ARTICLE

Twelve practical recommendations for developing and applying clinical predictive models

Guoshuang Feng,[1,*] Huiyu Xu,[2] Shibiao Wan,[3] Haitao Wang,[4] Xiaofei Chen,[5] Robert Magari,[6] Yong Han,[7] Yongyue Wei,[8] and Hongqiu Gu[9]
*Correspondence: glxfgsh@163.com

Received: November 6, 2024; Accepted: December 2, 2024; Published Online: December 6, 2024; https://doi.org/10.59717/j.xinn-med.2024.100105

GRAPHICAL ABSTRACT

PUBLIC SUMMARY

- This article proposes 12 key recommendations for the application of medical prediction models in clinical practice.
- Developing prediction models isn't just software output; it requires avoiding pitfalls at multiple steps.
- Publication alone can't ensure prediction model use; clinical value and physician acceptance are essential.
- Promoting prediction models requires internal and external validation, impact evaluation, and expert support.

ARTICLE

Twelve practical recommendations for developing and applying clinical predictive models

Guoshuang Feng,[1,*] Huiyu Xu,[2] Shibiao Wan,[3] Haitao Wang,[4] Xiaofei Chen,[5] Robert Magari,[6] Yong Han,[7] Yongyue Wei,[8] and Hongqiu Gu[9]

[1]Big Data Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China
[2]State Key Laboratory of Female Fertility Promotion, Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing 100191, China
[3]Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE 68198, United States
[4]Thoracic Epigenetics Section, Thoracic Surgery Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, United States
[5]Statistical Science, Southern Methodist University, Dallas, TX 75205, United States
[6]Department of R & D, Beckman Coulter, Brea, CA 33196, United States
[7]Key Laboratory of Tumor Molecular Diagnosis and Individualized Medicine of Zhejiang Province, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou 310014, China
[8]Center for Public Health and Epidemic Preparedness & Response, School of Public Health, Peking University, Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China
[9]China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Capital Medical University, Beijing 100050, China
*Correspondence: glxfgsh@163.com

Received: November 6, 2024; Accepted: December 2, 2024; Published Online: December 6, 2024; https://doi.org/10.59717/j.xinn-med.2024.100105

Citation: Feng G., Xu H., Wan S., et al., (2024). Twelve practical recommendations for developing and applying clinical predictive models. The Innovation Medicine **2(4)**: 100105.

Prediction models play a pivotal role in medical practice. To ensure their clinical applicability, it is essential to guarantee the quality of predictive models at multiple stages. In this article, we propose twelve recommendations for the development and clinical implementation of prediction models. These include identifying clinical needs, selecting appropriate predictors, performing predictor transformations and binning, specifying suitable models, assessing model performance, evaluating reproducibility and transportability, updating models, conducting impact evaluations, and promoting model adoption. These recommendations are grounded in a comprehensive synthesis of insights from existing literature and our extensive clinical and statistical experience in the development and practical application of prediction models.

INTRODUCTION

The medical field is witnessing a surge in the development of predictive models. Despite numerous publications, only a limited number of models are actively utilized in clinical settings. This gap primarily stems from concerns about the quality of the models. While most prediction models report moderate to excellent performance, there is often a significant risk of bias. Studies conducted by Wynants et al.,[1] Liu et al.,[2] and Kaiser et al.,[3] evaluating 606, 15, and 42 models respectively, revealed high bias risks, with proportions of 90%, 83%, and 81%. Moreover, practical considerations such as clinician acceptance, the model's utility for patients, and its applicability across various promotional scenarios play crucial roles in the successful clinical implementation of these models.

The TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement provides guidelines for reporting on the development, validation, or updating of prediction models, whether for diagnostic or prognostic purposes.[4] Although TRIPOD has released guidelines for enhancing the quality of article publications regarding predictive models, these recommendations primarily offer general suggestions without providing specific guidance for each stage. Furthermore, TRIPOD predominantly focuses on improving the quality of article publications rather than the clinical implementation of models.

In this article, we propose twelve recommendations to enhance the application of predictive models in clinical practice, whether for diagnostic or prognostic purposes. These recommendations emphasize the significance of clinical requirements, meticulous predictor choice, robust variable selection methods, rigorous performance validation, and appropriate impact evaluation techniques. Our focus is primarily on prediction models for binary outcomes (i.e., whether a patient experiences an event or not), which are commonly encountered in medical prediction studies. The recommendations presented in this article integrate statistical strategies with the author team's extensive experience in developing and applying three specific models: *OvaRePred* model (for evaluating and predicting female ovarian reserve),[5] *PCOSt* model (a screening tool for polycystic ovary syndrome (PCOS), a reproductive endocrine disorder in women),[6] and *POvaStim* model (for guiding follicle-stimulating hormone (FSH) dosage).[7] These models are currently in use at multiple hospitals and physical examination institutions in China, and they are also undergoing preparations for extensive clinical implementation.

IDENTIFYING THE NECESSITY OF MODEL DEVELOPMENT

Before initiating model development, it is vital to assess whether there is a genuine clinical need and potential for practical application. Identifying a precise and relevant clinical problem is of paramount importance, encompassing both the model's clinical utility and its feasibility for actual implementation.[8] For instance, a pediatrician considered utilizing a gaming platform to evaluate children's attention and develop a diagnostic model for attention deficit hyperactivity disorder (ADHD). However, the current clinical diagnosis of ADHD largely relies on cost-effective and convenient scales. In contrast, the proposed tool appeared more time-intensive and costly, offering limited practical value.

Clinical relevance and practical value alone do not necessarily justify the immediate development of a new model. If existing models are adequate, it is more prudent to validate them externally rather than redevelop them. The proliferation of numerous models is resource-intensive and may place clinicians in a difficult position, unsure of which model to choose, which hinders the clinical application of the model. Furthermore, if an externally validated model underperforms, developing a new model hastily is not advisable. Instead, updating the existing model should be prioritized, as this can expand its applicability.

If an existing model has demonstrated robust external validation but failed to achieve widespread clinical adoption, practical considerations need to be addressed: Are there limitations in the accessibility of predictive variables in real-world settings? Is the model too complex for clinicians to accept? Alternatively, might the absence of user-friendly prediction tools (e.g., web-based interfaces) be a contributing factor? Addressing these queries can facilitate the practical application of the model.

As an example, when developing the *OvaRePred* model, we first established its clinical significance and acknowledged its critical role in assessing female fertility. A thorough literature review revealed a global lack of clinically applicable tools for evaluating ovarian reserve and predicting major fertility-related events. This gap underscored the need to develop an ovarian reserve assessment and prediction tool. To ensure its seamless integration into clinical workflows, we developed a web-based prediction tool for clinician use. The model has achieved significant milestones, such as obtaining a patent and transitioning into commercial use. Currently deployed in clinical settings for over a year, it has demonstrated promising initial results, supported by impact evaluation.

