Predicting and understanding risks of venous thromboembolism to individualize prevention

Smeets, M.J.R.

Part 3

Methodological guidance for research on prediction models

Chapter 9

DO'S and DON'TS in clinical prediction research for venous thromboembolism

Banne Nemeth, Mark J.R. Smeets, Suzanne C. Cannegieter, Maarten van Smeden

Res Pract Thromb Haemost. 2024 May; 8:e102480

Abstract

Clinical prediction modeling has become an increasingly popular domain of venous thromboembolism (VTE) research in the recent years. Prediction models can help healthcare providers make decisions regarding starting or withholding therapeutic interventions, or referrals for further diagnostic workup, and can form a basis for risk stratification in clinical trials. The aim of the current guide is to assist in the practical application of complicated methodological requirements for well-performed prediction research by presenting key DO'S and DON'TS, while expanding the understanding of predictive research in general for (clinical) researchers who are not specifically trained in the topic; throughout we will use prognostic VTE scores as an exemplar.

Introduction

Clinical prediction modeling has become a popular domain of research in Venous Thromboembolism (VTE) research. Generally, a distinction is made between diagnostic and prognostic prediction models. The first type of models estimate the probability, for individual patients, of a VTE being present (but undiagnosed or neither completely ruled-in or ruled-out) at that particular point in time. Examples of these types of models are the Wells and YEARS criteria.[1,2] Prognostic models are instead used to estimate the probability that a patient will develop a VTE within a certain time frame, of which examples are the Padua Prediction and the Caprini score for hospitalized medical patients.[3,4] Besides diagnosing and predicting VTE, prediction models can help healthcare providers make decisions regarding starting or withholding prophylactic/therapeutic interventions, refer for further diagnostic workup, and can form a basis for risk stratification in clinical trials.

VTE can be prevented by administration of chemical thromboprophylaxis, which comes at a risk of (major)bleeding.[5] Therefore, exposure to chemical thromboprophylaxis is only warranted in patients in whom the risk of VTE outweighs that of bleeding. In the last decades, earlier prognostic studies have identified several prognostic factors (**table 1**) that can be used to predict VTE risk, such as immobilization, surgery, cancer and use of oral contraceptives.[6] Also, biomarkers such as D-dimer levels[7], factor VIII activity[8] and genetic predictors such as factor V Leiden mutation[9] are increasingly accessible and add valuable information to an individual's VTE risk profile. In all, the need for risk estimation in clinical practice and the availability of factors with potentially strong prognostic and diagnostic information make prediction of VTE relevant and feasible and hence an appealing approach towards better patient care.

Table 1. Prediction research terminology

Terms	Meaning
Impact study	Randomized controlled study in which the impact of a prediciton model, with subsequent intervention, are trialled.
Implementation	Implementing the prediction model in routine clinical practice.
Model coefficients	Each predictor in a prediciton model has a coefficient. This is the value by which the prognositc index (Y) of the model increases for one unit increase (or from 0 to 1 for dichotomous predictors) of the predictor.
Optimism	Meaning that the predictive performance measures are to optimistic because of overfitting.
Overfitting	Meaning that the model coefficients are too closely fitted (overfitted) on the derivation sample. This could result in a poor fit of the model in new populations.
Predictive ability of a predictor	The strength of the association between a predictor and the outcome of interest.
Predictive performance of a model	
Discriminitation	Performance measure of the prediction model which indicates its ability to distinguish patients who will develop the outcome of interst from those that will not.
Calibration	Performance measure of the prediction model which assesses how well predicted probabilities allign with observed proportions.
Predictors, prognostic factors	Variables included in the multivariable prediction model.
Updating	Changing the intercept and/or coefficients of the prediction model based on new data. Including new predictors to the model is also a form of updating.

While prediction research has become increasingly popular in recent years, several papers have shown methodological shortcomings and a lack of uniform and adequate reporting for many of these models.[10,11] Moreover, there is a lack of validation studies and, consequently, limited information about the performance of the prediction models in clinical practice.[5,12] For example, at least 13 risk prediction models for VTE for hospitalized medical patients have been published so far, which are hardly used in current clinical practice due to, possibly among other reasons, lack of good discriminative abilities, lack of proper validation studies and reluctance of healthcare providers to use such scores (e.g., due to personal beliefs).[13–15] This situation clearly illustrates the strong interest in this topic but also an overgrowth of unused prediction models whereas the primary goal of most of these models, i.e., adequate prevention of VTE in high-risk patients, is still out of reach.

