Medical Decision Making

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

Journal:	*Medical Decision Making*
Manuscript ID	Draft
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Ferroni, Patrizia; San Raffaele Rome University, IRCCS San Raffaele Pisana Zanzotto, Fabio; Universita degli Studi di Roma Tor Vergata, Department of Enterprise Engineering Scarpato, Noemi; San Raffaele Rome University Riondino, Silvia; IRCCS San Raffaele Pisana; Universita degli Studi di Roma Tor Vergata, Department Systems Medicine, Medical Oncology Nanni, Umberto; Universita degli Studi di Roma La Sapienza, Department of Computer, Control, and Management Engineering Antonio Ruberti Roselli, Mario; Universita degli Studi di Roma Tor Vergata, Department of Systems Medicine, Medical Oncology Guadagni, Fiorella; San Raffaele Rome University, IRCCS San Raffaele Pisana
APPLICATION AREAS:	ONCOLOGY
DETAILED METHODOLOGY:	Clinical Prediction Rules < CLINICAL RESEARCH METHODOLOGY, Risk Stratification < CLINICAL RESEARCH METHODOLOGY, Decision Support Techniques < INFORMATICS, Survival Analysis < STATISTICAL METHODS

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 Rome 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 Rome 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: 4366

[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 VTE risk predictors for ambulatory cancer patients that might be of advantage over the currently recommended model (Khorana score).

Design: Kernel machine and random optimization (RO) models were used to produce VTE risk predictors yielding the best classification performance over a training (3-fold cross validation) and testing set.

Results: Clinical attributes of the patient dataset were divided into 9 groups according to clinical significance. Our analysis produced 6 RO models in the training set, which yielded better hazard ratios (HRs) compared with baseline models (HRs ranging from 1.45 to 3.36) and were all significant in terms of VTE risk prediction. With only one exception, the superiority of these models over their baseline counterparts was validated in the testing set, in which the probability of VTE occurrence in patients classified as at-risk by 2 RO models (HRs 4.48 and 6.92) was 2 to 3-fold higher than that observed using the pure Khorana score (HR 2.16). Of interest, the best fitting model was one in which the strongest weight was retained by blood lipids, body mass index and performance status, with a weaker association with tumor site/stage and drugs.

Conclusions: Although the monocentric validation of the predictors here presented might represent a limitation, these results demonstrate that a model based on kernel learning machines and RO may outperform the currently recommended score, and has the unquestionable advantage to be dynamically recalculated and integrated with local data. 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 substantially changed: global approaches have been pressured by a growing availability of electronic health records (EHR) and by the consequent demand to provide *precision medicine*. The intuition is that precision medicine can produce better approaches to disease treatment and prevention by taking into consideration individual biological variability, environmental exposure and lifestyle.

Oncology is a field that could significantly benefit from a precision medicine-based approach, both in the development of targeted therapies, which represent a key to successful patient treatment, and in other clinical contests, in order to improve treatment delivery and clinical outcome.

One of the major challenges that oncologists are presently facing is the risk assessment of venous thromboembolism (VTE). The development of VTE, in fact, may result in treatment delays with detrimental effects on the overall outcome for cancer care and patient's quality of life.[1] Hence, the use of appropriate thromboprophylaxis in cancer patients treated with chemotherapy could provide an opportunity to substantially improve their clinical management.[2] Nonetheless, all current consensus guidelines do not recommend routine prophylaxis for the primary prevention of VTE in ambulatory cancer patients receiving chemotherapy,[3,4] although "*it may be considered for selected high-risk patients*" following "*discussion with the patient about the uncertainty concerning benefits and harms*".[4] These statements emphasize how selecting patients for prophylactic anticoagulation is perceived as a growing necessity in cancer patient management, fostering the demand for risk assessment models.

Predicting VTE risk for cancer patients is, thus a compelling challenge where precision medicine can play a crucial role. In fact, VTE risk differs not only among patients, but even in the same patient over the course of cancer natural history.[5] The highest risk is in the first 3-6 months after initial diagnosis possibly as a result of combined anti-cancer therapies in the same time range.[6,7]

However, implementing an effective VTE risk predictor for cancer patients is very difficult. Khorana and colleagues developed and validated an interesting model for predicting chemotherapy-associated VTE using a combination of routinely available variables.[8] This model takes into account the site of cancer (2 points for very-high-risk stomach, or pancreas cancer; 1 point for high-risk lung, or genitourinary cancer and 0 point for all other solid cancer sites), platelet count ≥350 x 10^9/L, leukocyte count ≥11 x 10^9/L, hemoglobin ≤10 g/dL and/or use of erythropoiesis-stimulating agents and body mass index (BMI) ≥35 kg/m^2 (1 point each).[8] To date, this is the sole model for VTE risk assessment in ambulatory cancer patients treated with chemotherapy. As such, the Khorana score has been proposed in the guidance statement of the Scientific and Standardization Committee of the International Society for Thrombosis and Haemostasis.[9] Nonetheless, although validated by independent groups,[10,11] the Khorana score fails to classify 40% to 60% of patients (intermediate risk), in whom clinical decision making remains challenging. Consistent with these observations, expanded risk scoring models, including either laboratory tests,[10] or the anti-cancer drug used,[12] were proposed. Despite these efforts, VTE risk prediction for chemotherapy-treated cancer outpatients is still sub-optimal.

