Annals of Clinical Epidemiology 2022;4(3):72–80

Annals of Clinical Epidemiology

SEMINAR

Introduction to Clinical Prediction Models

Masao Iwagami[1,2], Hiroki Matsui[3]

[1] *Department of Health Services Research, Faculty of Medicine, University of Tsukuba*

[2] *Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine*

[3] *Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo*

ABSTRACT

Clinical prediction models include a diagnostic prediction model to estimate the probability of an individual currently having a disease (e.g., pulmonary embolism) and a prognostic prediction model to estimate the probability of an individual developing a specific health outcome over a specific time period (e.g., myocardial infarction and stroke in 10 years). Clinical prediction models can be developed by applying traditional regression models (e.g., logistic and Cox regression models) or emerging machine learning models to real-world data, such as electronic health records and administrative claims data. For derivation, researchers select candidate variables based on a literature review and clinical knowledge, and predictor variables in the final model based on pre-defined criteria (e.g., thresholds for the size of relative risk and p-values) or strategies such as the stepwise regression and the least absolute shrinkage and selection operator (LASSO) regression. For validation, the clinical prediction model's performance is evaluated in terms of goodness of fit (e.g., R^2), discrimination (e.g., area under the receiver operating characteristic curve or c-statistics), and calibration (e.g., calibration plot and Hosmer-Lemeshow test). Performance of a new variable added to an existing clinical prediction model is evaluated in terms of reclassification (e.g., net reclassification improvement and integrated discrimination improvement). The model should be validated using the original data to examine internal validity through methods such as resampling (e.g., cross-validation and bootstrapping) and using other participants' data to examine external validity. For successful implementation of a clinical prediction model in actual clinical practice, presentation methods such as paper-based (nomogram) or web-based calculator and an easy-to-use risk score should be considered.

KEY WORDS

risk score, derivation, validation, regression, machine learning

1. INTRODUCTION

A clinical prediction model, also called as a prediction rule or risk score, aims to predict a specific health condition or disease for each individual [1]. Clinical prediction models include diagnostic and prognostic prediction models [2].

Diagnostic prediction models aim to estimate the probability of an individual currently having a specific health condition (often a disease) [2], such as pulmonary embolism (PE). For example, according to the Wells' criteria for PE [3, 4], a patient who has malignancy on treatment (1 point), is visiting an emergency department with clinical signs and symptoms of deep venous thrombosis (3 points) and hemoptysis (1 point), and has increased pulse rate of >100/min (1.5 points) is calculated to have a risk score of 6.5, categorized as being at "high risk" for PE. Thus, in addition to D-dimer measurement, this patient should receive more expensive and invasive examinations, such as contrast-enhanced computed tomography, ventilation-perfusion lung scanning, and right heart catheterization, to confirm the diagnosis of PE.

Prognostic prediction models aim to estimate an individual's probability of developing a specific health outcome over a specific time period, such as myocardial infarction (MI) and stroke. For example, according to the QRISK-3 risk calculator in UK primary care (https://

qrisk.org/three/index.php) [5], an individual who is a 55-year-old white male with type-2 diabetes, is a current heavy smoker, and is of height 170 cm and weight 70 kg is estimated to have a 21.1% risk of having MI or stroke within the next 10 years. Thus, this patient should be advised to stop smoking and start statins, in line with guidance from the National Institute for Health and Care Excellence, which recommends starting statins for patients with ≥10% risk from the viewpoints of both benefit-risk balance and cost-effectiveness [6].

As seen in these successful examples, clinical prediction models are useful to identify people at high risk for a certain outcome, and to intervene them efficiently. Clinical prediction models can be developed using routinely collected clinical data or real-world data. However, to establish a valid prediction model, researchers need to account for several knacks and pitfalls. This paper introduces ways to develop and validate a clinical prediction model.

2. PREPARATION OF STUDIED DATA

Similar to any observational study, data obtained from the study population of interest (e.g., a group of patients visiting hospitals with a certain condition [3, 4] or the general population with or without diseases [5]) are needed to develop and validate a clinical prediction model. Primary data collection [3, 4] or secondary use of existing data—such as data from electronic health records [5], administrative claims [7], registries [8], biobanks [9], and clinical trials [10]—are possible.

2.1 Variables in the Dataset

The dataset needs to contain an outcome variable (also called a dependent variable) as the reference standard, as well as predictor variables (also called predictors or independent variables/parameters/values) for individual participants. The outcome variable is often a binary variable (i.e., presence or absence of an outcome) or a time-to-event outcome (i.e., presence or absence of an outcome, in addition to the follow-up time until the incidence of the outcome or end of follow-up); it could also be a continuous or multi-categorical (nominal or ordinal) variable. Predictor variables may be continuous or categorical. Continuous variables should ideally be kept continuous, while applying linear or non-linear (e.g., fractional polynomial or spline) functions, depending on their relationship with the outcome. They can be categorized using cut points based on the model's predicted probabilities or risks, although dichotomization of these variables to optimize p-values is discouraged [11]. As there might be missing data for some variables, an appropriate strategy to deal with such missing data is needed during analyses [12]. Missing data and multiple imputation were featured in a previous paper in this seminar series [13].