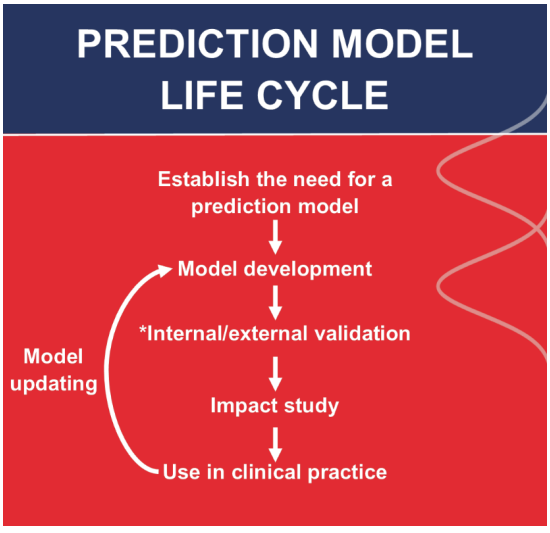
The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement was designed to improve the reporting of prediction research.[16] In addition,

the recently published PROBAST tool was designed to Assess the Risk of Bias and Applicability of Prediction Model Studies.[17] The aim of the current guide is to assist in the practical application of complicated methodological requirements for well-performed prediction research. We will do this by presenting key DO'S and DON'TS, while expanding the understanding of predictive research in general for (clinical) researchers who are not specifically trained in the topic; throughout we will use prognostic VTE scores as an exemplar.

Prediction life cycle

To develop a prediction model, a series of steps should be taken before it is ready to be used. This often means that multiple studies have to be performed (**figure 1**). Additionally, since the prediction model can be seen as a medical device, implementation (**table 1**) in clinical practice is also dependent on compliance with existing regulations.

Figure 1. Prediction model life cycle



*Both internal and external validation are part of model development. Model updating can take place at any stage after in/external validation.

1.Establish the need for a prediction model

DO	DON'T
- Establish the need for a prediction model - Define a clear purpose, target population and time horizon - Find and critically appraise other prediction models that have been developed within the same domain	- Develop a new model when similar models already exist

Just like in other types of research, the initiation of a study (in this case the development of a prediction model) starts with a hypothesis. For prediction research this hypothesis is generally that a prediction model can improve clinical decision making and thereby improve patient outcomes (e.g., deciding which hospitalized patients will benefit from thromboprophylaxis and which can go safely without). To get the aim of the prediction model clear, it is necessary to formulate the purpose to predict, the target population and to specify the time horizon. Let us assume there is a clinical need for predicting the occurrence of VTE (the purpose) in patients hospitalized in an academic center for treatment of a certain medical condition but not receiving thromboprophylaxis (the target population). It is decided that for this patient population the relevant outcome is VTE occurrence up to 3-months after the first day of hospitalization (the time horizon). For the inclusion of patients, we need to keep in mind the primary aim of the model, that is to estimate individual risks for VTE among patients that are not already on thromboprophylaxis because of, for example, a previous VTE.

Before embarking on the task to develop a new prediction model a literature search should be performed whether a similar prediction model already exists, for instance, by performing a systematic review of the prediction model landscape.[18] Clearly, developing a new prediction model defies the purpose when an existing prediction model with a similar *purpose*, *target population* and *time horizon* is already available but needs to be further evaluated and possibly refined. Authors should consider to first critically appraise and validate previously developed prediction models and, if necessary, consider to perform model updating by re-estimating/adjusting some or all of the models' coefficients or adding additional predictors (**table 1**). This way, any newly available data is used most efficiently, as new information is added to the information (based the development data and any validation studies) which is already embedded in the existing model. A testing procedure for model updating is discussed in detail elsewhere.[19]

An example of a situation in which new models were constantly added rather than building on the existing ones, is that of models aimed to identify hospitalized medical patients at high-risk for VTE. At least 13 such exist, of which only a limited number has so far received proper external validation.[13] Consequently, none of these models have been uniformly integrated in guidelines (instead, it is generally advised to assess the risks of VTE[12,20]) and hence, have not been widely implemented in clinical practice.[5,12]

2.Data availability

 DO	 DON'T
- Assess data/study type - Calculate the required sample size	- Develop a model if data/study type doesn't meet prediction aims - Develop a prediction model if sample size is unsatisfying

Do the data match the prediction aims?

The design of data collection should match the anticipated prediction aims. We consider here four groups of commonly used methods of data collection and their suitability for development of prediction models.

- *Cross sectional data collection*: can be suitable for the development of diagnostic prediction models to quantify the presence of a certain target disease at the moment of prediction but is almost never suitable for developing prognostic prediction models. For most prognostic prediction studies, it is necessary to know the value of a predictor (**table 1**) before the outcome of interest occurs. In cross sectional studies, this is not the case as both predictor and outcome are measured simultaneously. In such a study design, the value of a predictor can even be influenced by the outcome. An example in which cross sectional data can be used would be a study on the predictive value of genetic variants and the risk of VTE since the genetic variants were surely present before the VTE.
- *Case-control data collection*: Absolute risks estimates cannot directly be obtained from a case control study (unless the data are from a nested-case control study[21]), making it generally less suitable for development of prediction models. Furthermore, when data is obtained through patient questionnaires, researchers need to be aware of recall bias. Nevertheless, an important strength of case-control studies is that the number of cases is generally much higher than in any cohort design, this may allow for a larger number of potential predictors to be considered.