A precision medicine approach might help to overcome many of the problems encountered so far. Nonetheless, the general problem of precision medicine, which arises also in the case of VTE risk prediction for oncological patients, is that this method considers a huge amount of clinical variables.[13] This is both the power and a possible drawback of precision medicine and highlights the urgent need for a new generation of computational theories and tools to assist researchers in extracting knowledge from the growing volumes of digital data.[14]

Based on these considerations, we hypothesized that machine-learning (ML) models can help in solving this problem. ML is gaining popularity in medicine and in bioinformatics,[15-19] as it can derive patterns in clinical and biochemical knowledge (for a recent review see[20]). Moreover, ML has been also applied to learn VTE risk predictors for the general population,[21] and could thus represent a solid base on which to build the next generation of precision medicine approaches in oncology.

Therefore, aim of the present study was to analyze the performance of a different approach from that generally used in the development of risk assessment models based on the arbitrary assignment of a score according to association analyses (i.e. Khorana score). To this purpose, we used kernel learning machines,[22] (as suggested by Jensen and Bateman[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 of ambulatory cancer patient prior to chemotherapy start.

METHODS

Learning VTE Risk Predictors within a Precision Medicine Approach

In the context of precision medicine, we introduced a new methodology to produce VTE risk predictors that exploit personalized data. Our methodology is based on a particular class of learning machines, namely, kernel machines,[22] and on a model to devise relative importance of different groups of clinical attributes in final prediction decisions, namely, *RO*.[24]

VTE risk predictors are binary classifiers that, given a patient *x*, have to determine whether or not *x* will develop a VTE event in the future. In ML (see[18] for details), binary classifiers are functions *f(x)=y* that take as input instances *x* and emits a class *y*∈*{1, -1}*. Instances *x* are represented as vectors of feature values $\vec{x} = (x_1, \dots, x_n)$. Hence, in our settings, *x* is a patient, and *y=1* is the prediction of the occurrence of VTE in the future. Finally, each feature of vectors $\vec{x}$ represents one of the clinical attributes. Therefore, the challenge is to build binary classifiers that make a good use of the information stored in these vectors of feature values.

Inducing binary classifiers *f(x)* by observing training data *T* is the major objective of ML. This activity is called *learning*. Hence, the output *y* of the learnt classifier depends on x and on T, that is *f(x,T)*. During learning, specific algorithms discover regularities in training data by comparing instances. In our study, we use a particular class of learning algorithms called kernel machines.[22] These machines compare instances $x^{(a)}$ and $x^{(b)}$ by doing a dot product between their unit vectors, that is, $\langle \vec{x}^{(a)}, \vec{x}^{(b)} \rangle$. This dot product is called *kernel* and is often referred as $K\left(\vec{x}^{(a)}, \vec{x}^{(b)}\right)$. The kernel of unit vectors is close to 1 if vectors are similar. Thus, roughly, kernel machines tend to classify novel examples by computing the similarity with training examples. In fact, the learnt function is:

$$f(x, T) = sign(\langle \vec{w}, \vec{x} \rangle) \qquad (1)$$

where $\vec{w}$ is a linear combination of vectors of training examples in *T* and $\vec{w}$ is the result of the learning phase.

There is a large body of research in ML to induce the best classifiers from training data, but a real problem with medical data is represented by the heterogeneity of patients' clinical attributes. These attributes participate to the final classification decision with different weights: the vector $\vec{w}$ in Equation (1). However, these weights are rigidly derived from a linear combination of training examples. This is not sufficient to determine the relative weights between groups of very different clinical attributes.

Finding optimal ways to combine heterogeneous groups of different attributes is thus a major problem both in precision medicine and in ML in general.[25]

In this study, we used a simple method: combining kernel machines to learn predictors and *RO*[24] to optimize their performances by changing the relative weight of groups of features. With RO, our method finds the combination of groups of attributes that yields to the best classification performance of our predictors over a validation set.

Optimizing asks for a clear definition of the evaluation procedure. Evaluating classifiers is an important part of the learning process. Classifiers are evaluated on testing data sets that are completely separated from the training data. For unbalanced classes, performances are evaluated with *positive predictive value* (*PPV)*, with *sensitivity*, and with a combination of the two. In ML, these measures are called Precision (P), Recall (R), and f-measure, respectively. Hereafter, we indicate the value of the f-measure for a function $f(\vec{x})$ on the testing data set V as:

$$Per(f(\vec{x}),V) = \frac{2PR}{P+R}$$

where P and R are Precision and Recall of $f(\vec{x})$ on V, respectively.

We have thus a way to optimize predictors' evaluation.

Our method that combines kernel machines and RO is the following. First, clinical attributes are divided in groups according to clinical considerations. Each group has an associated sub-vector $\vec{g}_i$ in feature vectors representing patients. These vectors are obtained as a juxtaposition of sub-vectors, that is, $\vec{x} = [\vec{g}_1, \vec{g}_2, \dots,$ $\vec{g}_m]$. Second, the relative weight among groups of features is determined with a vector $\vec{\omega} = (\omega_1 \dots,\ \omega_m)$ of group weights. Hence, the kernel between two vectors of instances $\vec{x}^{(a)}, \vec{x}^{(b)}$according to a weight vector $\vec{\omega}$ is defined as follows:

$$K\left(\vec{x}^{(a)}, \vec{x}^{(b)}, \vec{\omega}\right) = \sum_i \omega_i \left\langle \vec{g}_i^{\,(a)}, \vec{g}_i^{\,(b)} \right\rangle \Big/ \sum_i \omega_i$$