If the original data source includes a large number of variables, which may or may not contribute to a clinical prediction model, researchers may select candidate variables (i.e., potentially relevant predictors) for model development based on a literature search of known risk factors and/or clinical knowledge in the field. For example, to develop the QRISK-3, researchers selected around 20 candidate variables from a large number of variables recorded in a UK primary care database. These variables included established risk factors already used in the previous risk score (QRISK-2 [14]) and new candidate variables recently highlighted in the literature or guidelines, such as corticosteroid use, severe mental illness, and diagnosis of HIV/AIDS [5].

2.2 Sample Size Consideration

Sample size consideration is important to develop robust models. For binary or time-to-event outcomes, a rule of thumb for the required sample size is to ensure at least 10 events for each predictor parameter [15, 16]. For example, to develop a clinical prediction model consisting of 15 parameters, the dataset should contain at least 150 patients with the outcome of interest. However, experts in this field suggest the following:

> The actual required sample size is context specific and depends not only on the number of events relative to the number of candidate predictor parameters but also on the total number of participants, the outcome proportion (incidence) in the study population, and the expected predictive performance of the model. [15]

Even when the sample size is already fixed in an existing dataset, the sample size calculations (using the *pmsampsize* package in Stata or R) might help determine if the sample size is sufficient and how many predictors can be considered before overfitting becomes a concern [15].

2.3 Splitting or Resampling the Data for Model Development and Validation

Fig. 1 illustrates common patterns of data usage for model development, internal validation (i.e., examination of the extent to which the model can perform in the original study sample), and external validation (i.e.,

examination of the extent to which the model can perform for other participant data than that used for model development [1]).

(i) Split-sample method

With a split-sample method, the model or equation is developed using a part of one dataset as the "derivation" stage, and its internal validity is evaluated in another part of the original study sample. The studied dataset is split randomly into two, one for model development and another for internal validation, in ratios such as 3:1 [5]. However, the split sample is generally not recommended for several reasons such as statistical inefficiency and instability of results in small datasets [11]. A methodological study comparing the split-sample and resampling methods suggested that resampling methods, especially the bootstrap method, provided more stable estimates than the split-sample methods [17].

(ii) Resampling method

Resampling methods include (k-fold) cross-validation and bootstrapping. Cross-validation involves developing and validating prediction models by dividing the data in turn and averaging the results. For example, in a five-fold cross-validation, a prediction model is created using four of the five data divisions, and the remaining one data division is used for validation (**Fig. 1**). This is repeated five times, and the average of the model estimates (e.g., regression coefficients in a regression model) and model performance (e.g., c-statistic) are presented as the final results.

In the bootstrap method, "bootstrap samples" of the same size are randomly sampled, with replacements from the original data many (e.g., more than 100) times. Models can be developed using bootstrap samples and validated using the original data or data from people not included in the bootstrap sample [18]. Resampling methods are also recommended to increase the statistical power, because all the data can be used for model development and validation [15]. Nowadays, packages of existing statistical software (e.g., Stata or R) have become more easily available [18].

(iii) Machine learning methods

Recently, machine learning methods, such as support vector machine (SVM), neural network (NN), and random forests, have been used increasingly to develop clinical prediction models [19]. However, caution is needed as machine learning methods are much more likely to cause "overfitting" (meaning that a model is developed to predict the data too well, so that it cannot be generalized to other data) than traditional regression models such as logistic regression and Cox regression models [20]. To avoid overfitting, in machine learning methods, it is recommended to use two datasets for the derivation stage alone (**Fig. 1**): a training set that is used

