Journal of Clinical Epidemiology 105 (2019) 136–141

Journal of Clinical Epidemiology

COMMENTARY

How variation in predictor measurement affects the discriminative ability and transportability of a prediction model

R. Pajouheshnia[a,*], M. van Smeden[b], L.M. Peelen[a], R.H.H. Groenwold[b]

[a]*UMC Utrecht Julius Center, Utrecht University, Utrecht, The Netherlands*
[b]*Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands*

Accepted 10 September 2018; Published online 14 September 2018

Abstract

Background and Objective: Diagnostic and prognostic prediction models often perform poorly when externally validated. We investigate how differences in the measurement of predictors across settings affect the discriminative power and transportability of a prediction model.

Methods: Differences in predictor measurement between data sets can be described formally using a measurement error taxonomy. Using this taxonomy, we derive an expression relating variation in the measurement of a continuous predictor to the area under the receiver operating characteristic curve (AUC) of a logistic regression prediction model. This expression is used to demonstrate how variation in measurements across settings affects the out-of-sample discriminative ability of a prediction model. We illustrate these findings with a diagnostic prediction model using example data of patients suspected of having deep venous thrombosis.

Results: When a predictor, such as D-dimer, is measured with more noise in one setting compared to another, which we conceptualize as a difference in ''classical'' measurement error, the expected value of the AUC decreases. In contrast, constant, ''structural'' measurement error does not impact on the AUC of a logistic regression model, provided the magnitude of the error is the same among cases and noncases. As the differences in measurement methods between settings (and in turn differences in measurement error structures) become more complex, it becomes increasingly difficult to predict how the AUC will differ between settings.

Conclusion: When a prediction model is applied to a different setting to the one in which it was developed, its discriminative ability can decrease or even increase if the magnitude or structure of the errors in predictor measurements differ between the two settings. This provides an important starting point for researchers to better understand how differences in measurement methods can affect the performance of a prediction model when externally validating or implementing it in practice.

Keywords: Prediction models; Measurement error; Discrimination; Area under the curve; Transportability

1. Introduction

Before prediction models are implemented in clinical practice, they should be externally validated, i.e., tested in individuals who were not a part of the data set used to develop the model [1–4]. Ideally, a model should perform well (in terms of its discriminative ability and calibration [5]) when validated in new sets of patients from different settings, e.g., from different clinical settings, geographical locations, or time periods. However, prediction models commonly perform differently—generally poorer—in new settings compared to what was observed in the development data set [6]. We then say that the transportability of the prediction model is low. Notably, failure for a model to transport well across settings indicates that the model cannot be readily implemented for new individuals [7]. Therefore, it is important that we understand what causes a prediction model to perform differently across settings. Discussions about variation in performance across data sets often focus on differences in patient characteristics [8–10]. Herein, we argue that variation in prediction model performance can also be explained (in part) by differences in how predictors are measured across settings, regardless of whether patient characteristics are similar or different.

The way that predictors are measured often varies from the development setting to validation or implementation settings. This occurs when predictor values are determined using different methodologies, protocols (e.g., fasting vs. nonfasting cholesterol measurements), or equipment, are

Declaration of funding and competing interests: This work was funded by the Netherlands Organisation for Scientific Research (project 9120.8004 and 918.10.615). Rolf Groenwold receives funding from the Netherlands Organisation for Scientific Research (project 917.16.430). The funding bodies had no role in the design, conduct or decision to publish this study, and there are no conflicts of interest to declare.

* Corresponding author. UMC Utrecht Julius Center, Utrecht University, Utrecht 3508 GA, The Netherlands. Tel.: 088 75 681 81; fax: 088 75 680 99.
E-mail address: R.Pajouheshnia@umcutrecht.nl (R. Pajouheshnia).

https://doi.org/10.1016/j.jclinepi.2018.09.001
0895-4356/

What is new?