In this new setting, the kernel machine learns a classifier f that depends on $\vec{x}$, on the training set T and, finally, on $\vec{\omega}$, that is, $f(\vec{x}, \vec{\omega}, T)$. Next, we used *RO* to find $\vec{\omega}_{max}$ that maximizes the performance of the classifier $f(\vec{x}, \vec{\omega}, T)$ on a validation set V:

$$\vec{\omega}_{max} = \underset{\vec{\omega}}{\operatorname{argmax}}\, Per(f(\vec{x}, \vec{\omega}, T), V)$$

Basically, the method sets an initial random vector $\vec{\omega}$, learns $f(\vec{x}, \vec{\omega}, T)$ with the kernel machine, and determines its performance $p = Per(f(\vec{x}, \vec{\omega}_0, T), V)$. Then, it starts a cycle where it randomly generates a perturbation vector $\vec{a}$, learns $f(\vec{x}, \vec{\omega} + \vec{a}, T)$, computes $p' = Per(f(\vec{x}, \vec{\omega} + \vec{a}, T), V)$ and, if $p' > p$, updates $\vec{\omega} \leftarrow \vec{\omega} + \vec{a}$ and $p \leftarrow p'$. The cycle stops when after *n* perturbation vectors, no one produced $p' > p$. The final $\vec{\omega}$ is retained as $\vec{\omega}_{max}$.

Our method to find the best VTE risk predictors has two major benefits: first, it selects the best predictors on training data $f(\vec{x}, \vec{\omega}_{max}, T)$; second, it determines relative weights $\vec{\omega}_{max}$between groups of clinical attributes. These weights give useful insights on how predictors take their decisions.

Patient dataset for VTE risk assessment

Patient dataset for VTE risk assessment 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).

The patient dataset consisted of 1179 consecutive ambulatory cancer patients with primary or relapsing/recurrent solid cancers, who were prospectively followed under the appropriate Institutional ethics approval and in accordance with the principles embodied in the Declaration of Helsinki to investigate possible predictors of chemotherapy-associated VTE. All patients were required to be at the start of a new chemotherapy regimen and no patient received thromboprophylaxis, according to current guidelines. Eligibility criteria are detailed in Supplementary Table 1. Clinical characteristics and laboratory attributes of patients are summarized in Supplementary Table 2.

Baseline blood samples were drawn at time of enrolment prior to chemotherapy start and tested for routine blood chemistry (Accelerator Total Lab Automation, Abbott Laboratories, Abbott Park, IL, USA) and complete and differential blood cell counts (Coulter LH 750, Beckman Coulter, Miami, FL) in the facilities of the BioBIM of the IRCCS San Raffaele Pisana. Estimated glomerular filtration rate (eGFR) was calculated using the simplified Modification Diet of Renal Disease study (MDRD) equation.[26]

VTE was diagnosed at the Medical Oncology ward of the Department of Systems Medicine, PTV during scheduled chemotherapy visits, or at the occurrence of clinically suspected VTE. Deep venous thrombosis (DVT) was confirmed by venography or color-coded duplex sonography (in proximal DVT only). Pulmonary embolism (PE) was diagnosed by spiral computed tomography displaying 1 or several low-attenuation areas that partly or completely filled the lumen of an opacified vessel. Within 1 year of study entry, VTE occurred in 8% (29 PE and 65 DVT) of patients. 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.

All patients provided written informed consent, previously approved by our Institutional Ethics Committees.

Experimental settings

To test our methodology and to test default methods, the patient dataset was used in the following ways: 1) we divided clinical attributes in 9 groups; 2) we randomly divided the patient dataset in training and testing set, 3) we rescaled continuous clinical attribute values to lay in the range [-0.5,0.5]; and, finally, 4) we filled missing clinical attribute values with the average of the attribute observed in the training set.

Regarding the patient dataset division, group distribution was performed according to the clinical significance of the attributes included in the patient dataset. In particular, demographic variables and tumor site and stage were individually considered given that they are generally recognized among the most important risk factors for VTE.[5,27] Hematological attributes, including complete and differential blood cell counts,[8,28] as well as neutrophil and platelet to lymphocytes ratios,[29] were grouped all together. Similarly, individual attributes concerning fasting blood lipids, glycemic indexes and liver and kidney function were clustered within three individual groups. BMI and Eastern Cooperative Oncology Group Performance Status (ECOG-PS) were considered within the same group, as the former might represent not only an obesity index, but is also indicative of underweight in patients with cancer cachexia and, as such, can affect the performance status of a particular patient. Supportive and anti-cancer drugs were collectively considered under the general definition of "drugs". In some experiments, tumor site, BMI, hemoglobin or erythropoiesis supporting agents, white blood and platelet counts were categorized as previously suggested,[8] and grouped as Khorana score, which served as reference for subsequent analyses. Details on groups of clinical attributes are reported in Figure 1.