Hence, case-control studies can be valuable as so-called *predictor finding studies (sometimes called prognostic factor (**table 1**) studies*[22]*)* which aim to identify individual or a combination of predictors that predict the outcome of interest.

- *Prospective data collection*: Prospective cohort studies and randomized controlled trials (RCTs) are often considered preferable study types for prediction research as data is collected similarly to how this would be done following implementation. One drawback of cohort studies might be typically incompleteness of some of the variables in the data set. For example, when we look at development studies of models designed to estimate the risk of VTE recurrence, for the DASH score,[23] only 802/1818 (44%) patients had complete data, as compared with 629/929 (68%) for the Vienna model[24] and 336/646 (52%) for the HERDOO2 model.[25] This led to a substantial loss of data (lower sample size) contributing to an increased risk of overfitting and optimism (**table 1**).[26] Furthermore, limiting the cohort to complete cases only can lead to selection bias, which can affect model performance.[27] While RCT data are usually more complete than cohort studies, the in- and exclusion criteria of most RCTs can limit their generalizability.[28] In addition, RCTs are prone to large treatment effects which can greatly hamper model performance.[29] Approaches to account for treatment effects in prediction models have been discussed elsewhere.[30,31]
- *Routine healthcare data collection*: large registries such as claims databases and electronic patient record data are increasingly available and used for prediction model development. Although routine healthcare data usually come with a relatively large sample size, registry studies are prone to misclassification and missing data which also can hamper model development and may affect overall predictive performance (**table 1**) of the model.[32]

In general, the population that has been used for model development must (ideally) closely resemble the population at the point of intended application of the model. If these populations differ, it is likely that the predictive performance is compromised due to deviations in the incidence of the predicted outcome or variations in patient characteristics between both populations, known as case-mix differences.[33] A discussion about the use of datasets from multiple populations (e.g. multi-center studies) is found elsewhere.[34]

Sample size

Simplified rules of thumb (for example the Events Per Variable ratio) to calculate the minimal sample size for prediction models have been widely implemented in the literature.[35] However, these rules have been shown to be inappropriate for sample size criteria for prediction models, as they are not

based on convincing scientific reasoning[36] and perform poorly in large-scale simulation studies.[37–39] Therefore, several new sample size calculation approaches, with accompanying software to simplify the calculations, have recently been developed by leading researchers in the field.[40–42] Because these calculations take the number of candidate predictors, the total sample size and the events fraction into account, we recommended to use these new approaches.

3. Preparing the data

 DO	 DON'T
- Perform a literature study on possible candidate predictors - Be aware of predictor measurement heterogeneity - Balance practical aspects and expected predictive value - Consider performing multiple imputation	- Dichotomize predictors without absolute need - Solely include candidate predictors based on univariable association with the outcome - Exclude candidate predictors because of a non-causal relationship with the outcome

Selecting candidate predictors

Predictor variables that are considered for inclusion in the prediction model at the start of development, so-called candidate predictors, can be of various types, such as patient demographics (e.g., age and sex), patient history (e.g., heart failure, previous VTE), biomarkers (e.g., D-dimer levels) and genetic phenotypes (e.g., Factor V Leiden mutation). Ideally, candidate predictors have a strong expected incremental predictive value (i.e., a strong correlation with the outcome on top of other known predictive factors). Information about the predictive ability (**table 1**) of predictors can be obtained from predictor finding studies,[22] which comprise a large portion of the prediction modeling literature,[11] from earlier published risk prediction models, meta-analyses, and expert opinion. Predictors don't have to be causally related to the outcome of interest to be a valuable predictor. A classic example of this is that grey hair is an excellent predictor of mortality, even though there is no causal relationship. Similarly, an example related to VTE would be *socioeconomic status*, which has no obvious direct causal relation to the development of VTE, but has been shown to have value in *predicting* VTE.[43]

To select candidate predictors, one should balance between (expected) predictive value and practical aspects of measuring the predictor in the setting where the prediction model will be applied. For instance, predictors that are invasive, expensive or time-consuming to measure may

not be adequate for inclusion in a prediction model which intended point of application is a primary care setting (i.e., a thrombin generation assay might function as a good predictor but is impractical compared with a point-of-care D-dimer test). The measurement of the predictor in the study setting should also mimic that in the situation where it is applied as closely as possible to prevent *measurement heterogeneity* that may severely hamper the performance of the prediction model.[44] For instance, in developing the HERDOO2 score, a rule to guide treatment duration for women with unprovoked venous thrombosis, the authors showed that replacement of the VIDAS® D-dimer assay with other D-dimer assays decreased the predictive ability of the HERDOO2 score.[45] Another example of measurement heterogeneity applies to discrete (categorical) predictors such as heart failure, which can be classified as either type II, III or IV. One must make sure the same criteria are used during model development, validation and in practice.