CHOOSING APPROPRIATE PREDICTORS

The selection of predictors should be approached from both clinical and statistical perspectives. Clinically, predictors should be chosen based on expert knowledge or supported by literature. Variables that have been previously identified as prognostic or hold clinical importance are recommended as primary candidates. It is of great importance that these predictors are objective, precisely defined, standardized, and reproducible to guarantee both the generalizability and practical applicability of the model.

Clinical perspective

Typical predictors include demographics, disease type and severity, disease history, clinical features, laboratory tests, and genetic traits. From an applicability perspective, readily available predictive variables are preferred. If performance differences are negligible, a model with easier implementation may be more readily adopted in a clinical setting. For instance, the Framingham risk model, despite its moderate area under the receiver operating characteristic (ROC) curve, is extensively used due to its accessible and cost-effective predictors.[9]

Uniformity in predictors is vital for the broad adoption of prediction models. When key predictors lack standardization, such as varying measurement values among different manufacturers, it is essential to establish uniform standards in advance. For instance, with Anti-Müllerian hormone (AMH) being a crucial marker in the *OvaRePred* model, different hospitals using instruments from various AMH manufacturers necessitated the establishment of conversion relationships among common AMH values. To address this, we developed conversion relationships among the measurement values from three major AMH manufacturers using Passing–Bablok regression.[10] This approach ensures that prediction models remain effective and usable across different hospitals, regardless of the brand of instruments they use and the resulting variations in AMH absolute values.

Objective predictors with well-defined and reliable measurements should be prioritized, as subjective predictors, such as the interpretation of imaging results, may introduce inconsistency due to inter-observer variability. For example, in the establishment of *AAFA* model for assessing ovarian reserve,[11] four predictors were included: AMH, antral follicle count (AFC), FSH, and age. However, we acknowledged the subjectivity involved in assessing AFC through ultrasound evaluation by clinicians. Consequently, we developed a revised *AFA* model using only three predictors: AMH, FSH, and age.[12] This modification enhances the objectivity of the model while maintaining comparable predictive performance and consistent predictions. If subjective predictors are still necessary, it is imperative to involve multiple raters for independent assessment and evaluate inter-rater consistency.

Easily obtainable predictors can enhance the applicability and usability of predictive models, thereby improving their practicality. For instance, while applying the *AFA* model, we found that FSH is only obtainable during menstrual cycle day 2 to 3, significantly limiting its accessibility. To facilitate wider applicability, we developed a streamlined version of prediction model that relies solely on AMH and age,[13] allowing blood draws at any time point in the menstrual cycle and simplifying implementation. Concurrently, we conducted comprehensive comparisons of several models to validate the availability and reliability of our streamlined model.

Prioritizing predictors that yield reliable and stable results is crucial. During the development the *PCOSt* model, the initial predictive variables included AMH, body mass index (BMI), the upper limit of menstrual cycle length (UML), and basal androstenedione (A4).[14] However, we encountered a challenge with the A4 indicator, as its measurement primarily used the chemiluminescence method in our study. Unlike gold-standard mass spectrometry, this method demonstrated a weak correlation (R^2=0.285) with mass spectrometry results in women, raising concerns about its precision.[15] Consequently, we excluded the A4 indicator and refined the predictors to include AMH, BMI, and UML, significantly enhancing the model's stability and reliability.[16]

Statistical perspective

From a statistical perspective, it is crucial to consider the distribution of the predictors. If the predictor exhibits minimal variability, it is not advisable to include it in the analysis, particularly if there is insufficient subject knowledge supporting its inclusion. For instance, in the case of a binary predictor where one class constitutes 98% of the data, it is generally unnecessary to incorporate such a variable into the model unless it is known to be highly predictive.

Another challenging issue for clinicians is whether predictors can be simultaneously considered as candidate variables in the presence of collinearity (i.e., strong correlation among predictors). For instance, when the correlation of AMH and AFC is above 0.7, can we keep both as candidates? This determination relies on the specific application context of the model.

It is important to note that collinearity does not impact the performance of prediction models. Instead, it influences the coefficient estimates and *p*-values, thereby affecting the interpretability of the model. Therefore, if the primary goal is purely predictive and without the need to understand the influence of each independent variable, predictors with high correlation can be retained. However, for models requiring interpretability, it is advisable to remove one correlated predictor. In addition, while principal component analysis (PCA) can be used as a statistical technique for summarizing information from correlated predictors, it compromises interpretability of variable coefficients. Therefore, we do not recommend PCA as the preferred method from a practical perspective.

DETERMINING ADEQUATE SAMPLE SIZE

Before determining the sample size for a study, it is crucial to first ask: Why is sample size determination necessary? In clinical trials, calculating an appropriate sample size ensures the primary research hypothesis can be tested effectively while controlling costs. However, the development of prediction models prioritizes predictive performance over statistical inference.

Furthermore, many prediction models are built using pre-existing datasets from hospital information systems, which eliminates additional data collection costs. For instance, the *PCOSt* model was developed using data from the hospital's laboratory information system, incorporating all 11,720 available cases. Although this approach did not involve determining the minimal required sample size, it made full use of existing resources without incurring extra financial costs.

The determination of sample size in the development of predictive models can be approached through two primary methods: rules of thumb and calculation approach based on specific criteria.

Rules of thumb for sample size determination

The rule-of-thumb of 10 EPV (10 events per variable) and 10 EPP (10 events per candidate predictor parameter) is widely advocated due to its simplicity, with a preference for 10 EPP to avoid potential misinterpretations of the term "variable". For example, when a continuous variable such as "age" is included in a model, it corresponds to a single parameter for estimation. However, adding a quadratic term for age increases the number of parameters to two. Similarly, categorizing age into five distinct groups requires the estimation of four parameters.

Despite their simplicity, rules of thumb have sparked extensive debates due to their inconsistency.[17-20] Some simulations suggest increasing 10 EPP to 20 or even 50 to reduce bias.[21-23] This inconsistency arises from the fact that the required events depend on various factors, such as the effect size of predictors, outcome proportion in the study population, and distribution of predictors.[24]

Rational calculation approach

The calculation approach was proposed by Riley et al and includes five steps: [24-26]

Step 1: Calculate a sample size to ensure a precise estimate of the overall outcome risk, with a recommended margin of error of 0.05 for the overall outcome proportion.

Step 2: Calculate a sample size to ensure a small mean absolute prediction error (MAPE), recommended to be below 5%.

Step 3: Calculate a sample size to ensure a global shrinkage factor (a metric for quantifying overfitting) greater than 0.9, minimizing overfitting.

Step 4: Calculate a sample size to ensure that the absolute difference between the apparent and adjusted $R^2_{\text{Nagelkerke}}$ is no more than 0.05.

Step 5: Select the largest of the four calculated sample sizes as the final sample size.