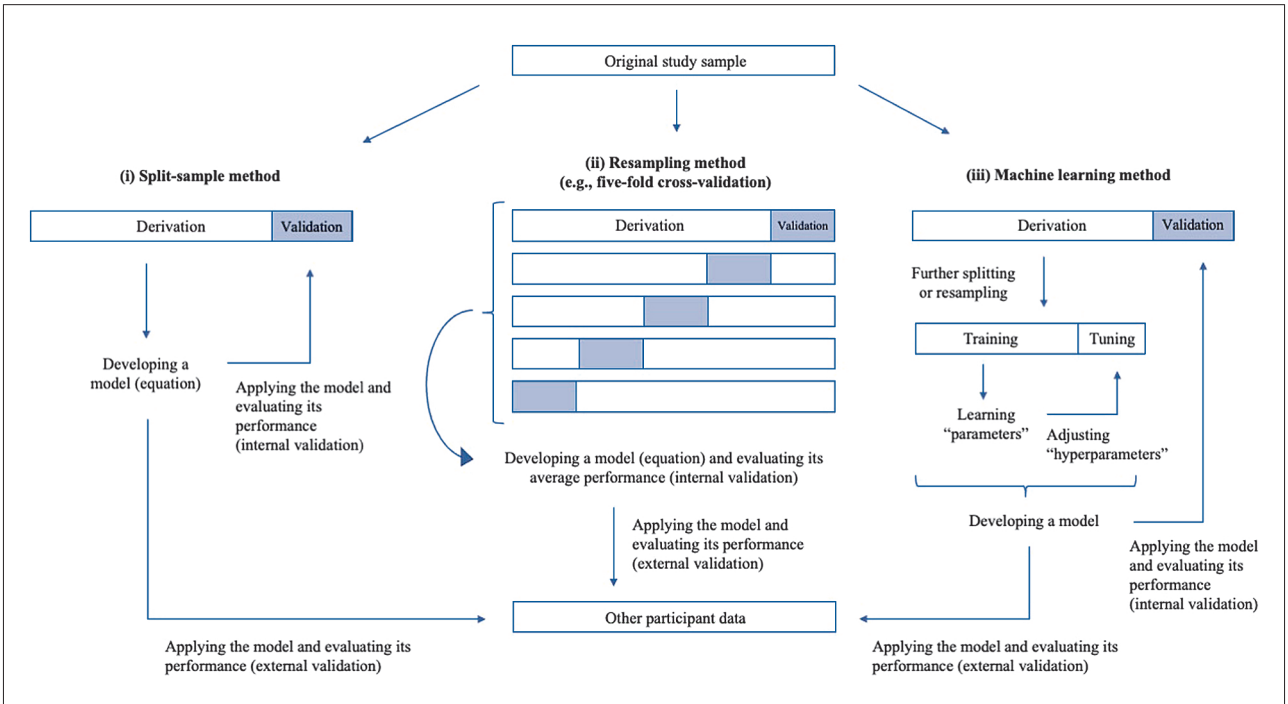

Fig. 1 **Common patterns of data usage for model development, internal validation, and external validation**

Note: In the field of machine learning, the derivation (further divided into training and tuning) and validation datasets may be called the "training" (further divided into "training" and "validation") and "test" datasets, respectively.

to learn "parameters" (e.g., support vectors for SVM and weights for NN) and a tuning set to adjust "hyperparameters" (e.g., C and sigma hyperparameters for SVM and learning rate for NN) [20]. Resampling methods may be preferred over split-sample methods to obtain the training and tuning datasets, especially when the derivation data are small. Notably, in the field of machine learning, the derivation (further divided into training and tuning) and validation datasets may be called "training" (further divided into "training" and "validation") and "test" datasets, respectively.

(iv) External validation

Ideally, within the same study or as a new study, the developed model should be validated using other participant data than that used for model development [1]. However, the definition of "other participant data" seems to be inconsistent across studies and among different researchers. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement suggests the following:

> External validation may use participant data collected by the same investigators, typically using the same predictor and outcome definitions and measurements, but sampled from a later period (temporal or narrow validation); by other investigators in another hospital or country, sometimes using different definitions and measurements (geographic or broad validation); in similar participants but from an intentionally different setting (for example, model developed in secondary care and assessed in similar participants but selected from primary care); or even in other types of participants (for example, model developed in adults and assessed in children, or developed for predicting fatal events and assessed for predicting non-fatal events). [1]

The TRIPOD team also suggests that temporal splitting of the original single dataset (e.g., into data for financial years 2014–18 and for 2019 [21]) can be considered an intermediate stage between internal and external validation [11]. The original single dataset can also be split based on study sites for model development and validation, which may be called "internal-external validation" [22]. In a methodological study comparing different strategies to split samples based on the study site and/or period, the study results (e.g., c-statistic) were very similar regardless of the strategies. However, the study's authors still recommend splitting the sample by study site and period in turn and pooling or meta-analyzing the results to examine geographic and temporal transportability, if possible [23].

3. DEVELOPMENT OF A CLINICAL PREDICTION MODEL

3.1 Selection of a Regression Model or Machine Learning Model(s)

For model development, researchers need to select a type of regression model (equation) or machine learning model(s) and plan a strategy to select predictors to be used in the final model. Researches tend to select common regression models/equations, such as the logistic regression model [7] and Cox regression model [5], whereas some researchers compare different models to suggest the best model in the context of the study [21, 24]. A recent systematic review suggests there is no performance benefit of machine learning over logistic regression for clinical prediction models [25]. However, this may be because many studies included in the systematic review used only a limited number and variation (e.g., binary) of variables. More research is needed to examine whether the conclusion is the same even if the number and variation of candidate variables and data volume are increased.