Selection of candidate predictors on the basis of statistical criteria, particularly when based on the univariable association with the predicted outcome of interest, should be avoided as much as possible (variable selection is further discussed under "model development" below).

Predictor modeling

Discrete predictor variables such as patient sex, use of oral contraceptives or stoke, can be included in the prediction model as categorical dummy variables with a reference category. For instance, sex can be added as the variable "female" which takes on the value 1 for females and 0 for males (reference). Categories that are rare or categories in which the outcome does not occur can be collapsed with adjacent categories. An example of this would be genetic variants, as there are many which are associated with an (small) increased risk of VTE and the prevalence of each variant is low.[9] Hence, including every variant as separate predictor results in a model with many predictors and potential overfitting as result. In this case, when the associations between the variants and VTE are relatively similar, they can be combined in a single predictor.

Categorization of continuous predictor variables (such as patient age or D-dimer value) is generally not advised to avoid unnecessary reduction in the predictive performance of the prediction model, which can amount to the equivalent of discarding one-third of the dataset.[46] Instead, continuous predictors can be added as such in the prediction model, and possible non-linear relationship can be modelled via fractional polynomials and cubic splines (presentation of more non-linear and more complex prediction models is discussed below).[47,48] The predictive ability of a continuous or categorical predictor may vary with the values of other predictors, which can be modelled by adding

interaction terms to the prediction model. However, significant interaction terms do not necessarily improve predictive performance and considering many interaction terms can cause the model to become overfitted.[49]

Missing data

Candidate predictors that have missing values can hamper development of a prediction model. With a complete case analysis, only subjects that have complete information on each of the included predictors and outcome are included. A couple of predictors with a large number of missing values (or a large number of predictors with a small number of missing values) may contribute to a large decrease in the effective sample size in a complete case analysis. For predictors with a large number of missing values including them as candidate predictors may therefore not be worthwhile.[50] Another drawback of a complete case analysis is that it assumes the missing values to be missing completely at random, which they seldom are and which will lead to selection bias in that case.[51] Relying on a less stringent but still critical assumption of data missing at random[52,53], multiple imputation has become an increasingly popular approach to handle missing data on predictors, as it preserves the size of the original data set.

4. Model development

DO	DON'T
- Choose a suitable statistical model - Be conservative in data driven selection of predictors - Apply shrinkage - Focus on optimism corrected performance	- Perform univariable, forward or stepwise predictor selection - Use the Hosmer-Lemeshow test to evaluate calibration performance

Model building

Once all candidate predictors have been selected, and missing data have been accounted for, a prediction model can be developed. Both logistic regression and survival models are used for the prediction of binary outcomes. To decide which model should be used, researchers have to determine whether there is significant lost to follow-up (right censoring) in the data.[49] Often when the follow-up is short, lost to follow-up is negligible and logistic regression is a valid modelling approach. For example, in a previous published prediction model for VTE in the postpartum period, the authors limited follow-up to 6 weeks following delivery. In this case, a logistic regression was

performed.[54] However, when there is significant loss to follow-up, survival analysis are required.[49] In the absence of competing risks (i.e., when VTE is the outcome of interest, death is a competing risk), the Cox regression model can be used.[55] When competing risk(s) is present, the Fine-Gray model is recommended.[55,56] An example of such an approach can be find in a study from Pabinger at al. in which the authors developed and validated a prediction model for cancer-associated VTE with significant competing risk.[57] Alternative modeling approaches, such as increasingly popular machine learning techniques (e.g. as random forest, neural networks and support vector machines) have yet to show their benefit in the context of clinical prediction models.[58]

The prediction model can now be derived by applying the statistical model (or machine learning technique) on the data of all candidate predictor variables included. Before the model is finalized, variable selection is often done to reduce the number of variables in the final prediction model. While variable selection generally does not result in predictive performance benefit, models with fewer predictors can be more user friendly and practical to use in a clinical setting. However, many popular variable selection strategies, such as univariable selection, forward selection, stepwise, and backward elimination have been shown to increase the risk of model overfitting.[59–61] To reduce this risk, the general advice is to use conservative approaches to variable selection, such as backward elimination with a high p-value criterion (e.g., a p-value of 0.20).[49] Additionally, one may force some (well established) candidate predictors in the final model regardless of their predictive performance.