While the calculation approach is theoretically more rational than the rules

Table 1. The number of EPP for different event proportion, number of parameters and *c* statistics

c statistic	Event proportion	Number of parameters					
		5	10	15	20	25	30
0.8	0.1	8	8	8	8	8	8
0.8	0.2	10	9	9	9	9	9
0.8	0.3	19	10	10	10	10	10
0.8	0.4	30	15	12	12	12	12
0.8	0.5	39	19	15	15	15	15
0.9	0.1	4	4	4	4	4	4
0.9	0.2	10	5	5	5	5	5
0.9	0.3	19	10	7	7	6	7
0.9	0.4	30	15	10	8	8	8
0.9	0.5	39	19	13	10	10	10

of thumb, as it considers multiple factors, numerous practical challenges remain. First, although the authors provided recommended values for parameters required for sample size calculation, using a fixed parameter value is not advisable. For instance, applying the same error margin of 0.05 to event proportions of 5% and 50% is inappropriate. Therefore, in practical applications, it is still necessary to integrate clinical knowledge and previous literature to make a comprehensive determination. Second, however, obtaining prespecified parameters from existing literature presents significant challenges. In a systematic review by Paula Dhiman et al.,[27] among 152 evaluated models from 62 studies, only one reported R^2 and four reported MAPE or RMSE (Root Mean Square Error). Therefore, in practical computations, the determination of parameters involves a degree of subjectivity.

Determining EPP using calculation approach

To streamline clinical applications, we calculated the sample size and corresponding number of EPP for different *c* statistics, event proportions, and numbers of parameters based on the five steps described earlier. The results are summarized in Table 1. Our analysis revealed that when the number of parameters is 10 or fewer, the number of EPP is significantly influenced by both the number of parameters and the event proportion. However, when the number of parameters exceeds 10, the number of EPP is primarily determined by the event proportion. In practical scenarios, where the number of parameters typically exceeds 10, 20 EPP are generally sufficient. Furthermore, when the event proportion is below 0.2, 10 EPP are also adequate.

Table 1 provides a rough estimate of sample size based on EPP, making it convenient for practical application. For example, if previous studies indicate that the *c* statistic is approximately 0.8, the event proportion is about 0.2, and 20 parameters are planned to be included in the model, the required sample size would be 9 EPP, equivalent to 180 events. Consequently, the total sample size would be 180/0.2 = 900 cases.

Practical considerations

Finally, three key considerations should be noted:

1. The number of predictor parameters refers not to those included in the final model but to those initially considered for inclusion. For instance, if 20 parameters were initially considered, and only 8 were ultimately included after variable selection, the EPP principle should be applied based on the original 20 parameters.

2. The rules of thumb and calculation approach are primarily designed for regression models, such as logistic regression. However, when employing machine learning techniques like neural networks, the number of parameters often far exceeds that of regression models. Consequently, machine learning methods typically require a much larger sample size than regression models. Therefore, it is recommended to aim for the largest sample size feasible when utilizing machine learning approaches.[28]

3. The aforementioned methods primarily focus on model development, whereas the requirements for sample size in external validation may differ. Researchers have proposed four criteria for accurately estimating the sample size needed for external validation,[29,30] including the precise estimation of the observed/expected (O/E) ratio, calibration slope, *c* statistic, and standardized net benefit at specific probability thresholds.

HANDLING MISSING VALUES AND OUTLIERS

Clinical prediction models often rely on real-world data from hospitals (e.g., electronic medical records, laboratory information systems), which frequently encounter issues such as missing data and outliers. These problems must be appropriately addressed to avoid compromising data quality.

Missing data mechanisms

Missing data can arise from various mechanisms, and understanding these mechanisms is essential because the choice of statistical methods for handling missing data depend on the underlying assumptions about the missingness. The MCAR (missing completely at random) mechanism typically occurs due to random errors (e.g., laboratory accidents). The MAR (missing at random) mechanism refers to missing values that depend on other observed variables, indicating a relationship with them. An MNAR (missing not at random) mechanism arises when the missingness is related to the missing values themselves or other unobserved predictors.

Methodologies for handling missing values

Extensive research has been conducted on various methodologies for handling missing values.[31-34] For situations involving data that are MAR or MCAR, multiple imputation (MI) is often considered as the optimal solution, regardless of the proportion of missing data.[35,36] In practice, however, a more commonly used method is complete case (CC) analysis,[37,38] in which individuals with missing data on any predictors or outcomes are excluded from the analysis. Simulation studies have shown that under most missing data mechanisms (MCAR, MAR when missingness is related to other predictors, and MNAR), CC analysis provides reasonably accurate coefficient estimates when 10% to 20% of the data are missing. However, significant bias arises under MAR when missingness depends on the outcome. Additionally, CC analysis underestimates the performance quantified by adjusted R^2 statistic.[39] Considering the convenience of MI techniques provided by statistical software, we recommend adopting MI as the primary method.

Simulation study on missing data

Using data from the *PCOSt* model, we conducted simulations to evaluate the impact of various missing data mechanisms, missing proportions, and handling methods on model performance. The dataset comprises 11,720 cases, with PCOS as the outcome variable (prevalence: 10.45%) and predictors including AMH, AFC, A4, and BMI. The correlation coefficients between AMH and AFC, between AMH and A4, and between AMH and BMI were 0.65,

Table 2. The deviation of *c* statistic and R^2 for different missing mechanisms, missing proportions and handling methods

missing mechanisms	missing proportion	*c* statistic deviation(%) and 95%*CI*		R^2 deviation(%)and 95%*CI*	
		CC	MI	CC	MI
MCAR	0.1	0.06(−0.48,0.60)	0.06(−0.48,0.60)	0.06(−2.40,2.43)	0.06(−2.40,2.43)
MCAR	0.2	0.00(−0.84,1.02)	−0.36(−0.60,−0.24)	0.06(−3.80,4.08)	−2.65(−3.85,−1.27)
MCAR	0.3	0.00(−1.08,1.20)	−0.60(−0.90,−0.36)	0.03(−4.88,5.51)	−3.95(−5.20,−2.55)
MAR on AFC	0.1	−0.96(−1.44,−0.60)	−0.36(−0.60,−0.24)	−3.27(−6.37,−0.72)	−2.30(−3.42,−1.15)
MAR on AFC	0.2	−2.64(−3.48,−2.04)	−0.72(−0.96,−0.48)	−8.67(−13.55,−4.38)	−4.57(−5.97,−3.17)
MAR on AFC	0.3	−5.76(−6.83,−4.68)	−1.20(−1.32,−0.96)	−18.02(−24.12,−11.80)	−6.96(−8.12,−5.65)
MAR on A4	0.1	−0.60(−1.08,0.00)	−0.24(−0.48,0.00)	−2.79(−5.57,0.07)	−1.74(−2.86,−0.59)
MAR on A4	0.2	−1.56(−2.40,−0.66)	−0.48(−0.72,−0.24)	−7.08(−11.53,−2.35)	−3.51(−4.75,−2.01)
MAR on A4	0.3	−3.12(−4.32,−2.16)	−0.72(−0.96,−0.48)	−13.86(−19.70,−8.77)	−5.25(−6.71,−3.57)
MAR on BMI	0.1	−0.24(−0.84,0.24)	−0.18(−0.36,0.00)	−1.86(−4.60,1.15)	−1.47(−2.42,−0.47)
MAR on BMI	0.2	−0.72(−1.44,0.00)	−0.36(−0.60,−0.18)	−4.35(−8.06,−0.71)	−2.92(−4.05,−1.83)
MAR on BMI	0.3	−1.44(−2.28,−0.60)	−0.60(−0.84,−0.48)	−8.08(−12.71,−3.83)	−4.42(−5.43,−3.11)
MAR on PCOS	0.1	0.12(−0.12,0.24)	0.00(−0.12,0.00)	2.57(1.50,3.73)	−0.12(−0.66,0.50)
MAR on PCOS	0.2	0.12(−0.12,0.24)	−0.12(−0.12,0.00)	5.46(3.95,7.19)	−0.29(−1.06,0.69)
MAR on PCOS	0.3	0.12(−0.12,0.36)	−0.24(−0.24,−0.12)	8.73(6.80,10.95)	−0.65(−1.62,0.35)
MNAR	0.1	−0.96(−1.44,−0.48)	−0.36(−0.60,−0.12)	−3.07(−5.93,−0.02)	−2.09(−3.29,−0.81)
MNAR	0.2	−2.52(−3.29,−1.92)	−0.72(−0.96,−0.48)	−8.4(−12.96,−3.81)	−4.13(−5.51,−2.76)
MNAR	0.3	−5.64(−6.77,−4.56)	−1.20(−1.44,−0.84)	−17.25(−23.66,−10.96)	−6.34(−7.81,−4.81)