3.2 Selection of Predictor Variables

There are two main strategies to select predictors in the final model. The first is to increase model performance as much as possible. In general, it is expected that a variable with high prevalence (but not too high, i.e., over 50%) in the study population and large relative risk on the outcome can contribute to increased performance of the clinical prediction model. Adding such variables is expected to increase the model performance, as long as overfitting or multicollinearity (i.e., strong correlations between the variables, resulting in decreased performance of the model) does not occur. If researchers suspect overfitting or multicollinearity associated with certain variables, they should compare the performance of models with and without that variable in the validation stage to examine which model is better. Generally, overfitting is likely to occur if the total number of predictors in the model is too large relative to the number of outcomes, as well as if a variable is represented too much in the original study sample, such as patient ID.

Some researchers pre-define their own criteria to select predictors in the final model. For example, researchers developing the QRISK-3 planned to retain a variable if it had an adjusted hazard ratio of less than 0.90 or greater than 1.10 (for binary variables) and was statistically significant at the 0.01 level [5]. Consequently, among their candidate variables, HID/AIDS was dropped from the final model.

The second strategy is to develop a parsimonious model with a smaller number of variables, which is more efficient and easier to use in clinical practice. A traditional approach may be a stepwise regression, including forward selection (meaning that candidate predictors are added to the regression model one by one), backward elimination (meaning that candidate predictors are subtracted one by one from the regression model with all candidate predictors), or their combination, with some criteria to retain each predictor, such as the Akaike and Bayesian information criteria. Another increasingly popular approach is the least absolute shrinkage and selection operator (LASSO) regression, which constrains the sum of the absolute values of regression coefficients and can effectively exclude predictors from the final model by shrinking their coefficients to exactly zero [16].

3.3 Presentation of a Clinical Prediction Model

Once the prediction model is developed, researchers need to consider how to present it. Taking a logistic regression model as an example, the finally developed equation would look as follows:

$$\log [p/(1-p)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

where p is the probability of having an outcome, X_1 to X_k are all predictors, and β_0 to β_k are the regression coefficients;

$$\Leftrightarrow p/(1-p) = e^{(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta kXk)} = e^{\beta 0} e^{\beta 1X1}\, e^{\beta 2X2} \ldots e^{\beta kXk},$$

where $e^{\beta 1}$ to $e^{\beta k}$ correspond to an odds ratio for each predictor;

$$\Leftrightarrow p = e^{(\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta kXk)}/[1 + e^{(\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta kXk)}],$$

where p falls within a range of 0 to 1 (i.e., 0% to 100%).

Thus, by applying the estimated regression coefficients and information on each predictor of the individuals (e.g., 0 for male and 1 for female) to this equation, the probability of having the outcome (p) can be calculated for each individual.

However, it is theoretically possible but practically difficult for clinicians and patients to calculate the probability of individuals by identifying and using regression coefficients presented in a medical article. Therefore, researchers are expected to present a risk calculator as a paper-based tool (called a nomogram [26]) or a web-based tool, such as the QRISK-3 risk calculator (https://qrisk.org/three/index.php) [5]. Their algorithms and command programs are also fully open to the public (https://qrisk.org/three/src.php). Nowadays, creation of such web-app tools has become easier, so researchers tend to make their own apps and publish their weblinks in their abstracts or manuscripts [27].

Another method of presentation is to create a simple (integer) risk score, which can be calculated easily at bedside or outpatient, such as the Wells' criteria for PE [3, 4] and the Framingham risk score [28]. For this purpose, an estimated regression coefficient for each factor is often rounded to an integer value after multiplying it by a scaling factor (k) [29]. For example, in a study setting the scaling factor as 4, if a regression coefficient for a risk factor (e.g., type-2 diabetes on the incidence of cardiovascular disease) is 0.45, it can be approximated to an integer score of 2 after multiplying 0.45 by 4 (i.e., 1.8, which is closer to 2). A regression coefficient for another risk factor may be 0.1, which is approximated to 0 after multiplying 0.1 by 4 (i.e., 0.4, which is closer to 0). Therefore, it does not contribute to the risk score. Another approach is to define the constant (B), that is, the number of regression units that correspond to 1 point in the final score system. For example, Framingham investigators often set up B to be equivalent to the regression coefficient for a five-year increase in age [30]. Notably, despite its ease of use, such a simple risk score's predictive performance may be lower than that of the original (paper-based or web-based) risk calculator without approximating the regression coefficients.

4. VALIDATION OF A CLINICAL PREDICTION MODEL

There are several aspects in evaluating the performance of a developed clinical prediction model, such as goodness of fit, discrimination, calibration, and reclassification. Researchers should evaluate two or more of these aspects in the same study; for example, researchers evaluated goodness of fit, discrimination, and calibration for QRISK-3 [5].