By applying regression shrinkage approaches (often called “regularization” in the context of machine learning) the risk of model overfitting, that occurs when the prediction model captures idiosyncrasies in the data that do not generalize to other settings, can be further reduced. In brief, shrinkage approaches introduce a small bias in the regression coefficients, generally towards the zero effect, to reduce prediction error. Several approaches to shrinkage, such uniform shrinkage, Ridge, Lasso and Firth’s correction have been suggested.[47,62,63] For more details about shrinkage we refer to Pavlou et al.[64]

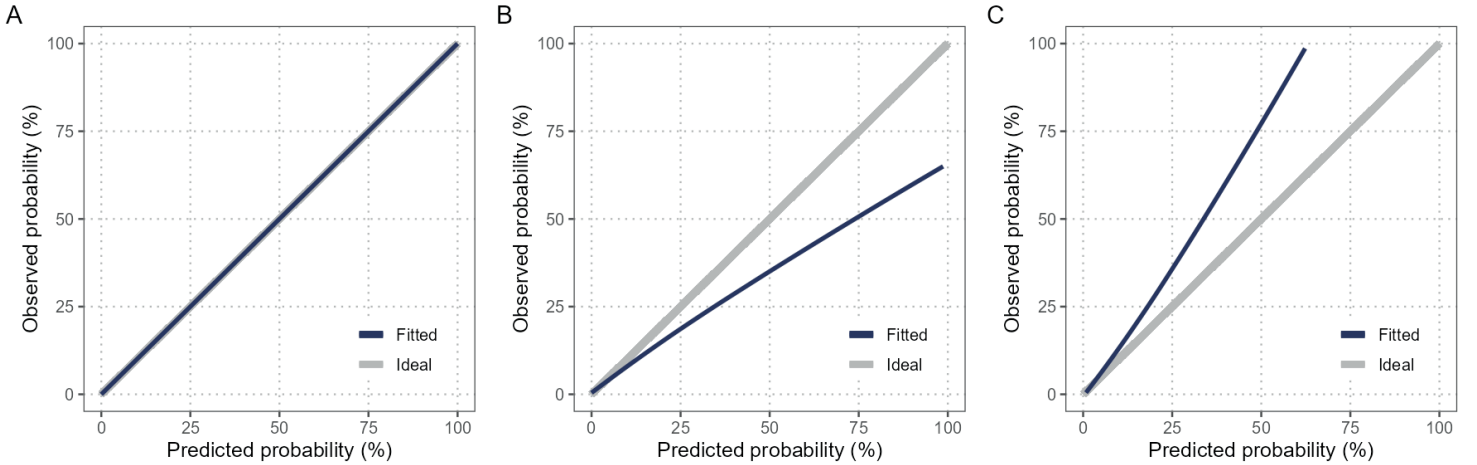
Model performance measures

For prediction models with a binary outcome, the performance of the prediction model is often expressed in terms of risk calibration, discrimination (**table 1**) and overall performance.[65]

Performance measures for prediction models with more than two categories are described elsewhere.[66]

Calibration is the ability of the prediction model to accurately estimate the risks (do *x* of 100 subjects with a predicted risk of x% truly experience the predicted event?). A common way to investigate calibration is by plotting a calibration curve that depicts the relation between estimated risks (horizontal axis) versus the observed outcome frequencies (vertical axis). In **figure 2.** we show 3 different calibration plots. Plot A depicts a model with perfect calibration. In plot B and C, a model is shown which systematically over- and underestimates the risk of patients compared to the observed risks, respectively.[67] Often, summary measures of the calibration, i.e., the calibration intercept and slope are also reported.[65] The closer the calibration intercept is to 0, the better as this means that the predicted average risk is similar to the average risk observed in the population. For the calibration slope, the optimal value is 1. However, this value can be misleading as a model can have a slope of 1, even when it systematically over- or underestimates the risks.[67] Hence, the calibration plot should always be shown. We advise against the use of the Hosmer-Lemeshow statistic for checking the calibration as it is known to have many drawbacks, including low statistical power,[68] making it an unreliable approach.

Figure 2. A) model with perfect calibration, **B**) model which overestimates the risks, **C**) model which underestimates the risks.



Discrimination is the ability of the prediction model to discriminate between individuals who develop an event versus who don't. For logistic regression models (binary outcomes), a common statistic for discrimination is the concordance probability or c-statistic (area under the ROC curve), which is determined by comparing the predicted risk of each patient in the population who developed the outcome with those who did not.[69] This c-statistic then represents the proportion of instances in which the predicted risk is higher for patients that do develop the outcome. So, when the c-statistic is 0.5 this means that the model is just as likely to assign a higher risk to patients who do and do not develop the outcome. Thus, the closer the c-statistic is to 1, the better the discrimination. In the setting of time-to-event outcomes, the concordance probability is often called the c-index.[47] This c-index is determined in a similar way as the c-statistic, with the addition that now it is also taken into account whether patients that developed the outcome earlier, have a higher predicted risk than patients that developed the outcomes later in time. Again, the closer a c-index is to 1, the better the performance of the model. Generally, reporting the c-statistic or c-index with the 95% confidence interval is sufficient and no additional information is conveyed by also reporting the ROC-curve.