0.38, and 0.07, respectively. We designed six scenarios for missing data, including MCAR for AMH, MAR for AMH (missingness dependent on AFC, A4, BMI, and PCOS), and MNAR for AMH. In each scenario, the missing rates were set at 10%, 20%, and 30%. For each type of missingness, we applied both CC analysis and MI method to handle it. To ensure robustness, we repeated the simulations 1000 times for each scenario to calculate the average values and percentage bias in *c* statistics and R^2 compared to the true value. The results are summarized in Table 2.

For the MCAR of AMH, CC analysis showed minimal bias in both the *c* statistic and R^2, whereas the MI method exhibited a slight bias, with the bias in the *c* statistic and R^2 increasing as the proportion of missing data rose.

The MAR scenario is more complex. When AMH missingness depended on AFC, the bias in the *c* statistic and R^2 was the highest, with CC analysis showing greater bias than the MI method. When AMH missingness was dependent on A4, the bias in the *c* statistic and R^2 was reduced. This bias decreased further when missingness was dependent on BMI. Notably, in cases where AMH missingness was dependent on PCOS, the bias in the *c* statistic was minimal, while R^2 was overestimated with CC analysis.

In the MNAR scenario, both CC analysis and the MI method exhibited a high bias in the *c* statistic and R^2, approximately equivalent to the results when AMH missingness depended on AFC.

Based on the simulation results, we recommend the following handling approaches for different scenarios:

1. MCAR scenario: when the missing proportion is within 30%, CC analysis yields nearly unbiased performance, and the MI method also shows low bias. Both methods can be considered. However, due to the reduction in sample size with the CC method, its confidence interval tends to be wider.

2. MAR scenario I: if the missingness of a predictor depends on other variables, the correlation between the predictor and these variables significantly impacts model performance. The stronger the correlation and the higher the missing proportion, the greater the impact. In such cases, it is advisable to use the MI method to mitigate bias.

3. MAR scenario II: if the missingness of a predictor depends on outcome, its effect on the *c* statistic is relatively minor, regardless of whether CC analysis or the MI method is used.

4. MNAR scenario: The impact of missing values on performance is similar to that described in MAR scenario I. Although the MI method may not be optimal for MNAR, it is still superior to CC analysis.

Practical challenges in determining missing mechanisms

In practical research, determining the missing data mechanism is often the most challenging task. MCAR is uncommon in medical research, and its assumptions can usually be tested. For instance, in our investigation of PCOS among women, we observed a missing data rate of approximately 11% related to the upper limit of the menstrual cycle. Notably, the missing rate for women over 35 was twice that of women under 35, indicating that the missing data mechanism does not conform to MCAR assumptions. However, even if we find a correlation between the missingness of predictors and the values of other variables, we can only infer that it is not MCAR; it remains unclear whether the missingness conforms to the MAR assumption, as we lack information on whether the missingness is directly related to the values of the predictors themselves (e.g., occurring primarily at higher values of the upper limit of the menstrual cycle). Verifying MNAR remains elusive and typically relies on reasonable assumptions informed by clinical knowledge. For example, when examining the discrepancies in cancer stage diagnosis among five prevalent cancers,[40] a notable prevalence of missing data was observed in tumor stage variables. It is plausible to infer that the frequency of missing data is higher among patients with advanced stages than among those with early stages, suggesting a possible MNAR mechanism.

Handling outliers

Outliers can significantly distort parameter estimation, making their detection an essential step in predictive modeling. Various methods for outlier detection have been documented,[41-44] typically classified into univariate and multivariate techniques.

In predictive modeling, relying solely on univariate outlier evaluation may not always suffice. For instance, consider measurements of height and weight, with values of 150 cm and 90 kg, respectively. While neither variable

appears anomalous on its own, their combination in a single individual might represent a multivariate outlier, as it is highly unusual for a person with a height of 150 cm to weigh 90 kg. Consequently, we recommend employing multivariate outliner detection methods in predictive models, such as the Mahalanobis Distance or Nearest Neighbor-based techniques.[45,46]

Once outliers are detected, the first step is to identify their underlying cause. Based on our extensive analytical experience, we have found that most outliers result from data entry errors, which are often correctable. When outliers are not attributable to such errors, the most pragmatic approach is to exclude them. This recommendation is based on two key considerations: first, outliers typically constitute a very small proportion of the dataset used for model development (often less than 0.1% in our experience), so their exclusion has a negligible impact on model building; second, predictive models are designed for the general population, and occasional outliers in application will not compromise overall performance.

PERFORMING PREDICTOR TRANSFORMATION AND BINNING

When developing predictive models, continuous predictors are commonly encountered. Most regression-based predictive models assume a linear relationship between continuous predictors and the outcome (or a link function of the outcome). This assumption implies that a one-unit increase in the predictor has the same effect on the outcome regardless of the predictor's value. For instance, the risk associated with increasing age from 24 to 25 years is assumed to be equivalent to the risk associated with increasing age from 44 to 45 years. However, in reality, this relationship may not strictly follow linearity; the risk of a change from 44 to 45 years is likely to differ significantly. Therefore, when incorporating continuous predictors into the model, it is crucial to carefully evaluate and determine the appropriate functional form.

Predictor transformation

If a linear relationship exists between a continuous predictor (or its transformed form, such as a logarithmic transformation) and the outcome, the predictor in its original or transformed form can be directly included in the model. However, in many cases, the relationship is either unknown or cannot be adequately captured by a simple transformation. In such situations, it is advisable to employ methods such as fractional polynomials, which use a restricted set of power transformations, or restricted cubic splines (RCS), which fit cubic polynomials to segments of the data and join them smoothly at predefined knots.[47-49] These methods offer greater flexibility in capturing associations between predictor variables and outcomes. Nevertheless, caution is needed as they may lead to overfitting and reduced interpretability.