To apply ML models, the patient dataset was randomly divided in two groups:

1) Training (Tr): 70% of the cases
2) Testing (Ts): 30% of the cases

The Training set was used to optimize the parameters $\vec{\omega}$ with RO and to learn the final risk predictor using the selected $\vec{\omega}_{max}$. RO was applied using 3-fold cross validation. The training set was divided in three parts and three runs were performed. For each step of the RO, we computed the performance of three learnt risk predictors on one of the three parts of the training set. These risk predictors were learnt using the remaining two parts. RO stops when the algorithms cannot improve a local maximum and selects $\vec{\omega}_{max}$ (see Table 1). The 3 ML-RO are the top-3 f-measure in the training set for two different experiments including or not the "Khorana" group (Table 2). Then, we used $\vec{\omega}_{max}$ to learn the final risk predictor $f(\vec{x}, \vec{\omega}_{max}, Tr)$. The testing set was used to compute the final performance of our risk predictors (Table 2).

Statistical analysis

Time-to-event (TTE) was calculated from the enrolment date until VTE or the most recent follow-up visit (median TTE: 3 months). 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). Cox-proportional hazards analyses were performed by a free web-based application (http://statpages.org/).

This study had no external funding source.

RESULTS

The weights of attribute classes for the ROs models are reported in Table 1. Table 2 summarizes the results achieved using the top-3 models out of 5 runs obtained with RO using Khorana score (*ML+RO-1-K, ML+RO-2-K, and ML+RO-3-K*), the top-3 models out of 5 runs obtained with RO without Khorana score (*ML+RO-1, ML+RO-2, and ML+RO-3*), and 4 different baseline models: 1) *Khorana k≥3*: pure Khorana Score with cutoff at 3;[9] 2) *Khorana-ML*: an SVM VTE event predictor trained with a polynomial kernel of degree 2 that uses only the Khorana Score 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 Khorana score and *Basic-ML* does not use it. As shown, a ML approach with RO was capable of improving VTE risk prediction compared to *Khorana k≥3* or Khorana-ML as demonstrated by comparable precision (or positive predictive value – PPV) and considerably higher recall (or sensitivity) values, translating in a substantial improvement of the F-measure.

These results were confirmed by Cox-proportional hazards survival analyses, in which Hazard Ratio (HR) and 95% Confidence Intervals denoted the ratio of the probabilities of VTE occurrence in patients classified as *at-risk* or *low-risk* by the ML models applied to the dataset. As shown in Table 2, Khorana score, analyzed either as pure Khorana score (*Khorana k≥3*) or Khorana-ML, failed to achieve the statistical significance in risk estimation analysis when applied to the training set, whereas both basic-ML models, with (Basic-ML-K) or without (Basic-ML) inclusion of the Khorana score, yielded weak, but significant HRs (basic-ML: HR=1.69, p=0.040; Basic-ML-K: HR=1.91, p=0.019). With only one exception (ML+RO-2), risk estimation for all ROs models (*ML+RO-1-K, ML+RO-2-K, ML+RO-3-K, ML+RO-1, and ML+RO-3*) in the training set yielded HRs ranging from 2.03 to 3.36, which were all significant in terms of VTE risk prediction.

Table 1: Weights of attribute classes 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
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

Table 2: Results of basic predictors and predictors based on machine-learning with random optimization

Method	Precision (PPV)	Recall (Sensitivity)	F-Measure	HR (95%CI)	Precision (PPV)	Recall (Sensitivity)	F-Measure	HR (95%CI)
Khorana (k>=3)*	0.122	0.075	0.093	1.86 (0.75-4.63)	0.136	0.111	0.122	2.16 (0.65-7.18)
Khorana-ML	0.065	0.448	0.114	0.85 (0.51-1.41)	0.063	0.593	0.113	0.55 (0.26-1.19)
Basic-ML-K	0.096	0.642	0.167	1.91 (1.12-3.29)	0.099	0.852	0.177	3.23 (1.12-9.33)
ML-RO-1-K	0.126	0.761	0.217	3.04 (1.80-5.14)	0.105	0.741	0.184	2.61 (1.10-6.17)
ML-RO-2-K	0.119	0.791	0.207	3.24 (1.80-5.84)	0.100	0.778	0.177	2.55 (1.03-6.33)
ML-RO-3-K	0.115	0.687	0.197	2.10 (1.28-3.43)	0.112	0.704	0.193	2.73 (1.19-6.24)
Basic-ML	0.091	0.537	0.155	1.69 (1.02-2.78)	0.078	0.593	0.137	1.09 (0.51-2.35)
ML-RO-1	0.117	0.716	0.202	3.36 (1.94-5.84)	0.082	0.556	0.143	1.21 (0.57-2.59)
ML-RO-2	0.115	0.731	0.198	1.45 (0.89-2.36)	0.122	0.889	0.214	6.92 (2.08-23.0)
ML-RO-3	0.115	0.702	0.197	2.03 (1.23-3.35)	0.119	0.815	0.208	4.48 (1.70-11.8)

*Patients with brain cancer (n=7) were excluded from the analysis (Khorana score not applicable)

Validation step was then performed on the testing set. As summarized in Table 2, all ML models including the Khorana score resulted in an overall improvement of the performance measures for VTE risk prediction, both in terms of F-measure and HRs compared to the pure Khorana score, although the best fitting model in terms of clinical risk prediction was represented by *ML+RO-3-K* (HR=2.73, p=0.017). On the other hand, the ML approach not including the Khorana score yielded significant results in the survival analyses only in *ML-RO-2* (p=0.002) and *ML-RO-3* (p=0.003) in which patients classified as at-risk had approximately 7 and 5-fold higher risks of developing VTE during chemotherapy administration than patients classified at no-risk.

