA, Kuderer NM, Culakova E, Lyman GH, Francis CW. Development and validation of a predictive model for chemotherapy-associated thrombosis. Blood. 2008;111(10):4902–7.
8. Khorana AA, Otten HM, Zwicker JI, Connolly GC, Bancel DF, Pabinger I; Subcommittee on Haemostasis and Malignancy. Prevention of venous thromboembolism in cancer outpatients: guidance from the SSC of the ISTH. J Thromb Haemost. 2014;12(11):1928–31.

9. Ay C, Dunkler D, Marosi C, Chiriac AL, Vormittag R, Simanek R, Quehenberger P, Zielinski C, Pabinger I. Prediction of venous thromboembolism in cancer patients. Blood. 2010;116(24):5377–82.
10. Mandalà M, Clerici M, Corradino I, Vitalini C, Colombini S, Torri V, De Pascale A, Marsoni S. Incidence, risk factors and clinical implications of venous thromboembolism in cancer patients treated within the context of phase I studies: the 'SENDO experience'. Ann Oncol. 2012;23(6):1416–21.
11. Verso M, Agnelli G, Barni S, Gasparini G, LaBianca R. A modified Khorana risk assessment score for venous thromboembolism in cancer patients receiving chemotherapy: the Protecht score. Intern Emerg Med. 2012;7(3):291–2.
12. Crowley MP, Eustace JA, O'Shea SI, Gilligan OM. Venous thromboembolism in patients with myeloma: incidence and risk factors in a "real-world" population. Clin Appl Thromb Hemost. 2014;20(6):600–6.
13. Srikanthan A, Tran B, Beausoleil M, Jewett MA, Hamilton RJ, Sturgeon JF, O'Malley M, Anson-Cartwright L, Chung PW, Warde PR, Winquist E, Moore MJ, Amir E, Bedard PL. Large retroperitoneal lymphadenopathy as a predictor of venous thromboembolism in patients with disseminated germ cell tumors treated with chemotherapy. J Clin Oncol. 2015;33(6):582–7.
14. Lim SH, Woo SY, Kim S, Ko YH, Kim WS, Kim SJ. Cross-sectional Study of Patients with Diffuse Large B-Cell Lymphoma: Assessing the Effect of Host Status, Tumor Burden, and Inflammatory Activity on Venous Thromboembolism. Cancer Res Treat. 2016;48(1):312–21.
15. Chen JH, Podchiyska T, Altman RB. OrderRex: Clinical order decision support and outcome predictions by data-mining electronic medical records. J Am Med Inform Assoc. 2015;pii:ocv091. doi:10.1093/jamia/ocv091.
16. Mani S, Chen Y, Li X, Arlinghaus L, Chakravarthy AB, Abramson V, Bhave SR, Levy MA, Xu H, Yankeelov TE. Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy. J Am Med Inform Assoc. 2013;20(4):688-95.

17. Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, Roelofs E, van Elmpt W, Boutros PC, Granone P, Valentini V, Begg AC, De Ruysscher D, Dekker A. Predicting outcomes in radiation oncology—multifactorial decision support systems. Nat Rev Clin Oncol. 2013;10(1):27–40.
18. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A, Robles V. Machine learning in bioinformatics. Brief Bioinform. 2006;7(1):86–112.
19. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak. 2011 Jul 29;11:51.
20. Deo RC. Machine learning in medicine. Circulation. 2015; 132 (20):1920–30.
21. Kawaler E, Cobian A, Peissig P, Cross D, Yale S, Craven M. Learning to predict post-hospitalization VTE risk from EHR data. AMIA Annu Symp Proc. 2012; 2012:436–45
22. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.
23. Jensen LJ, Bateman A. The rise and fall of supervised machine learning techniques. Bioinformatics. 2011;27 (24):3331–2.
24. Matyas J. Random optimization. Automat Rem Contr. 1965;26:246–53.
25. Bennett KP, Momma M, Embrechts MJ. MARK: A boosting algorithm for heterogeneous kernel models. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 24–31, 2002.
26. Gönen M, Alpaydın E. Multiple kernel learning algorithms. J Mach Learn Res. 2011;12:2211–68.
27. Saar-Tsechansky M, Provost F. Handling Missing Values when Applying Classification Models. J Mach Learn Res. 2007;8(December 2007):1623–57.
28. Vergati M, Della Morte D, Ferroni P, Cereda V, Tosetto L, Riondino S, La Farina F, Guadagni F, Roselli M. Increased Risk of Chemotherapy-Associated Venous Thromboembolism in Elderly Patients with Cancer. Rejuvenation Res. 2013;16(3):224–31.