Predictor binning

Another commonly employed approach in practice is binning (categorizing a variable into multiple groups),[50] although it raises concerns about information loss and the challenge of determining appropriate cut-points.[51,52] Previous studies have compared fractional polynomials, RCS, and binning, revealing that fractional polynomials and RCS exhibit slightly higher *c* statistics values than binning (approximately 0.01 higher in the development set and 0.02 higher in the validation set).[53] However, clinicians often prefer the simplicity and interpretability of the binning method when applying models in practice. Therefore, a trade-off between interpretability and goodness-of-fit should be weighed, making binning a viable option when the difference in model performance is minimal.

In a study examining the association between hypertension (binary outcome) and adiponectin (a continuous variable), we evaluated the impact of three approaches for handling adiponectin on model performance, as measured by the *c* statistic. These approaches included modeling a linear relationship, employing RCS to capture potential nonlinear relationships, and binning adiponectin into three groups based on RCS-derived shape (Figure 1). The corresponding *c* statistics were 0.41, 0.66, and 0.69, respectively. Notably, despite observing a U-shaped relationship between the two variables, incorporating adiponectin as a quadratic term in the model did not yield statistically significant results, and the resulting *c* statistic of 0.66 was lower than that obtained through the binning method. This discrepancy may be attributed to the skewed distribution of adiponectin values, particularly within

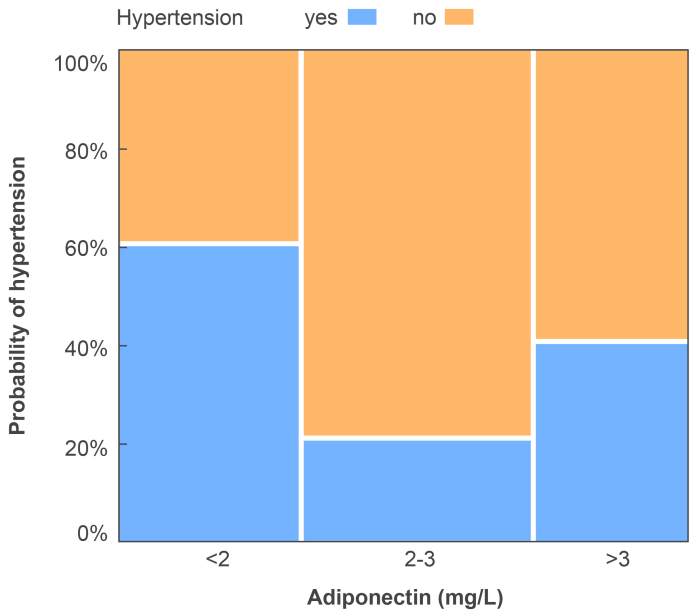

Figure 1. Illustration of nonlinear relationship between hypertension and adiponectin.

certain ranges.

When employing the binning method, it is essential to carefully determine both the number and placement of cut-points. From a practical standpoint, we propose the following recommendations:

First, if prior clinical knowledge or consensus exists, categorization should be guided by such expertise.

Second, in the absence of prior knowledge, statistical techniques such as tree-based methods or spline techniques can be employed to identify suitable cut-points.

Third, it is crucial to ensure that each category encompasses a sufficient number of cases.

Finally, dividing categories based on quantiles is generally not advisable.

For instance, during the development of a prediction model for PCOS, we observed a skewed distribution of testosterone levels, with approximately 80% of values falling below 0.7 nmol/L (Figure 2). There's natural categorical distinction, making it easy to find a cut-point. Given the substantial size of the development dataset, we divided values exceeding 0.7 into four groups, starting at 0.7 with increments of 0.2. Considering the specific distribution pattern observed in our study population, we believe it is reasonable to classify testosterone levels into categories rather than treating them as a continuous variable.

Recommendations

In summary, when incorporating continuous predictors into the regression model, we recommend the following approach:

1. If the predictor (or its transformed form) exhibits a linear relationship with the outcome (or its link function), it is preferable to retain the variable as continuous in the model. In this case, using a linear model to incorporate the continuous variable or its transformed function (e.g., log transformation, quadratic transformation) is likely to yield optimal performance.

2. If a nonlinear relationship is observed and no clear transformation function is available, fractional polynomials or RCS can be utilized. Alternatively, predictors may be binned based on exploratory analysis using spline techniques.

3. When categorizing continuous predictors to improve interpretability, it is advisable to divide them into at least three categories while ensuring that each category includes an adequate sample size. Dichotomization is generally not recommended. Cut-points should be determined comprehensively by integrating clinical expertise and statistical methods (e.g., spline techniques). The use of data-driven cut-points or quantiles as cut-points should generally be avoided.

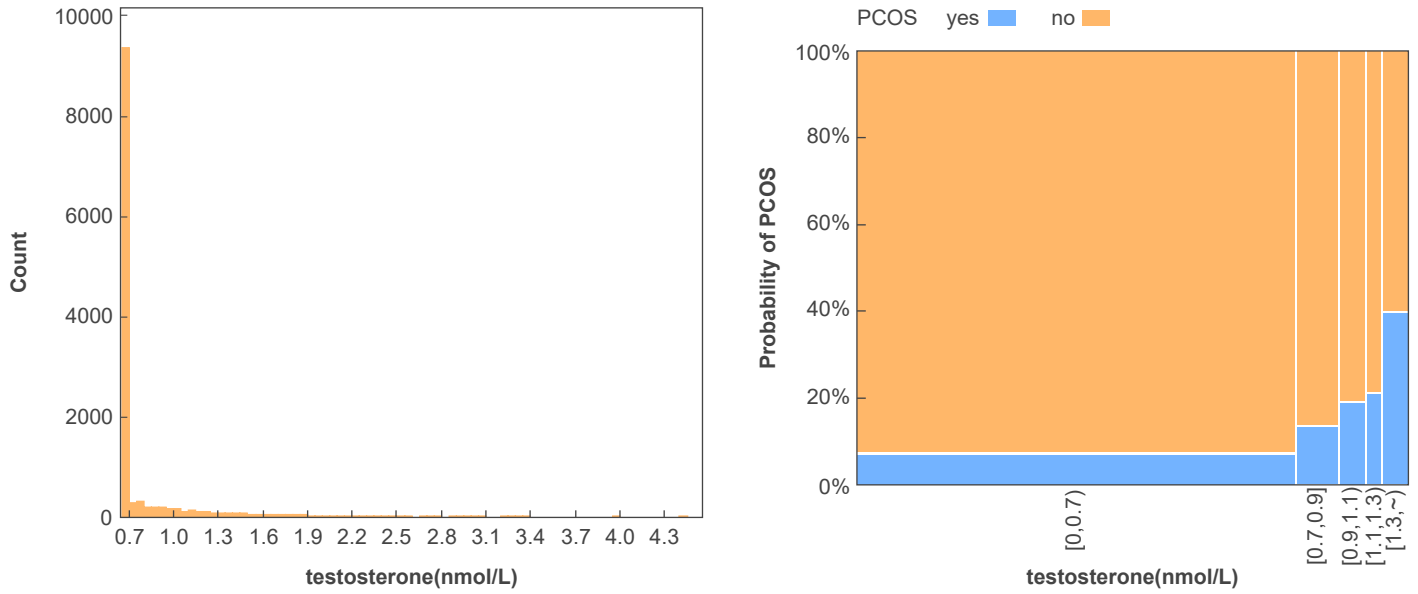

Figure 2. Illustration of the relationship between PCOS and testosterone The left panel shows the skewed distribution of testosterone levels, while the right panel depicts the association between testosterone categorized into five groups and the binary outcome.