4.1 Goodness of Fit

Goodness of fit suggests the extent to which the prediction model fits actual observations. A typical measure of this is a coefficient of determination (R^2). R^2 means the proportion of variability of the outcome that is explained by the prediction model among the total variability of the outcome. R^2 takes values between 0 and 1 (i.e., 0% and 100%), with larger values indicating better-fitting models. Because R^2 tends to naturally increase as the number of variables increases, the adjusted coefficient of determination (adjusted R^2), corrected for the number of variables in the model, is often used. In general, the (adjusted) R^2 is estimated in a linear regression model for a continuous variable as an outcome. In other regression models, modified R^2 has been proposed, such as pseudo-R^2 for a

logistic regression [31] and generalized R^2 for a Cox regression [32].

4.2 Discrimination

Discrimination refers to the prediction model's ability to discriminate between the presence and absence of the outcome, if the outcome is a binary variable. A representative indicator of this ability is a c-statistic, which is an estimate of the concordance index (c-index). The c-statistic is equal to an area under the curve of the receiver operating characteristic, which connects the dots suggesting the sensitivity on the vertical axis (Y-axis) and 1-specificity on the horizontal axis (X-axis) according to each cut-off value of the predicted probability/score (**Fig. 2**). The c-statistic takes values between 0.5 and 1, with larger values indicating better-discriminating models. For time-to-event outcomes typically modeled by the Cox regression, an estimator of the overall C, which suggests the ability to discriminate between a longer or shorter time until the incidence of outcome, was proposed by Harrel et al. [33] and improved by Uno et al. [34] to correct the bias dependent on the censoring distribution.

4.3 Calibration

Calibration involves examining the extent of agreement between the predicted risk of outcome incidence, as estimated by a clinical prediction model, and the actual risk of outcome incidence at a group level. In more detail, people in the study sample for validation are divided into multiple (e.g., five or 10) small groups of similar number of people, according to the size of the predicted probability (or risk score) for each individual. In each group, the average value of the predicted probability can be calculated, whereas the actual risk of outcome incidence or observed probability/proportion in the data (i.e., number of outcomes divided by the number of people in the group) can be also calculated. The plot of this information on the x- and y-axes, respectively, is called the calibration curve (**Fig. 3**). In addition, the Hosmer-Lemeshow test can test the null-hypothesis—whether there is no difference between the predicted and actual risks of outcome incidence. A large p-value suggests that calibration of a clinical prediction model is compatible with the event distribution in a dataset, while caution is needed as a smaller sample size is generally more likely to result in a larger p-value.

4.4 Reclassification

Reclassification refers to the ability of a variable (e.g., a new biomarker) that is added to an existing prediction model to better reclassify the risk of individuals [35]. It is

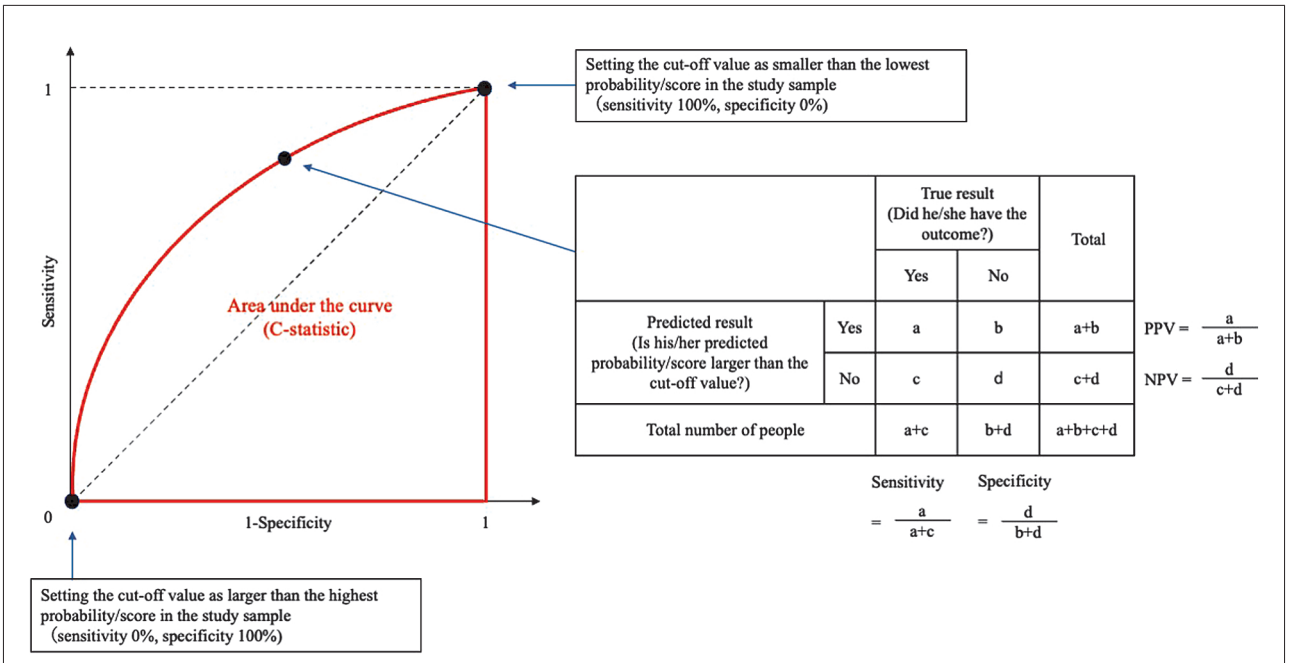

Fig. 2 **Graphical representation of a receiver operating characteristic curve and a two-by-two table (confusion matrix) for a certain cut-off value of the predicted probability/score**