Kaplan–Meier curves for patients in the testing set stratified on the basis of *Khorana k≥3* and Khorana-ML are reported in Figure 2. As shown, despite a high precision, the Khorana score used at a cut-off ≥3 points, as currently recommended,[9] resulted in a very low sensitivity (only 3 of 27 VTE recorded events occurred in patients classified as at-risk) with a sub-optimal negative predictive value (NPV=0.928) and a 6-month VTE-free survival rate not significantly different from that of low-risk patients (86% vs. 93%)(Figure 2A). Similar considerations can be drawn for the ML predictor using the Khorana feature alone (Figure 2B).

Figure 3 depicts the Kaplan-Meier curves for the two best fitting models obtained with RO with (*ML-RO-3-K*) or without (*ML-RO-3*) Khorana score in the testing set. As shown, optimizing the relative importance (weight) of groups of clinical attributes resulted in an approximately 3 to 7-fold improvement of VTE risk prediction. In particular, patients classified at-risk with *ML-RO-3* had a significantly lower 6-month VTE-free survival (87%) compared to patients classified as low-risk (99%)(Figure 3B). Kaplan-Meier survival curves of patients stratified with the other ML-RO models are reported in Supplementary Figure 1.

DISCUSSION

The present study was designed to address the challenging task of VTE risk prediction in chemotherapy-treated cancer outpatients. To this purpose, we used ML methods to build predictive models which consider different variable types, such as demographic, laboratory and clinical data (including therapies), routinely collected in EHRs, to retrospectively identify chemotherapy-associated VTE events in a general medical oncology centre population.

Here, for the first time to our knowledge, we propose a precision medicine model to design VTE risk predictors for oncological patients treated with chemotherapy. 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 validation set. Finally, we compared the predictive value of our learned models against the previously developed Khorana's risk assessment tool.

The results obtained demonstrated that this approach may be of advantage in the selection of VTE risk predictors over the currently accepted models and allowed us to draw a number of interesting considerations.

First, the analysis of the variables collected from each patient identified several risk factors, not previously included in VTE risk assessments as per current guidelines. In general, precision medicine approaches were better than generic ones. In fact, ML models using all the clinical attributes (Basic-ML-K, Basic-ML and ML-ROs) showed better F-measures and better HRs than generic models (pure Khorana score and Khorana-ML). This was verified on the training and, more importantly, on the testing set. Using additional clinical attributes is thus promising.

Second, our approaches ML+ROs appeared extremely useful in designing VTE risk predictors. By optimizing the relative importance of groups of clinical attributes, we selected better risk predictors. It is obvious that on the training set f-measures of ML+ROs were better than Basic-ML as RO was carried out on the training set. It is less obvious that all ML-ROs outperformed Basic-MLs on the testing set in terms of f-measure and that only ML-RO-1 was not superior to Basic-ML in terms of HR.

Most importantly, best scoring models in terms of both f-measure and HRs were also clinically plausible, as demonstrated by the finding that blood lipids and body mass index and performance status retained the strongest weight both in ML-RO-3-K and in ML-RO-2. This is consistent with the literature showing that low levels HDL-cholesterol[30] and ECOG-PS[27,29] proved good predictors of increased risk of VTE in chemotherapy-treated cancer patients, in multiple regression models. Moreover, the ML-RO-2 model showed a weak association with tumor site and stage, and with drugs. This is not surprising, since both clinical attributes have also been associated with an increased risk of developing VTE.[8,12] Indeed, advanced cancer, either locally (regional) or distant, represents per se an increased risk of VTE, [31] and we must acknowledge the role that anti-cancer drugs may play as thrombotic triggers in association with specific disease stages.[7,32] Indeed, anticancer therapies represent an important predisposing factor for VTE, capable of inducing an acquired thrombophilic condition,[7] at a point that certain anti-cancer agents have been proposed to be included in the Khorana's score in order to implement it, as in the case of the Protecht score.[12]

Finally, the low f-measures obtained with our VTE risk predictors could be explained with the fact that our patient dataset was extremely unbalanced. Indeed, VTE occurred only in 8% of the cases. Hence, applying ML models to this dataset was extremely difficult, consistently with Larrañaga et al.[19] Experiments with VTE predictors in general population have better performance,[21] but the test set generally used, consisted of VTE cases paired to non-VTE controls, resulting in a more balanced set. Hence, VTE predictors in these studies[21] cannot be compared to our study cohort, in which ambulatory cancer patients were consecutively enrolled and all VTE events were prospectively recorded by the oncologists during follow-up. Moreover, we must take into consideration that in hospitalized patients, cancer is connoted as one of the risk factors for VTE development, to such an extent that about 60% of occult cancers are diagnosed shortly after the diagnosis of an episode of unprovoked VTE.[33] Conversely, in an oncological out-patient population, such as the one analyzed in our study, the attribute "cancer" is expanded to take into account individual groups of clinical attributes (i.e., cancer site and stage or administered anti-cancer or supportive drugs) that, as already stated, portend different degrees of clinically significant VTE risk, and might "weight" differently in the context of a ML algorithm.