29. Ferroni P, Guadagni F, Riondino S, Portarena I, Mariotti S, La Farina F, Davì G, Roselli M. Evaluation of mean platelet volume as a predictive marker for venous thromboembolism in chemotherapy-treated cancer patients. Hematologica. 2014;99(10):1638–44.
30. Ferroni P, Riondino S, Formica V, Cereda V, Tosetto L, La Farina F, Valente MG, Vergati M, Guadagni F, Roselli M. Clinical significance of neutrophil lymphocyte ratio and platelet lymphocyte ratio in venous thromboembolism (VTE) risk prediction in ambulatory cancer patients treated with chemotherapy. Int J Cancer. 2015;136(5):1234–40.
31. Filice S, Castellucci G, Croce D, Basili R. KeLP: a Kernel-based Learning Platform for Natural Language Processing. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 19–24, 2015
32. Ferroni P, Roselli M, Riondino S, Guadagni F. Predictive value of high-density lipoprotein (HDL)-cholesterol for cancer-associated venous thromboembolism during chemotherapy. J Thromb Haemost. 2014;12(12):2049–53.
33. Dickmann B, Ahlbrecht J, Ay C, Dunkler D, Thaler J, Scheithauer W, Quehenberger P, Zielinski C, Pabinger I. Regional lymph node metastases are a strong risk factor for venous thromboembolism: results from the Vienna Cancer and Thrombosis Study. Haematologica. 2013;98(8):1309–14.
34. Ferroni P, Riondino S, Guadagni F, Roselli M. Impact of chemotherapy on venous thromboembolism. Haematologica. 2013;98(8):e153–e154
35. Carrier M, Le Gal G, Wells PS, Fergusson D, Ramsay T, Rodger MA. Systematic review: the Trousseau syndrome revisited: should we screen extensively for cancer in patients with venous thromboembolism? Ann Intern Med. 2008;149(5):323–33.
36. Lustig DB, Rodriguez R, Wells PS. Implementation and validation of a risk stratification method at The Ottawa Hospital to guide thromboprophylaxis in ambulatory cancer patients at intermediate-high risk for venous thrombosis. Thromb Res. 2015 Dec;136(6):1099-102.

37. Bosson JL, Labarere J. Determining indications for care common to competing guidelines by using classification tree analysis: application to the prevention of venous thromboembolism in medical inpatients. Med Decis Making. 2006;26(1):63-75.

FIGURE LEGENDS

Figure 1. Groups of clinical attributes. NLR: Neutrophil/lymphocyte ratio; PLR: platelet/lymphocyte ratio; BMI: body mass index; ECOG-PS: Eastern Cooperative Oncology Group Performance Status; eGFR: estimated glomerular filtration rate. The group "Drugs" includes all supportive and anti-cancer agents listed in Supplementary Table 2.

Figure 2. Recall vs. Precision curves of the ML systems. The plot for Khorana was obtained by computing Recall and Precision for 4 cut-off values: 3 (the standard value), 2, 1 and 0. Numbers in open rectangles report the Area Under the Precision-Recall Curve (AUCPR). PPV: Positive Predictive Value.

Figure 3. Kaplan–Meier curves of VTE-free survival of chemotherapy treated ambulatory cancer patients in the testing set. Comparison between patients with low (dotted line) or high (solid line) risk of VTE based on a SVM VTE event predictor using only the KS as feature (Khorana-ML)(Panel A) or the two best fitting ML-RO models: ML-RO-3-K (Panel B) and ML-RO-2 (Panel C).