Key findings

- Variation in the way a predictor is measured across settings can lead to a decrease or even an increase in the discriminative ability of a prediction model across those settings.

What this adds to what was known?

- Differences in the methods used to measure a predictor can be described in terms of differences in the magnitude or structure of predictor measurement errors.
- In expectation, an increase in random error in the measurement of a continuous predictor from one setting to another will result in a decrease in the area under the ROC curve, when predicting a binary outcome.

What is the implication and what should change now?

- Future discussions about variation in prediction model performance across settings should also consider variation in predictors' measurements.

measured by different people with varying levels of training, or are directly measured in one setting and measured by patient recall in a different setting, for example. Surrogate values for predictors may also be used when measurements of a certain predictor are unavailable in a data set [11]. Altogether, this can have a large impact on the value of measurements for individual patients; the value of a blood pressure reading, for example, is known to vary greatly depending on how the measurement is taken [12]. Therefore, we can expect that differences in the distribution of predictor values across different studies are not only due to true variation in the characteristics of patients but also the ways that their characteristics were measured.

In this report, we describe how the discriminative ability of a prediction model varies across settings with variation in the measurement of predictors and illustrate the effect both numerically and in a case study about a diagnostic prediction model for deep venous thrombosis (DVT).

2. How the AUC is related to measurements of a predictor

2.1. Relating the AUC to the distribution of a continuous predictor

The area under the receiver operating characteristic curve (AUC) indicates how well a prediction model can discriminate between individuals who have/will have (cases) or do not have/will not have (noncases) the health outcome of interest (e.g., disease or health state) [13]. Assuming a continuous predictor (which may be a linear combination of several predictors) follows a normal distribution among cases and noncases separately, it has been shown that the AUC of a predictor of a binary outcome can be approximated by [14]:

$$\mathrm{AUC} = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \quad [1]$$

Here μ_1, μ_0, σ_1^2 and σ_0^2 refer to the means and variances of the predictor in the cases (μ_1, σ_1^2) and noncases (μ_o, σ_0^2), and Φ denotes the cumulative normal distribution function. From [1] we can see that the AUC is a function of the mean and variance of the values of a predictor that are observed for cases and noncases.

2.2. Relating the AUC to the distribution of a predictor measured with error

To understand how the measurement of a predictor can affect the AUC of a prediction model, we turn to an existing taxonomy for measurement error (for further detail see [15–17]). First, consider a candidate predictor, for example, height. In one sample, the height of patients is measured directly by a research assistant, providing an accurate measure of heights in the sample. In another sample, height is self-reported. Given that patients are likely to recall their height with a certain amount of error, the self-reported height value observed for an individual i (W_i) represents their accurately measured height (X_i) plus some additional error (U_i) [15]:

$$W_i = X_i + U_i \quad [2]$$

A common model for U_i is the "classical" measurement error model where U follows a normal distribution with a mean value of zero and a (constant) variance, τ. Under this model, the error in self-reported height is considered random, and on average the measurements are unbiased (E (W) = E (X)) but have additional variance, such that the expected variance of W is equal to the sum of σ^2 (the variance of the accurately measured predictor X) and τ. It follows that the expected value of the AUC of the predictor in the sample, measured with random error, is:

$$\mathrm{AUC} = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2 + 2\tau}}\right) \quad [3]$$

From this, we can see that as the amount of random error with which a predictor is measured increases, the discriminative ability of the predictor is expected to decrease, provided the other parameters in [3] remain constant.