There are, of course, some limitations that need to be acknowledged. First, we must recognize that the model here reported was designed and validated on a dataset, which was not actually extracted from the EHR of single patients, due to privacy restrictions in reference to identifiable individuals. As a matter of fact, the Medical Oncology Unit stores patients' data in a digital format in EHRs, under data protection legislation. These records 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 the Khorana score with EHRs extraction to readily stratify patients into intermediate-high and low risk of VTE.[34] 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, thus validation was limited to a single institution. However, primary aim of this study was not to present a new classifier that other Centers can adopt for clinical use, but rather to propose a different approach from that generally utilized in risk assessment models, based on the arbitrary assignment of a score according to association analysis. Here, we demonstrate that the use of ML algorithms and RO models might be of advantage in developing local classifiers capable of improving the original Khorana score, while retaining other advantages (e.g., recalculation based on data advance over time) in a perspective of precision medicine.

CONCLUSIONS

In conclusion, 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. This model showed to outperform the general well-assessed Khorana score. 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.

Furthermore, 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,[35] and might fill the gap left by current guidelines concerning VTE prophylaxis in intermediate risk patients. In this context, future application of our model might help oncologists in the delicate phase of decision making, by providing them with the great advantage of limiting observer subjectivity.

Ongoing research involves: 1) the use of other optimization methods such as simulated annealing and genetic algorithms; and 2) the definition of a VTE risk prediction system. Of course, the prediction system will require larger sets of cases and controls to be acquired in future research projects. 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. Agnelli G, Gussoni G, Bianchini C, Verso M, Mandalà M, Cavanna L, Barni S, Labianca R, Buzzi F, Scambia G, Passalacqua R, Ricci S, Gasparini G, Lorusso V, Bonizzoni E, Tonato M, on behalf of the PROTECHT Investigators. Nadroparin for the prevention of thromboembolic events in ambulatory patients with metastatic or locally advanced solid cancer receiving chemotherapy: A randomised, placebo-controlled, double-blind study. Lancet Oncol. 2009;10(10):943–9.
3. 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.
4. 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.
5. Sousou T, Khorana AA. New insights into cancer-associated thrombosis. Arterioscler Thromb Vasc Biol. 2009;29(3):316–20.
6. 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.
7. 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.
8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. Collins FS, Varmus H. A New Initiative on Precision Medicine. N Engl J Med. 2015;372(9):793–5.

14. Fayyad U, Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Database. AI magazine. 1996;17:37–54.

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. Gönen M, Alpaydın E. Multiple kernel learning algorithms. J Mach Learn Res. 2011;12:2211–68.

26. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Ann Intern Med. 1999;130(6):461–70.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. Ferroni P, Riondino S, Guadagni F, Roselli M. Impact of chemotherapy on venous thromboembolism. Haematologica. 2013;98(8):e153–e154

33. 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.

34. 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.

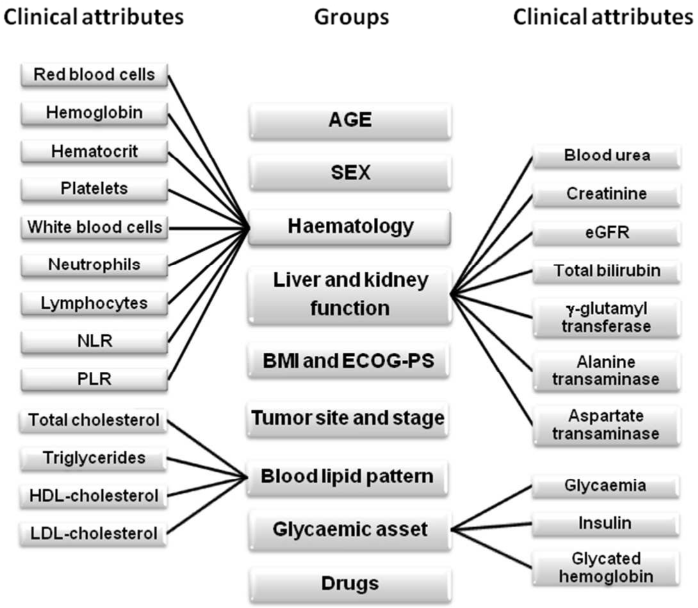
35. Bosson JL, Labarere J. Determining indications for care common to competing guidelines by using classification tree analysis: application to the prevention of venous thromboembolism in medical inpatients. Med Decis Making. 2006;26(1):63-75.

FIGURE LEGENDS

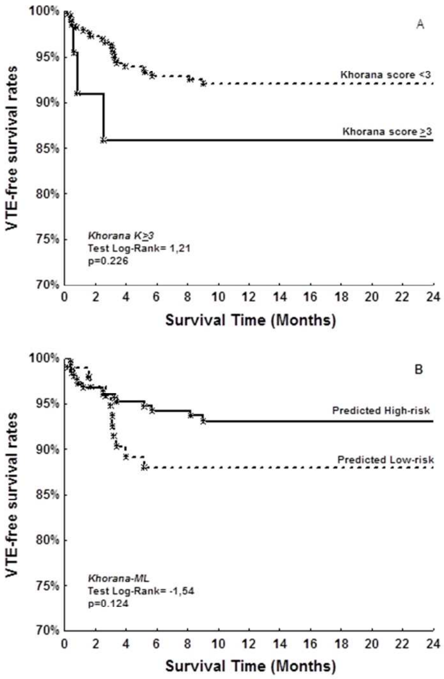
Figure 1. Groups of clinical attributes. NLR: Neutrophil/lymphocyte ratio; PLR: platelet/lymphocyte ratio; BMI: body mass index; ECOG-PS: Eastern Cooperative Oncology Group Performance Status; eGFR: estimated glomerular filtration rate. The group "Drugs" includes all supportive and anti-cancer agents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

Figure 2. 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 pure Khorana score (*Khorana k≥3*)(Panel A) or a SVM VTE event predictor using only the Khorana Score as feature (Khorana-ML).