Note: PPV = positive predictive value; NPV = negative predictive value.

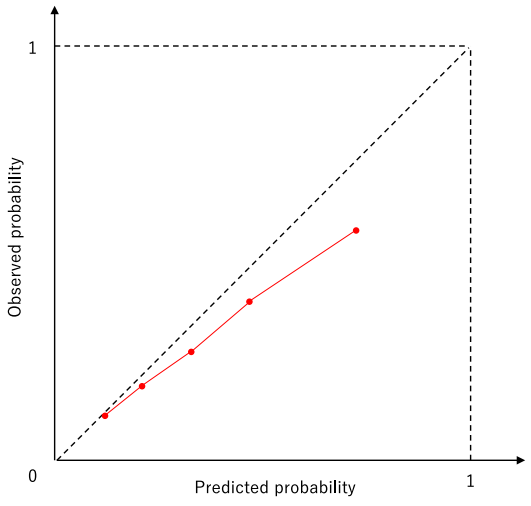

Fig. 3 **Hypothetical example of a calibration curve**

Note: The five dots suggest five groups with similar number of people according to the size of each individual's predicted probability or risk score. In this example, the probability of an event in high-risk patients is underestimated.

possible for a new biomarker to not improve discrimination (e.g., c-statistic), but it could better reclassify the existing prediction model [35]. Common indicators of the reclassification (ability) are net reclassification improvement (NRI) and integrated discrimination improvement (IDI).

To calculate the NRI, study participants are divided into multiple (e.g., two or three) risk categories according to their predicted probabilities in the prediction models with and without the newly added variable. Then, in each group of cases and non-cases (i.e., people with and without the outcome), the proportion of people moving into other risk categories based on the new model is calculated and combined by subtracting the proportion of unfavorable moves from the proportion of favorable moves. **Fig. 4** shows an example of a study examining the reclassification ability of genetic information (a polygenic risk score, which can be calculated for each individual by measuring their DNA from the blood and applying the results of a genome-wide association study) besides the QRISK-3 for coronary artery disease prediction in the UK Biobank [36]. The authors divided the study participants into two risk categories at 10%, because this threshold is used to start statins in UK primary care. Consequently, the reclassification improvement (RI) was calculated to be 0.043 (4.3%) for cases and −0.006 (−0.6%) for non-cases, resulting in an NRI 0.037 (3.7%). This could be interpreted as the prediction model improved by 3.7% because of additional genetic information.

One potential issue with the NRI is that the cut-off for the risk category could be chosen arbitrarily by researchers to achieve statistical significance. A solution to this is calculating a continuous NRI that does not create any category and simply counts the number and proportion of people moving in favorable and unfavorable directions in each group of cases and non-cases, respectively. In the aforementioned example of the UK Biobank, the continuous NRI was 0.296 by summing RI 0.149 for cases and RI 0.147 for non-cases [36].

However, a weakness of the continuous NRI is that it accounts for only the direction but not the amount of change in the predicted probability for each individual. Therefore, another solution is to calculate the IDI, an improvement in differences between the average predicted probabilities among cases and among non-cases (also known as Yates's discrimination slope) from the old model to the new model [35]. The IDI for the aforementioned example was 0.0064 [36].

5. IMPLEMENTING A CLINICAL PREDICTION MODEL IN ACTUAL PRACTICE

In recent years, there has been an increase in the number of studies on clinical prediction models. Most of these have developed and validated a clinical prediction model. However, for successful implementation of a clinical prediction model in actual clinical practice, more efforts may be needed to show that patients and/or clinicians can change their behaviors based on the results of clinical prediction. On the one hand, there has been little evidence indicating that informing patients of their risk score can directly change their health-related behaviors [37, 38]. On the other hand, there has been certain evidence that clinical prediction models can change physicians' decision making on examinations and treatments [38]. It is also possible that public health workers can use a clinical prediction model to identify people at risk for a certain outcome (e.g., initiation of long-term care) in the community and allocate limited health care resources or preventive opportunities to them efficiently. Additional analyses such as a decision-curve analysis and relative utility would offer insights on the clinical consequences or net benefits of using a prediction model at specific thresholds [11].