4. Finally, if binning is used to enhance interpretability, it is advisable to compare the model's performance with that of spline-based methods to ensure that binning does not significantly diminish model performance.

SPECIFYING APPROPRIATE MODELS

In medical research, the selection of prediction models should consider not only predictive performance but also the critical aspect of interpretability. Highly interpretable models are more likely to be successfully implemented and utilized in clinical settings. Medical prediction models can be broadly categorized into three types based on their flexibility and interpretability: regression-based methods, machine learning techniques, and non-parametric approaches. Figure 3 illustrates commonly used predictive models in medical research.

Regression-based models and machine learning methods

Lasso regression and linear regression models are highly valued for their exceptional interpretability. The clear interpretation of regression parameters makes these models easily understandable for clinicians. However, their primary limitation lies in the inability to automatically capture nonlinear relationships. When the relationship between predictor and outcomes is nonlinear, these models may be prone to prediction bias. Nevertheless, experienced statisticians can address this limitation by combining regression techniques with non-parametric methods, such as generalized additive models, to accommodate nonlinear complexities.

In contrast, machine learning methods offer remarkable flexibility, excelling at handling nonlinear relationships and often outperforming regression models in analyzing complex data. However, a significant drawback is their limited interpretability, as some machine learning methods lack transparency.[54] Although techniques such as SHAP partially enhance interpretability,[55–57] they generally do not achieve the same level of transparency as regression models.

Misconceptions in model selection

A common misconception is that machine learning methods always outperform regression-based models. This topic remains contentious, as some studies comparing machine learning methods to regression models have found no significant advantage for machine learning methods, especially in low-dimensional settings.[58,59] in practical analysis, regression-based models often underperform when thorough data exploration is lacking. However, skilled statisticians can design regression-based models that rival machine learning by conducting comprehensive data exploration, such as addressing nonlinearity, applying variable transformations, and performing

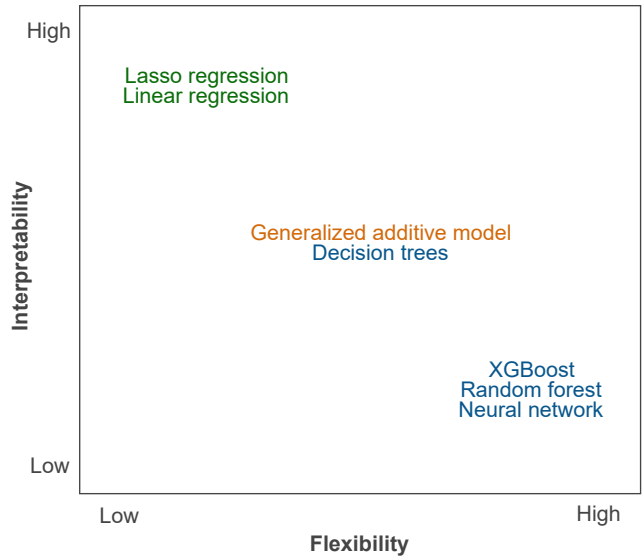

Figure 3. Illustration of the flexibility and interpretability of commonly used models The chart uses green to represent regression-based methods, blue for machine learning methods, and orange for non-parametric methods.

categorization. As machine learning becomes increasingly popular, there is growing concern about its exaggerated capabilities and potential overinterpretation,[60] necessitating caution.

Recommendations

We recommend that the selection of a predictive model be guided by the research objectives. If high interpretability is a priority, lasso regression is advisable. However, analysts must remain vigilant about addressing nonlinearity during the analysis. Conversely, if predictive performance is prioritized over interpretability, machine learning methods may be more suitable. Nevertheless, caution should be taken to avoid overfitting. Given the emphasis on interpretability in clinical settings, we advocate prioritizing regression-based models as the preferred choice when their performance is comparable to machine learning methods, thereby facilitating integration into clinical practice.

Additionally, selecting a model requires thorough consideration of research objectives and application scenarios, such as whether the model is intended

Table 3. Best subset selection results for four predictors

Model	Parameter numbers	R^2	RMSE	AICc	BIC
x_2	1	0.1896	5.9903	248.5413	252.7482
x_1, x_2	2	0.2621	5.7973	247.4881	252.8263
x_1, x_2, x_4	3	0.3482	5.5281	245.4366	251.7495
x_1, x_2, x_3, x_4	4	0.3679	5.5257	247.1038	254.2196

for patent applications or the development of predictive tools. In our previous study involving the *PCOSt* and *POvaStim* models, which are now utilized in clinical settings, we chose lasso regression over machine learning methods for the following reasons:

1. To ensure clinicians could easily understand and apply the model, we opted for algorithms with clear formulas, simplifying patent application and technology transfer.
2. This approach facilitated the development of web-based predictive tools.
3. The performance difference between lasso regression and machine learning methods was minimal.

CONSIDERING REASONABLE STRATEGIES FOR VARIABLE SELECTION

In practical applications of predictive models, variable selection methods are used to reduce model complexity without compromising performance. Instead of incorporating all potential candidate variables, these techniques typically select a subset of variables to optimize the model.[61]

Variable selection challenges

However, selecting the appropriate variable selection method can be challenging due to the wide variety of available approaches.[62] Sanchez-Pinto et al. conducted statistical simulations comparing eight regression- and tree-based variable selection methods.[63] Their findings revealed that regression-based methods achieve better parsimony in smaller datasets with fewer than 20 EPV, while tree-based methods perform better in larger datasets with more than 300 EPV. Hastie et al. compared several regression-based variable selection techniques, including best subset, forward selection, and lasso.[64] They concluded that the best subset method generally outperforms other methods under high signal-to-noise ratio (SNR) conditions, whereas lasso performs better under low SNR conditions. Hanke et al. found that in high-dimensional scenarios with correlated predictors and low SNR, lasso outperformed the best subset method.[65]

When using regression-based methods (e.g., logistic regression), it is important to consider that $R^2_{Nagelkerke}$ tends to be lower in binary models. So we recommend prioritizing the lasso method, particularly when dealing with a large number of predictors and strong correlations between variables. However, it is critical to note that variable selection should not solely rely on statistical techniques but should also incorporate clinical background expertise and evidence from previous literature.

Misconceptions in variable selection

One crucial recommendation is to avoid relying solely on *p*-values for variable selection. Two common misconceptions should be noticed:

First, conducting univariate analysis and subsequently including variables with *p*-values below a pre-set significance threshold (e.g., 0.05 or 0.10) in model construction.