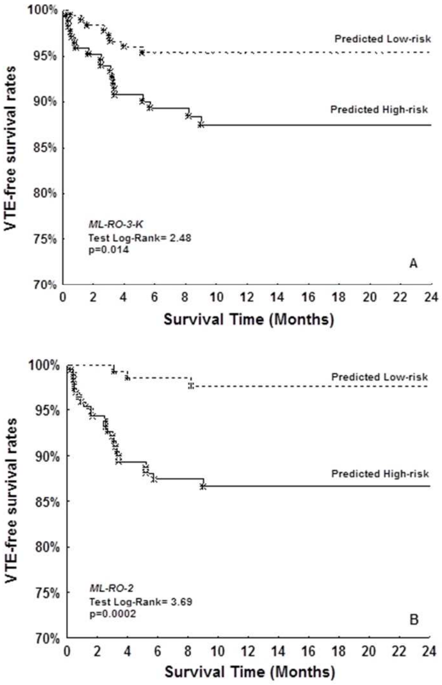
Figure 3. Kaplan–Meier curves of VTE-free survival of chemotherapy treated ambulatory cancer patients in the testing set. Comparison between patients with low (dotted line) or high (solid line) risk of VTE based on the two best fitting ML-RO models. Panel A: ML-RO-3-K. Panel B: ML+RO-2.



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 1.
254x190mm (96 x 96 DPI)



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 pure Khorana score (Khorana k≥3)(Panel A) or a SVM VTE event predictor using only the Khorana Score as feature (Khorana-ML).
254x190mm (96 x 96 DPI)



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 the two best fitting ML-RO models. Panel A: ML-RO-3-K. Panel B: ML+RO-2.
254x190mm (96 x 96 DPI)

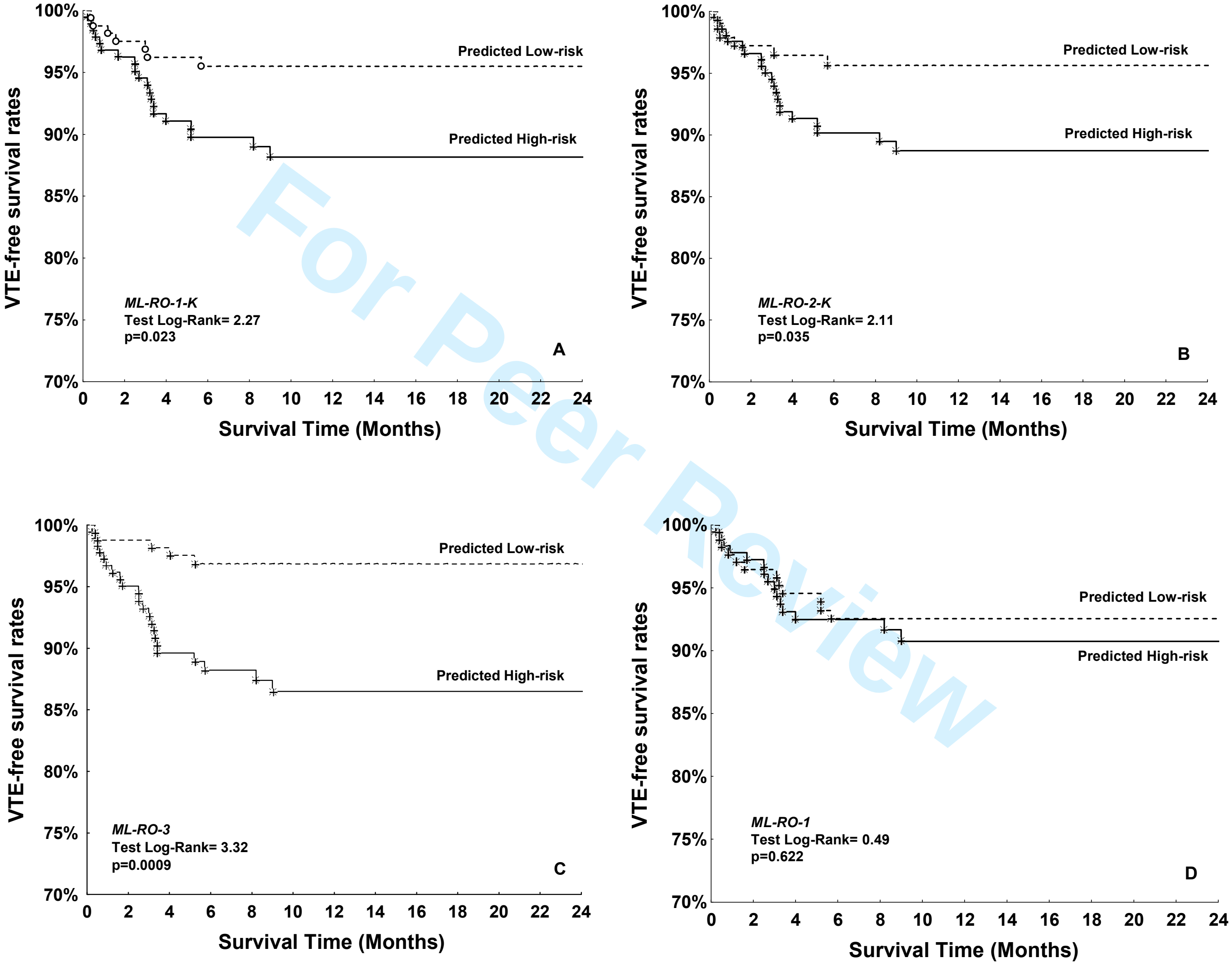
Supplementary Table 1: Inclusion and exclusion criteria