Other measures of overall performance such as the Brier score, and pseudo r-square are also commonly used to quantify model performance. Measures that quantify the incremental value of predictors or compare models (e.g. the Net Reclassification Index[70]) and measures that quantify clinical usefulness (e.g. decision curve analyses[71]), are beyond the scope of this text.

5. Model validation

 DO	 DON'T
- Perform an internal validation (or internal-external) and external validation before implementation	- Perform an internal validation by random split sample - Validate the model on datasets with fewer than 100 events

When developing a prediction model, internal validation is essential to obtain valid estimates of predictive performance. This is because estimates of performance on the same data that were used to develop the prediction model tend to be too optimistic.[49] With bootstrap (repeated sampling with replacement individuals from the original dataset) or cross-validation procedures (repeated partitioning of the dataset into a larger set to develop the model and a smaller set to evaluate it),

where all the prediction modeling steps (including the selection of variables) are repeated several times, the optimism of the predictive performance measures can be investigated and "corrected" to obtain more realistic measures of performance.[49]

Model validation can also be done by splitting the data into meaningful groups. For instance, if the prediction model is derived on data obtained from multiple academic centers, the natural unit for splitting is by center. This is called internal—external validation.[72] Every center is left out once, for validation of the prediction model that is based on the remaining studies. In contrast, split-sample approaches that randomly split the data into a training (to fit the model) and test set (to validate the model), should be avoided because they are statistically inefficient and a weak test of model performance.[73,74]

It is widely acknowledged that, next to any internal validations, a prediction model needs to be assessed in external validation studies with independent data. For external validation studies, a dataset that contains at least 100 subjects that experienced the event has been recommended as a minimum sample size.[75] More precise estimates of the required sample size can be obtained with the method(s) described by Riley et al.[76] For any external validation study, the degree of relatedness between the setting where the model was developed and where it was validated should be carefully reported to facilitate the interpretation of findings.[33] Furthermore, it should be noted that a prediction model can never be conclusively "validated" and that preferably, before implementation in a specific population, the model should be validated locally.[77] Lastly, following validation, the effectiveness and safety of implementing the CPM in practice needs to be established. In such an impact study (**table 1**) design, the combination of the CPM performance (i.e., the predicted risks as provided by the CPM) and the subsequent treatment options (which are depending on the predicted risks) will be investigated.[78,79]

6. Reporting

 DO	 DON'T
- Follow the TRIPOD statement for model reporting	- Forget to report the full model parameters, including the intercept

The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement provides guidance on the key items to report when describing developing, evaluating, validating or updating (**table 1**) of clinical prediction models.[16] Numerous systematic reviews of prediction models have shown that adequate reporting is often lacking, to an extent that external validation based on the reported information is often not possible.[10,11] Following the TRIPOD reporting guideline when developing a prediction model for VTE is therefore recommended.

Reporting of the prediction model should clearly summarize all modeling steps taken to derive the model and the characteristics of the final model, including the full model equation with intercept and regression coefficients. Reporting should also provide sufficient detail on the moment and time horizon of prediction, collection of data, such as the design of data collection (see above), as well as the methods and timing of predictor and outcome measurements. Estimates of predictive performance corrected for optimism should also be reported for readers to understand the potential predictive power of the prediction model.

Further, to facilitate the implementation of prediction models, Bonnett and colleagues have summarized four methods on how prediction models can be presented[80]; using a point score system, graphical score chart, nomogram or application/website. For a detailed explanation on the advantages and disadvantages (including clinical examples) we refer to their article.[80]

Concluding remarks

In this article we described some of the key do's and don'ts when developing a prediction model. Prediction model development is only the starting point of the life cycle of a prediction model. It requires external validation studies, implementation and impact studies,[79] before it is ready to be implemented in clinical practice.

References

1 Wells PS, Hirsh J, Anderson DR, Lensing AW, Foster G, Kearon C, Weitz J, D'Ovidio R, Cogo A, Prandoni P. Accuracy of clinical assessment of deep-vein thrombosis. *Lancet* 1995; 345: 1326–30.

2 van der Hulle T, Cheung WY, Kooij S, Beenen LFM, van Bemmel T, van Es J, Faber LM, Hazelaar GM, Heringhaus C, Hofstee H, Hovens MMC, Kaasjager KAH, van Klink RCJ, Kruip MJHA, Loeffen RF, Mairuhu ATA, Middeldorp S, Nijkeuter M, van der Pol LM, Schol-Gelok S, et al. Simplified diagnostic management of suspected pulmonary embolism (the YEARS study): a prospective, multicentre, cohort study. *Lancet* 2017; 390: 289–97.

3 Barbar S, Noventa F, Rossetto V, Ferrari A, Brandolin B, Perlati M, De Bon E, Tormene D, Pagnan A, Prandoni P. A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua Prediction Score. *J Thromb Haemost* 2010; 8: 2450–7.