Second, equating *p*-values with a variable's contribution to the model, assuming that smaller *p*-values indicate greater contributions.

These issues can be illustrated with a practical example.

In a study analyzing the relationship between four predictors (x_1, x_2, x_3, x_4) and the outcome, we observed that the *p*-value for x_4 was 0.25 in univariate analysis, suggesting that it could be excluded from further analysis based on a significance threshold of 0.05. However, when applying best subset selection using Bayesian Information Criterion (BIC), Corrected Akaike Information Criterion (AICc), and RMSE, the optimal model included x_1, x_2, and x_4 (Table 3). If we had excluded x4 solely based on its *p*-value from the univariate analysis, we would have missed this optimal model.

Furthermore, it is noteworthy that although the multiple regression results showed x_4 had a lower *p*-value than x_2 (0.04 *vs.* 0.08), the variable importance of x_4 was smaller than that of x_2 (0.29 *vs.* 0.30). This discrepancy highlights the inconsistency between the significance level indicated by *p*-values and the actual variable importance, underscoring the limitations of relying solely on *p*-values for variable selection.

Recommendations

Therefore, irrespective of the selection methods employed, it is not advisable to solely rely on *p*-values for variable selection, particularly in the following situations:

1. Clinical knowledge and previous research suggests that the variable may have a significant impact on the outcome.
2. The predictor exhibits strong correlations with other variables.
3. The predictor has weak correlations with other predictors, but the direction of correlation is consistent with that of other predictors (e.g., all negative or all positive correlations). For instance, in Table 3, there is consistent negative correlations among x_1, x_2, x_3 and x_4, with correlation coefficients of -0.15, -0.15, and -0.11, respectively.

ASSESSING MODEL PERFORMANCE

The assessment of medical prediction models typically involves three key aspects: discrimination, calibration, and clinical utility.[66,67] Discrimination measures the model's ability to distinguish between individuals with and without events, while calibration assesses the consistency between predicted risk and observed risk. An ideal model should exhibit both high discrimination and calibration to improve clinical decision-making.

Although it is essential to assess the discrimination and calibration of a model, current publications appear to be less optimistic. A study by Wessler et al. revealed that out of 796 prediction models, only 63% reported discrimination measure, and merely 36% included calibration evaluation metrics.[68] Similarly, Carrick et al. reviewed 62 validation models and found that while 98% reported discrimination, only 41.9% provided information on calibration.[69] This underscores a notable gap in the comprehensive evaluation of medical prediction models.

Discrimination metrics

Sensitivity and specificity are commonly used metrics for assessing discrimination; however, they are often insufficient when applying to imbalanced datasets with uneven class distributions. Even with high sensitivity and specificity, a model's practical utility is not guaranteed. For instance, as illustrated in Table 4, the sensitivity and specificity are 80% and 96%, respectively. However, the positive predictive value (PPV) is only 50%, indicating that of 100 individuals predicted to experience an event, only 50 will actually be positive.

Moreover, the sensitivity and specificity can vary depending on the decision threshold. For instance, in the *PCOSt* model with a PCOS prevalence of 10.45%, setting the decision threshold at 50% results in sensitivity and specificity of 0.34 and 0.98, respectively. However, when the threshold is adjusted to 10%, the sensitivity and specificity shift to 0.76 and 0.84, respectively. In scenarios with low event rates, a threshold of 0.5 often leads to low sensitivity, but this does not necessarily indicate poor performance. From this perspective, the area under the ROC curve (AUC), which aggregates performance across all thresholds, is often favored in practical applications.

Despite its widespread use, the AUC (also referred to as the *c* statistic) has limitations, particularly in imbalanced datasets.[70] For instance, in a dataset with 42 events and 312 non-events, predicting all subjects as non-events yields an AUC of 0.885, falsely suggesting strong performance. This high-

Table 4. Example of inconsistency between sensitivity and PPV

	True events	True non-events	Total
Predicted events	40	40	80
Predicted non-events	10	910	920
Total	50	950	1000

Table 5. Example of inconsistency between F1 score and specificity

	True events	True non-events	Total
Predicted events	900	80	980
Predicted non-events	10	10	20
Total	910	90	1000

lights the potential for misleading conclusions when relying solely on AUC.

Therefore, while a predictive model may exhibit high sensitivity, specificity, and AUC, these metrics alone do not guarantee robust practical performance. The F1 score combines sensitivity and PPV, partially compensating for the limitations of sensitivity. However, as the F1 score emphasizes events over non-events, it may introduce bias. For instance, as shown in Table 5, elevated sensitivity (0.92) and PPV (0.99) produce a high F1 score of 0.95, while specificity remains extremely low at 0.11. In such cases, the Matthews Correlation Coefficient (MCC) provides a more comprehensive and balanced evaluation of model performance. [71]

Therefore, when evaluating discrimination in imbalanced data, we recommend against relying solely on a single metric. Instead, we suggest using a combination of sensitivity, specificity, AUC, F1 score, and MCC to assess model performance from multiple perspectives.

Calibration metrics

Discrimination alone is insufficient for assessing the prediction performance of a model. Even if a model demonstrates high discrimination, its practical utility remains limited if there is a significant discrepancy between the predicted and actual values. From this perspective, calibration evaluation becomes indispensable. Calibration is typically visualized using a calibration plot, which divides predicted probabilities into a certain number of bins (usually 10) and assesses the consistency between the predicted probabilities and the actual proportion of positive events for each bin. Although the Hosmer-Lemeshow test can serve as a goodness-of-fit test for calibration plot, it is widely discouraged for its limited statistical power and poor interpretability.[72] To address issues arising from excessively large or small sample sizes, modified versions of the Hosmer-Lemeshow test have been proposed.[73,74] Additionally, recent studies have proposed another calibration metric, Integrated Calibration Index (ICI),[75,76] which quantifies calibration for binary outcome by calculating the weighted mean difference between the observed and predicted probabilities.

Combined performance measures

Some performance measures incorporate discrimination and calibration components, such as explained variation (R^2) and Brier score. R^2 is the most commonly used measure for continuous outcomes, while $R^2_{\text{Nagelkerke}}$ variant is typically employed for binary models.[77,78] The Brier score represents the mean squared error between the actual outcomes and the estimated probabilities, with a lower score indicating better model performance.[79] However, it should be noted that a lower Brier score does not necessarily imply higher calibration, as it reflects both discrimination and calibration simultaneously.[80]

Decision curve analysis

While discrimination and calibration are essential for evaluating a prediction model's performance, they do not provide clinicians with insights into their clinical utility. To address this limitation, decision curve analysis (DCA) has been developed as a method to summarize the performance of models in supporting clinical decision-making.[81,82] Decision curves plot the net benefit (NB) of using the model against various probability thresholds, representing levels of predicted risk considered positive and warrant intervention. By comparing the NB of different models, DCA helps clinicians identify which recommendation yields the highest NB and facilitates better clinical decisions.