Inclusion criteria	Age >18 years
	Willingness to provide written informed consent
	Histologically confirmed diagnosis of malignancy
	Eastern Cooperative Oncology Group Performance Status (ECOG-PS) 0-2
	Absolute neutrophil count ≥2,000 mm^{-3}
	Platelet count ≥100,000 mm^{-3}
	Hemoglobin level ≥9.5 g/dl
	Bilirubin level ≤1.5x upper normal limit (UNL)
	Alanine-aminotransferase and aspartate-aminotransferase ≤2.5x UNL in the absence, or ≤5x UNL in the presence of liver metastasis
	Normal renal function (serum creatinine ≤1.2 mg/dL)
Exclusion criteria	Therapeutic doses of any heparin before enrolment
	Concomitant treatment with anticoagulant or antiplatelet drugs

Supplementary Table 2: Clinical and Laboratory attributes of the patient dataset

Age, Mean ± SD (range)	62 ± 12 (18 – 85)	**Haematology and biochemical attributes**	
Sex, N (%)			
Males	575 (49%)	**Blood cell counts**	
Females	604 (51%)	Red blood cells	4.3 ± 0.6
BMI, Mean ± SD	25.5 ± 4.5	Haematocrit	35.8 ± 9.2
ECOG-PS, N (%)		Hemoglobin	12.5 ± 1.6
0	940 (80%)	White blood cells	7.3 ± 2.9
1	228 (19%)	Neutrophils	4.9 ± 2.7
2	11 (1%)	Lymphocytes	1.7 ±0.9
Primary tumor, N (%)		Platelets	254 ± 97
Colorectal	316 (26.7%)	Mean platelet volume	8.6 ± 1.1
Gastric	53 (4.5%)	Neutrophil/lymphocyte ratio	3.9 ± 4.2
Esophageal	10 (0.9%)	Platelet/lymphocyte ratio	188.1 ± 146.3
Pancreatic	43 (3.7%)		
Biliary	18 (1.5%)	**Routine blood chemistry**	
Lung		Blood urea nitrogen	36.6 ± 15.1
Non small cell	183 (15.5%)	Creatinine	0.9 ± 0.3
Small cell	32 (2.7%)	eGFR	91.0 ± 25.6
Mesothelioma	5 (0.4%)	Glucose	112.6 ± 43.6
Breast	262 (22.2%)	Insulin	27.9 ± 32.0
Prostate	39 (3.3%)	Glycated hemoglobin	6.1 ± 3.1
Ovarian	33 (2.8%)	Total bilirubin	0.6 ± 0.5
Genitourinary	71 (6.0%)	Alanine transaminase	22.5 ± 19.3
Head-Neck	47 (4.0%)	Aspartate transaminase	22.9 ± 17.1
Sarcoma	24 (2.0)	γ-glutamyl transferase	60.7 ± 129.2
Brain	7 (0.6%)	Triglycerides	136.9 ± 76.6
Unknown	14 (1.2%)	Total cholesterol	191.9 ± 47.0
Other*	22 (1.9%)	High-density lipoproteins	47. 8 ± 14.0
Stage of disease, N (%)		Low-density lipoproteins	116.7 ± 39.8
Primary	462 (39%)		
Relapsing/metastatic	717 (61%)		
Anti-cancer drugs, N (%)**			
Platinum compounds	580 (49.2%)		
Fluoropyrimidine	453 (38.4%)		
Anthracycline	201 (17.1%)		
Taxanes	212 (18%)		
Paclitaxel	89 (7.6%)		
Bevacizumab	153 (13.0%)		
Gemcitabine	170 (14.4%)		
Irinotecane	157 (13.3%)		
Pemetrexed	77 (6.5%)		
Herceptin	59 (5.0%)		
Anti-tyrosine kinase	14 (1.2%)		
Aromatase inhibitors	22 (1.9%)		
Supportive drugs, N (%)			
Erythropoiesis stimulating agents	39 (3.3%)		
Prophylactic myeloid growth factors	65 (5.5%)		
Corticosteroids	307 (26%)		

BMI: body mass index; ECOG-PS: Eastern Cooperative Oncology Group Performance Status. eGFR: estimated glomerular filtration rate.*Including melanoma (n=10), cancer of the small intestine (n=6), neuroendocrine tumors (n=2), thymomas (n=2) and one thyroid cancer. **11% neoadjuvant, 29% adjuvant and 60% metastatic treatments.


100%
95%
90%
85%
80%
75%
70%
VTE-free survival rates
0 2 4 6 8 10 12 14 16 18 20 22 24
Survival Time (Months)
Predicted Low-risk
Predicted High-risk
ML-RO-1-K
Test Log-Rank= 2.27
p=0.023
A
ML-RO-2-K
Test Log-Rank= 2.11
p=0.035
B
ML-RO-3
Test Log-Rank= 3.32
p=0.0009
C
ML-RO-1
Test Log-Rank= 0.49
p=0.622
D