4 Caprini JA. Thrombosis Risk Assessment as a Guide to Quality Patient Care. *Disease-a-Month* 2005; 51: 70–8.

5 Kahn SR, Lim W, Dunn AS, Cushman M, Dentali F, Akl EA, Cook DJ, Balekian AA, Klein RC, Le H, Schulman S, Murad MH. Prevention of VTE in nonsurgical patients: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012; 141: e195S-e226S.

6 Lijfering WM, Rosendaal FR, Cannegieter SC. Risk factors for venous thrombosis - current understanding from an epidemiological point of view: Review. *British Journal of Haematology* 2010; 149: 824–33.

7 Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, Kovacs G, Mitchell M, Lewandowski B, Kovacs MJ. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. *N Engl J Med* 2003; 349: 1227–35.

8 Nemeth B, Adrichem RA van, Hylckama Vlieg A van, Bucciarelli P, Martinelli I, Baglin T, Rosendaal FR, Cessie S le, Cannegieter SC. Venous Thrombosis Risk after Cast Immobilization of the Lower Extremity: Derivation and Validation of a Clinical Prediction Score, L-TRiP(cast), in Three Population-Based Case–Control Studies. Sattar N, editor. *PLoS Med* 2015; 12: e1001899.

9 de Haan HG, Bezemer ID, Doggen CJM, Le Cessie S, Reitsma PH, Arellano AR, Tong CH, Devlin JJ, Bare LA, Rosendaal FR, Vossen CY. Multiple SNP testing improves risk prediction of first venous thrombosis. *Blood* 2012; 120: 656–63.

10 Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu L-M, Moons KG, Altman DG. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology* 2014; 14: 40.

11 Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, Moons KGM. Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLOS Medicine* Public Library of Science; 2012; 9: e1001221.

12 Schünemann HJ, Cushman M, Burnett AE, Kahn SR, Beyer-Westendorf J, Spencer FA, Rezende SM, Zakai NA, Bauer KA, Dentali F, Lansing J, Balduzzi S, Darzi A, Morgano GP, Neumann I, Nieuwlaat R, Yepes-Nuñez JJ, Zhang Y, Wiercioch W. American Society of Hematology 2018 guidelines for management of venous thromboembolism: prophylaxis for hospitalized and nonhospitalized medical patients. *Blood Adv* 2018; 2: 3198–225.

13 Cobben MRR, Nemeth B, Lijfering WM, Cannegieter SC. Validation of risk assessment models for venous thrombosis in hospitalized medical patients. *Research and Practice in Thrombosis and Haemostasis* 2019; 3: 217–25.

14 Greene MT, Spyropoulos AC, Chopra V, Grant PJ, Kaatz S, Bernstein SJ, Flanders SA. Validation of Risk Assessment Models of Venous Thromboembolism in Hospitalized Medical Patients. *Am J Med* 2016; 129: 1001.e9-1001.e18.

15 Kappen TH, van Loon K, Kappen MAM, van Wolfswinkel L, Vergouwe Y, van Klei WA, Moons KGM, Kalkman CJ. Barriers and facilitators perceived by physicians when using prediction models in practice. *Journal of Clinical Epidemiology* 2016; 70: 136–45.
16 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; 350: g7594–g7594.
17 Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* American College of Physicians; 2019; 170: W1–33.
18 Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, Riley RD, Moons KGM. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017; 356: i6460.
19 Vergouwe Y, Nieboer D, Oostenbrink R, Debray TPA, Murray GD, Kattan MW, Koffijberg H, Moons KGM, Steyerberg EW. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med* 2017; 36: 4529–39.
20 Venous thromboembolism in over 16s: reducing the risk of hospital-acquired deep vein thrombosis or pulmonary embolism | Guidance | NICE. NICE; 2018.
21 Haaf KT, Steyerberg EW. Methods for individualized assessment of absolute risk in case-control studies should be weighted carefully. *Eur J Epidemiol* 2016; 31: 1067–8.
22 Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, Malats N, Briggs A, Schroter S, Altman DG, Hemingway H, for the PROGRESS Group. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med* 2013; 10: e1001380.
23 Tosetto A, Iorio A, Marcucci M, Baglin T, Cushman M, Eichinger S, Palareti G, Poli D, Tait RC, Douketis J. Predicting disease recurrence in patients with previous unprovoked venous thromboembolism: a proposed prediction score (DASH). *J Thromb Haemost* 2012; 10: 1019–25.
24 Eichinger S, Heinze G, Jandeck LM, Kyrle PA. Risk assessment of recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism: the Vienna prediction model. *Circulation* 2010; 121: 1630–6.
25 Rodger MA, Kahn SR, Wells PS, Anderson DA, Chagnon I, Le Gal G, Solymoss S, Crowther M, Perrier A, White R, Vickars L, Ramsay T, Betancourt MT, Kovacs MJ. Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy. *CMAJ* 2008; 179: 417–26.
26 Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol* 2003; 56: 441–7.
27 Gorelick MH. Bias arising from missing data in predictive models. *J Clin Epidemiol* 2006; 59: 1115–23.
28 Tzoulaki I, Siontis KCM, Ioannidis JPA. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. *BMJ* 2011; 343: d6829.
29 Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ* 2018; 363: k4245.
30 Groenwold RHH, Moons KGM, Pajouheshnia R, Altman DG, Collins GS, Debray TPA, Reitsma JB, Riley RD, Peelen LM. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *J Clin Epidemiol* 2016; 78: 90–100.
31 Sperrin M, Martin GP, Pate A, Van Staa T, Peek N, Buchan I. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Stat Med* 2018; 37: 4142–54.
32 Thygesen LC, Ersbøll AK. When the entire population is the sample: strengths and limitations in register-based epidemiology. *Eur J Epidemiol* 2014; 29: 551–8.