2.3. A general expression relating measurement error to the AUC of a continuous predictor

Measurement error can also affect the mean of the observed values, which may also vary between cases and noncases. Expression [2] can be extended as such:

$$W_i = \Psi_y + X_i\theta_y + \varepsilon_i \quad [4]$$

where y is an indicator to distinguish between cases (y = 1) and noncases (y = 0). Furthermore, we assume that $\varepsilon|y \sim N(0, \tau y)$. The mean and variance components of the observed predictor values in the cases and noncases can now be defined. Let the expected values of the predictor X be defined as $\overline{X}_0 = E(X|Y = 0), \overline{X}_1 = E(X|Y = 1)$, and similarly, the values for error-contaminated predictor W, be defined as $\overline{W}_0 = E(W|Y = 0), \overline{W}_1 = E(W|Y = 1)$. Hence, in expectation,

$$\overline{W}_1 = \Psi_1 + \theta_1 X_1 \quad [5]$$

$$\overline{W}_0 = \Psi_0 + \theta_0 X_0 \quad [6]$$

$$\Sigma_1^2 = \sigma_1^2\theta_1^2 + \tau_1 \quad [7]$$

$$\Sigma_0^2 = \sigma_0^2\theta_0^2 + \tau_0 \quad [8]$$

It follows that, under the same conditions required for expression [1] to hold, this expression can be extended to incorporate measurement error by substituting the means and variances of predictor X in the cases and noncases for values of predictor W (measured with error), such that

$$AUC = \Phi\left(\frac{\overline{W}_1 - \overline{W}_0}{\sqrt{\Sigma_1^2 + \Sigma_0^2}}\right) \quad [9]$$

3. Differences in measurement error between settings lead to changes in the AUC

As explained, the way that predictors are measured in samples from different settings varies, which could result in variation in measurement error. Notably, it follows from expression [9] that if differences in measurement between settings translate to differences in the structure or magnitude of the measurement error associated with the predictors, the AUC can vary across these settings. Figure 1 (scenarios S1 and S2) shows that when the amount of random error across settings increases, the expected value of the AUC decreases. In contrast, depending on the direction of the error, and whether it is present equally in the measurements of both cases and noncases, nonrandom error can cause the AUC to increase, decrease or it may have little effect. Therefore, we see that for the discriminative ability of a predictor to transport to a new setting, the measurement error of that predictor must also be transportable. Furthermore, in this example, the mean and variance of the predictors in the absence of measurement error remained constant across scenarios. In reality, it becomes extremely challenging to predict how the AUC will change across settings because the AUC is a function of predictor means, variances, and error, all of which can change between settings.

4. Case study: differences in measurements from development to validation

Data from 1,295 patients with possible DVT [18] were used to examine how differences in the measurement of a predictor across samples affect the discriminative ability of a diagnostic prediction model. First, a model to predict the presence of DVT was developed with a single predictor, D-dimer (log-transformed, continuous measurements of the biomarker were used), using logistic regression on a random half of the data (a split-sample procedure for illustration purposes). Characteristics of the population in each half of the data were on average very similar and closely resembled the full data set (see [18] for details). Next, we explored how measuring D-dimer with greater error in a validation sample could affect its discriminative power. Error of increasing magnitude was simulated and added to the D-dimer measurements in the remaining half of the data, and subsequently the AUC was calculated. The AUCs reported in Figure 2 denote the average AUC after replicating the entire procedure (from data splitting to calculating of the AUC) 1,000 times. Figure 2 shows that when measurements were conducted less accurately (i.e., with increasing amounts of noise or "classical error", see Section 3), there was greater overlap between the distributions of the predictor values of the cases and noncases in the validation sample. This translated to a strong reduction of the AUC, from 0.89 in the development sample to 0.67 in the validation sample with a 200% increase in log D-dimer variance relative to the actual variance. In contrast, a fixed increase in the D-dimer measurements (constant "structural error", see Section 3) caused a uniform shift in the predictor values, and thus the AUC remained unchanged with increasing error, as was also seen in Figure 1, scenario 3. Finally, the prediction model including D-dimer was extended with additional predictors: sex, oral contraceptive use, presence of a malignancy, recent surgery, absence of leg trauma, vein distension, and difference between calf circumferences, to reflect a published diagnostic model [18]. All predictors except D-dimer were assumed to be measured in the same fashion in the development and validation samples. The same trends were observed as in the univariable (D-dimer only) model; the AUC ranged from 0.90 in the development set to 0.70 in the validation sample with a 200% increase in log D-dimer variance relative to the actual variance and remained stable with fixed increases in the D-dimer measurements.