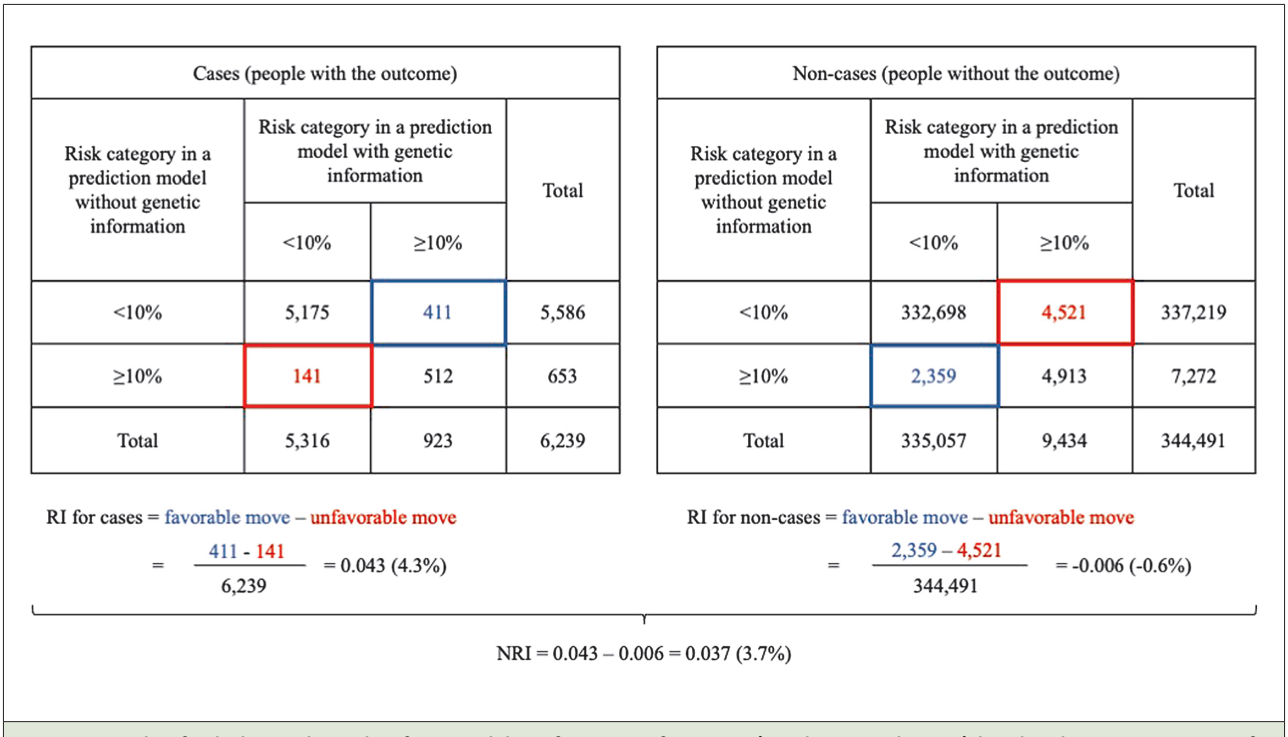

Fig. 4 **Example of calculating the reclassification ability of genetic information (a polygenic risk score) besides the existing QRISK-3 for coronary artery disease prediction in the UK Biobank [36]**

Note: NRI = net reclassification improvement; RI = reclassification improvement.

6. CONCLUSION

This paper provided an overview of how to develop and validate clinical prediction models—including a diagnostic prediction model and prognostic prediction model—by applying traditional regression models or emerging machine learning models to real-world data. At the derivation stage, researchers select candidate variables based on the literature review and clinical knowledge, as well as predictor variables used in the final model using predefined criteria (e.g., thresholds in the size of relative risk and p-value) or approaches such as the stepwise or LASSO regression. At the validation stage, performance of a clinical prediction model is evaluated in terms of goodness of fit, discrimination, calibration, and reclassification. Model validation should be performed for the original data to examine internal validity, and it may be performed using other participant data than that used for model development to examine external validity. Ultimately, a clinical prediction model is expected to change patient behaviors and/or clinicians' decision making as well as improve patient outcomes and/or public health.

ACKNOWLEDGMENTS

We would like to thank Dr. Tomohiro Shinozaki of the Department of Information and Computer Technology, Tokyo University of Science; Dr. Atsushi Goto of the Department of Health Data Science, Graduate School of Data Science, Yokohama City University; Dr. Sachiko Ono of the Department of Eat-loss Medicine, Graduate School of Medicine, The University of Tokyo; and Dr. Tadahiro Goto of TXP Medical Co. Ltd. and Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo, Dr. Yu Sun of the Graduate School of Comprehensive Human Sciences, University of Tsukuba, as well as Dr. Kazuaki Uda and Dr. Ryota Inokuchi of the Department of Health Services Research, University of Tsukuba, for their input and critical reading of the manuscript and feedback.