33 Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology* 2015; 68: 279–89.

34 Wynants L, Kent DM, Timmerman D, Lundquist CM, Van Calster B. Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. *Diagn Progn Res* 2019; 3: 6.

35 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; 49: 1373–9.

36 Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012; 98: 683–90.

37 van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, Reitsma JB. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology* 2016; 16: 163.

38 Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol* 2011; 64: 993–1000.

39 Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol* 2016; 76: 175–82.

40 van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, Reitsma JB. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res* SAGE Publications Ltd STM; 2019; 28: 2455–74.

41 Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine* 2019; 38: 1276–96.

42 Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, Collins GS. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Stat Med* 2019; 38: 1262–75.

43 Kort D, van Rein N, van der Meer FJM, Vermaas HW, Wiersma N, Cannegieter SC, Lijfering WM. Relationship between neighborhood socioeconomic status and venous thromboembolism: results from a population-based study. *J Thromb Haemost* 2017; 15: 2352–60.

44 Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol* 2019; 105: 136–41.

45 Rodger MA, Le Gal G, Langlois NJ, Gin B, Mallick R, Giulivi A, Freedman M, Kovacs MJ, REVERSE II investigators. "HERDOO2" clinical decision rule to guide duration of anticoagulation in women with unprovoked venous thromboembolism. Can I use any d-Dimer? *Thromb Res* 2018; 169: 82–6.

46 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006; 25: 127–41.

47 Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, NY: Springer; 2001.

48 Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15: 361–87.

49 Steyerberg EW. Clinical Prediction Models. New York, NY: Springer; 2009.

50 Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009; 338: b604–b604.

51 Heymans MW, Twisk JWR. Handling missing data in clinical research. *Journal of Clinical Epidemiology* 2022; 151: 185–8.

52 Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology* 2010; 63: 205–14.

53 Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* British Medical Journal Publishing Group; 2009; 338: b2393.

54 Sultan AA, West J, Grainge MJ, Riley RD, Tata LJ, Stephansson O, Fleming KM, Nelson-Piercy C, Ludvigsson JF. Development and validation of risk prediction model for venous thromboembolism in postpartum women: multinational cohort study. *BMJ* British Medical Journal Publishing Group; 2016; 355: i6253.

55 Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation* 2016; 133: 601–9.

56 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* Taylor & Francis; 1999; 94: 496–509.

57 Pabinger I, van Es N, Heinze G, Posch F, Riedl J, Reitter E-M, Di Nisio M, Cesarman-Maus G, Kraaijpoel N, Zielinski CC, Büller HR, Ay C. A clinical prediction model for cancer-associated venous thromboembolism: a development and validation study in two independent prospective cohorts. *The Lancet Haematology* 2018; 5: e289–98.

58 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* 2019; 110: 12–22.

59 Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol* 1999; 52: 935–42.

60 Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001; 21: 45–56.

61 Smith G. Step away from stepwise. *Journal of Big Data* 2018; 5: 32.

62 Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat Med* 2017; 36: 2302–17.

63 Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med* 2016; 35: 1159–77.

64 Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, Omar RZ. How to develop a more accurate risk prediction model when there are few events. *BMJ* British Medical Journal Publishing Group; 2015; 351: h3868.

65 Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology* 2010; 21: 128–38.

66 de Jong VMT, Eijkemans MJC, van Calster B, Timmerman D, Moons KGM, Steyerberg EW, van Smeden M. Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in Medicine* 2019; 38: 1601–19.

67 On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; 17: 230.

68 Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology* 2016; 74: 167–76.

69 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29–36.

70 Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest* 2012; 42: 216–28.

71 Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, Roobol MJ, Steyerberg EW. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *Eur Urol* 2018; 74: 796–804.

72 Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016; 69: 245–7.

73 Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; 54: 774–81.

74 Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res* 2017; 26: 796–808.

75 Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005; 58: 475–83.

76 Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, Snell KIE. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine* 2021; 40: 4230–51.

77 Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Medicine* 2023; 21: 70.

78 Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; 98: 691–8.

79 Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and Prognostic Research* 2018; 2: 11.

80 Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* 2019; 365: l737.