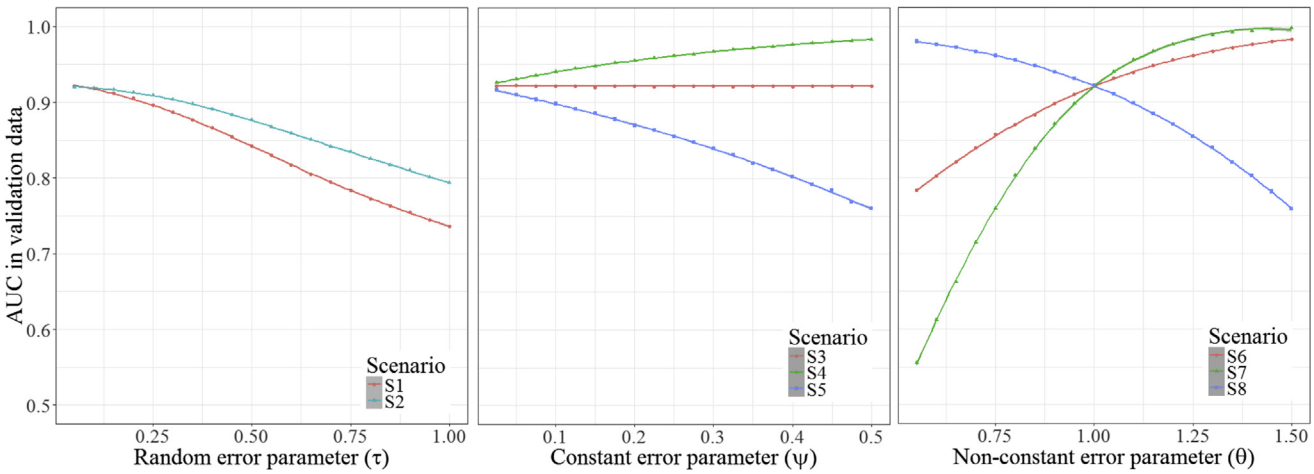

Fig. 1. The effect of measurement error in a single continuous predictor on the AUC when predicting a binary outcome. Data of size n = 10^5 were simulated such that the values of a continuous predictor X, measured without error, were normally distributed for noncases ($X_0 \sim N[1, 0.5^2]$) and cases ($X_1 \sim N[2, 0.5^2]$). Measurement error was simulated on top of the error-free measurements according to expression [4], such that values, W, were observed instead of the true values of X, and the AUC was estimated using expression [9]. Each point on the curves represents a sample with a certain amount and type of measurement error. The horizontal axes represent the measurement error parameter values used to vary the observed values, W, using expressions [5–8]. In scenario S1, random (classical) error was added by increasing the random error term τ. In scenario S2, random error was only added to the cases (τ_1). In scenarios S3–S5, a constant (structural) error value ψ, was added to the measurements to both cases and noncases (S3), cases only (ψ_1) (S4), and noncases only (ψ_0) (S5), respectively. In scenarios S6–S8, error proportional to the true value of X, θ, was added to the cases and noncases (S6), cases only (θ_1) (S7), and noncases only (θ_0) (S8), respectively.

5. Concluding remarks

Differences in the way predictors are measured across settings can cause the discriminative ability of a prediction model to appear to be worse, but perhaps surprisingly can appear to be better as well, in one setting compared to another. Thus, for a prediction model to transport well to new patient samples, i.e., from development, to validation, and finally implementation in daily practice, measurements methods should be comparable across each sample. We propose that differences in the way a predictor is measured across samples from different settings can be viewed in the context of measurement error. Prediction does not require the "true" value of a variable to be measured rather predictions are made using observed measurements [15]. Whether predictor measurements deviate from their "true" (e.g., biological) value becomes important only if this deviation from "truth" varies from one sample to another.