M.I. was supported by JSPS KAKENHI (Grant Number 19K19430).

CONFLICT OF INTERESTS

No potential competing interest relevant to this study is reported.

REFERENCES

1. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;13:g75941.
2. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol* 2021;132:142–5.
3. Wolf SJ, McCubbin TR, Feldhaus KM, Faragher JP, Adcock DM. Prospective validation of Wells criteria in the evaluation of patients with suspected pulmonary embolism. *Ann Emerg Med* 2004;44:503–10.
4. Wells PS, Anderson DR, Rodger M, Ginsberg JS, Kearon C, Gent M, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thromb Haemost* 2000;83:416–20.
5. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099.
6. National Institute for Health and Care Excellence. Cardiovascular disease: risk assessment and reduction, including lipid modification. https://www.nice.org.uk/guidance/cg181 (Accessed May 6, 2022).
7. Yamana H, Matsui H, Sasabuchi Y, Fushimi K, Yasunaga H. Categorized diagnoses and procedure records in an administrative database improved mortality prediction. *J Clin Epidemiol* 2015;68:1028–35.
8. Inohara T, Kohsaka S, Abe T, Miyata H, Numasawa Y, Ueda I, et al. Development and validation of a pre-percutaneous coronary intervention risk model of contrast-induced acute kidney injury with an integer scoring system. *Am J Cardiol* 2015;115:1636–42.
9. Welsh P, Welsh CE, Jhund PS, Woodward M, Brown R, Lewsey J, et al. Derivation and validation of a 10-year risk score for symptomatic abdominal aortic aneurysm: cohort study of nearly 500 000 individuals. *Circulation* 2021;144:604–14.
10. Pocock SJ, Ferreira JP, Gregson J, Anker SD, Butler J, Filippatos G, et al. Novel biomarker-driven prognostic models to predict morbidity and mortality in chronic heart failure: the EMPEROR-reduced trial. *Eur Heart J* 2021;42:4455–64.
11. Moons KG, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
12. Janssen KJ, Vergouwe Y, Donders, ART Harrell FE Jr, Chen Q, GrobbeeDE, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem* 2009;55:994–1001.
13. Morita K. Introduction to Multiple Imputation. *Annals Clin Epidemiol* 2021;3:1–4.
14. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475–82.
15. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.
16. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ* 2015;351:h3868.
17. Steyerberg EW, Harrell FE Jr, 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.
18. Noma H, Shinozaki T, Iba K, Teramukai S, Furukawa TA. Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism correction methods. *Stat Med* 2021;40:5691–701.
19. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ* 2020;369:m958.
20. Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: Users' guides to the medical literature. *JAMA* 2019;322:1806–16.
21. Ohbe H, Goto T, Nakamura K, Matsui H, Yasunaga H. Development and validation of early prediction models for new-onset functional impairment at hospital discharge of ICU admission. *Intensive Care Med* 2022.
22. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245–7.
23. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J Clin Epidemiol* 2016;79:76–85.
24. Osawa I, Goto T, Yamamoto Y, Tsugawa Y. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. *NPJ Digit Med* 2020;3:148.
25. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Calster BV. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12–22.
26. Ogura K, Fujiwara T, Yasunaga H, Matsui H, Jeon DG, Cho WH, et al. Development and external validation of nomograms predicting distant metastases and overall survival after neoadjuvant chemotherapy and surgery for patients with nonmetastatic osteosarcoma: A multi-institutional study. *Cancer* 2015;121:3844–52.
27. Kuno T, Sahashi Y, Kawahito S, Takahashi M, Iwagami M, Egorova NN. Prediction of in-hospital mortality with machine learning for COVID-19 patients treated with steroid and remdesivir. *J Med Virol* 2021;94:958–64.
28. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; 97:1837–47.
29. Cole TJ. Algorithm AS 281: scaling and rounding regression coefficients to integers. *Appl Stat* 1993;42:261–8.
30. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med* 2004;23:1631–60.
31. Hu B, Shao J, Palta M. Pseudo-R2 in logistic regression model. *Statistica Sinica* 2006; 16:847–60.
32. Royston P. Explained variation for survival models. *Stata J* 2006;6:83–96.
33. Harrell FE Jr, 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.
34. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30:1105–17.
35. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72, 207–12.
36. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, et al. Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. *JAMA* 2020;323:636–45.
37. Sheridan SL, Viera AJ, Krantz MJ, Ice CL, Steinman LE, Peters KE, et al. The effect of giving global coronary risk information to adults: a systematic review. *Arch Intern Med* 2010;170:230–9.
38. Usher-Smith JA, Silarova B, Schuit E, Moons KGM, Griffin SJ. Impact of provision of cardiovascular disease risk estimates to healthcare professionals and patients: a systematic review. *BMJ Open* 2015;5:e008717.