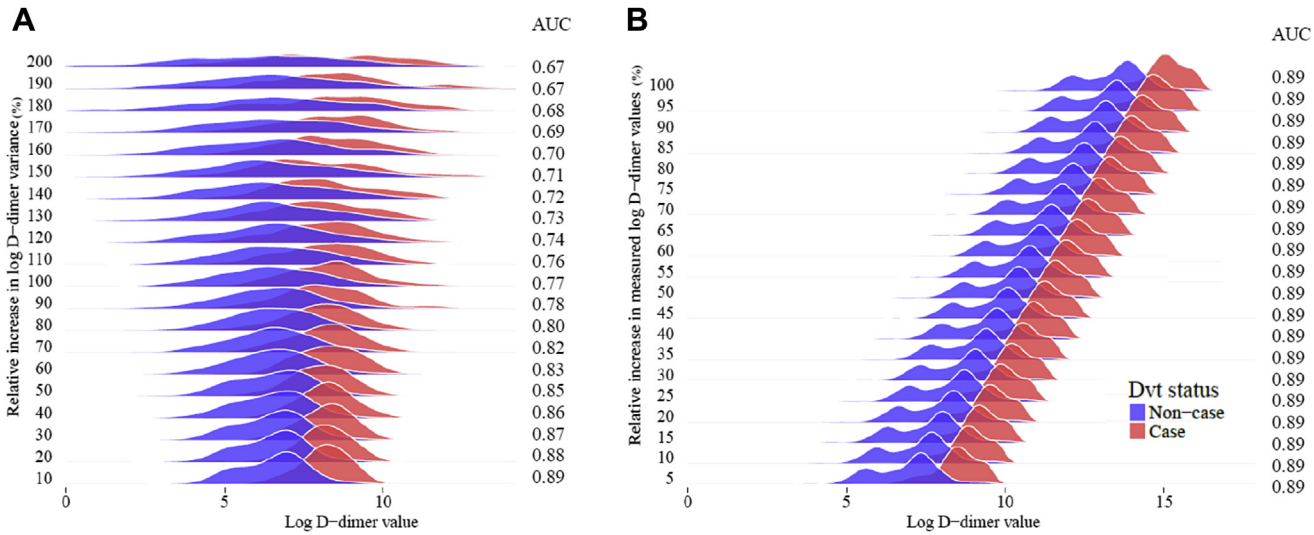

Fig. 2. (A–B) The transportability of the discriminative value of D-dimer for separating DVT cases from noncases, from the development sample to validation samples with increasing amounts of measurement error. The DVT data were randomly split 50:50, simulated error was added to the DVT values measured in the validation sample, and the AUC was calculated. The process was repeated 1,000 times, and the average of the AUCs was calculated. A: random (classical) error, where the error was randomly sampled from a normal distribution ($\varepsilon \sim N(0, \sigma_{D\text{-dimer}} * m)$, and m ranged from 0.1 to 2. B: constant (structural) error, such that $\varepsilon = \widehat{X}_{D-dimer} + m$, and m ranged from 0.05 to 1.

A number of studies have investigated factors that influence the performance of a prediction model. The effect of measurement error on prediction model performance within a single sample has been examined elsewhere [19,20]. Others have investigated how correlation between predictors in a model is related to model performance. A simulation study by Kundu et al. [21] found that differences in the correlations between predictors in validation samples compared to the development sample can result in differences in the AUC. Given that differences in measurement methods can affect the variance of predictor measurement in a sample, and the correlation between two variables is a function of their variances and covariance, our findings explain how differences in correlations can arise between samples, and how this can affect the AUC of a model.

There are several implications of the effect that variation in predictor measurement (and consequent differences in measurement error) has on model performance across settings. First, differences in performance can arise when predictors in a validation sample have been measured using methods that do not reflect current standards. This could happen if the validation sample comes from an historical cohort, in which outdated, more error-prone methods of measurement were used. Alternatively, if data were collected in a highly protocol-driven setting, such as for a randomized trial, measurements may be more precise (with less error) than in real practice. In such cases, evidence of poor model transportability is weakened by nonrepresentative measurements in the validation sample. Second, differences in measurements from development to validation might reflect true variation in clinical practice. In this case, we might conclude that poor performance in a validation sample is evidence of limited transportability of the model. Finally, it could be that the model itself is outdated, and since its development, the methods used to measure a predictor have improved. Poor performance in a contemporary validation sample would indicate that the model requires updating.

We present a starting point for the further exploration of the impact of differences in measurements on prediction model performance, and further attention is required in a number of areas. First, the mathematical relationship we present between measurement error and model discrimination is restricted to the case of a single continuous predictor and is based on strict assumptions. Future research could use computer simulations to further explore the impact of differences in measurements on the discriminative ability of multivariable prediction models across samples and could investigate the misclassification of categorical predictors. Second, although we have discussed model discrimination, model calibration requires separate attention. Khudyakov et al. demonstrated that measurement error in a predictor does not affect the calibration of a prediction model within the same sample [19]. However, differences in how a predictor has been measured across settings could negatively impact on the calibration of the model across settings. To examine this would require extensive simulations and is beyond the scope of this study. Third, in our case study, we do not consider that the variables (e.g., D-dimer) are likely to have already been measured with an amount of error, before the addition of simulated error. Given that we randomly split our data in development and validation sets, any existing error should have been similar across the sets, and thus the findings should not be affected. Fourth, we consider the effect of differences in predictor measurement in isolation from other factors that influence the discriminative ability of a prediction model. Variation in population characteristics across settings (or ‘‘patient spectrum’’), for example, has been shown to influence the performance and transportability of a diagnostic test or prediction model [10,22,23]. Further research is needed to determine the relative contribution of variation in the measurement of predictors to overall prediction model performance, compared with other aspects of the characteristics of a population. Finally, we do not comment on the use of correction methods for measurement error when developing or validating a prediction model, as this remains a topic for further investigation.

To conclude, if the measurement of predictors varies from sample to sample, we can anticipate changes in the discriminative ability of the model. Discussions about variation in prediction model performance across settings should therefore also consider variation in predictors’ measurements. The way predictors are measured when developing of validating a prediction model should mimic the way predictors are measured in practice to obtain realistic and relevant estimates of prediction model performance.

References

[1] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):e1001381.
[2] Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.
[3] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453−73.
[4] McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users’ guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-based medicine working group. JAMA 2000;284:79−84.
[5] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer Science & Business Media; 2008.
[6] Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015;68:25−34.
[7] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515−24.
[8] Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010;172:971−80.
[9] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68:279−89.

[10] Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ 2016;353:i3139.
[11] Smith T, Muller DC, Moons KGM, Cross AJ, Johansson M, Ferrari P, et al. Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies. Gut 2018; https://doi.org/10.1136/gutjnl-2017-315730.
[12] Handler J. The importance of accurate blood pressure measurement. Perm J 2009;13(3):51−4.
[13] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29−36.
[14] Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol 2012;12:82.
[15] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. 2 ed. New York: Chapman ⧹ & Hall/CRC Press; 2006.
[16] Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. New York: Chapman and Hall/CRC; 2004.
[17] Buonaccorsi J. Measurement error: models, methods and applications. New York: Chapman and Hall/CRC; 2010.
[18] Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005;94(1):200−5.
[19] Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. Stat Med 2015;34:2353−67.
[20] Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. Popul Health Metr 2012;10(1):20.
[21] Kundu S, Mazumdar M, Ferket B. Impact of correlation of predictors on discrimination of risk models in development and external populations. BMC Med Res Methodol 2017;17:63.
[22] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926−30.
[23] André Knottnerus J, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol 1992;45:1143−54.