A key aspect of DCA is understanding NB, which is analogous to profit. In financial terms, net profit is derived by subtracting expenses from income and then multiplying by the exchange rate. In a medical context, NB is calculated as the number of true positives minus the number of false positives, weighted by a factor (p_t/(1-p_t)), where p_t represents the threshold probability. This weight balances the benefits of true positives against the harms of false positives.

The determination of p_t depends on clinical considerations. For instance, in predicting prostate cancer, a p_t exceeding 10% would warrant biopsy intervention, as detecting aggressive cancer outweighs the risk of unnecessary biopsies. However, for the risk of pathologic fractures, surgery might not be considered unless the risk exceeds 25%.

Although DCA is a valuable tool for decision-making, it is primarily applicable to conditions where early intervention significantly impacts outcomes, such as early cancer detection and treatment. However, certain interventions are not solely clinician-driven. For instance, the *OvaRePred* model assesses and predicts ovarian reserve status. If the model identifies diminished ovarian reserve and predicts early onset of perimenopause, the clinician's role is limited to informing the patient, allowing her to decide when to consider pregnancy. In such cases, DCA is unnecessary for facilitating clinical decision-making.

EVALUATING MODEL REPRODUCIBILITY AND TRANSPORTABILITY

The objective of validation is to demonstrate the model's reproducibility and transportability.[83] Reproducibility assesses the model's validity in individuals with similar characteristics who were not part of the initial development cohort. Transportability, on the other hand, evaluates the model's effectiveness when applied to a new patient population with distinct characteristics.[84]

Internal validation

Model validation typically involves both internal and external validation processes.[85,86] Internal validation uses the same dataset as model development, primarily focusing on reproducibility. This process typically involves splitting the sample into two segments: a training set for constructing the model and a validation set for evaluating its validity.[87,88] While random splitting is a straightforward approach, it has been criticized by statisticians.[85] First, with limited sample sizes, splitting the dataset results in smaller subsets, increasing the risk of overfitting and producing unreliable models. Second, random splitting generates different results each time, which may lead researchers to repeat analyses and selectively report the most favorable outcomes.

Resampling techniques, such as cross-validation and bootstrapping, are more appealing for obtaining model stability. Previous studies have shown that bootstrapping is particularly effective, especially with small sample sizes, making it the preferred approach for assessing the reproducibility of prediction models.[89-91]

External validation

Although internal validation is convenient, it often yields overly optimistic results and tends to overstate model performance. Therefore, external validation becomes crucial before deploying a prediction model in clinical settings. However, current research reveals that most predictive models remain at the development stage, with insufficient emphasis on external validation.[92-95] This gap significantly hinders the effective implementation of predictive models in clinical practice.

External validation evaluates the performance of the original model on a new set of patients for whom the prediction model is intended. Geographical validation, widely regarded the preferred method, evaluates the model's transportability across different institutions or regions. However, this method poses significant implementation challenges, requiring extensive collaboration and data collection from multiple units. Temporal validation, often considered as a form of external validation over time, examines the effective-

Table 6. A Summary of model updating methods for regression-based models

Type		Method	Notation
Model updating	Recalibration	Method 1 (Intercept recalibration)	Adjustment of the intercept only, using the calibration intercept
		Method 2 (Logistic recalibration)	Adjustment of the intercept and the regression coefficients, using the calibration intercept and calibration slope
	Revision	Method 3	Adjustment of coefficients for predictors with different strength in updating set compared to original set, after recalibration by method 2
		Method 4	Adding additional predictors, after recalibration by method 2
	Extension	Method 5	Re-estimating all regression coefficients, using the data of updating set
		Method 6	Re-estimating all regression coefficients, and adding additional predictors
Meta model updating			Model averaging using meta technique
Dynamic updating		Periodic updating	Model updating at periodic time point
		Continuous updating	The continuous updating using Bayesian dynamic modeling or dynamic model averaging

ness of previously developed models on subsequent patient cohorts within the same center. For example, data from 2018-2019 could be used for model development, while data from 2020 could serve as the validation dataset. While simpler to implement, temporal validation primarily focuses on reproducibility. As a result, its effectiveness falls between internal and geographical validation.[96]

To ascertain whether external validation measures reproducibility or transportability, one approach is to compare the case mix (i.e., the distribution of predictor values) between the development and validation populations. If the external validation population closely resembles the development group, reproducibility is primarily assessed. Conversely, variations in case mix across different populations provide insights into transportability.[84] In an external validation study,[97] Ramspek et al. discovered that mortality prediction models developed for hemodialysis patients exhibited significantly improved discrimination when applied to a peritoneal dialysis subgroup. The authors attributed this enhanced discrimination mainly to greater case-mix heterogeneity: peritoneal dialysis patients had a wider age range and included individuals ranging from relatively healthy to extremely frail. From this perspective, substantial variation in predictor values enhances the discrimination ability of prediction models.

While the prevailing view advocates for external validation of predictive models in academic publications, some researchers have raised objections.[85] We argue that if the data used for model development is geographically representative of the patient population the prediction model is intended to serve, it is possible to evaluate model's performance even without external validation. For instance, in constructing our predictive models, we obtained data from the world's largest fertility center located in Beijing, where over 60% of patients originated from other provinces. Therefore, we assert that the development data is nationally representative. Even without geographical validation, temporal validation can effectively evaluate the performance of the predictive models in the Chinese population.

UPDATING PREDICTIVE MODELS

When a prediction model fails to demonstrate satisfactory performance in external validation, researchers often choose to discard the model and develop a new one. However, the frequent development of new models not only wastes resources but also poses challenges for physicians in selecting the appropriate model for practical use.[98-100] If a reasonable model exhibits good discrimination (even if slightly miscalibrated), the consensus is to update the existing model and improve its performance with the new data rather than completely redeveloping it.[4,72,101-103]

Three types of regression-based model updating methods have been proposed: model updating (including model recalibration, model revision, and model extension), meta model updating, and dynamic updating (Table 6).

Model updating

Model recalibration involves intercept recalibration and logistic recalibration, primarily aimed at improving calibration rather than discrimination.[103,104] In cases where there are differences in outcome incidence between the development and validation datasets, adjusting only the intercept of the original model (intercept recalibration) can effectively address poor calibration. Logistic recalibration, on the other hand, is typically employed when regression coefficients in the development model are overfitted, often due to the inclusion of too many predictors. Recalibration is generally sufficient if the sample size of the new dataset is relatively small and the case mix is similar between the development and new datasets, making calibration the primary concern.[84,105,106]

Model revision and model extension can improve both calibration and discrimination and are suitable when there are substantial differences between the development and new datasets.[100,104,107] If some predictors exhibit different effects in the new dataset compared to the original dataset, method 3 involves re-estimating the coefficients based on logistic recalibration, while method 4 extends the original model by incorporating additional predictors based on logistic recalibration. If methods 3 and 4 prove inadequate, more extensive revisions may be required. Method 5 re-estimates the intercept and regression coefficients for all predictors using the new datasets, while method 6 allows for the inclusion of new predictors. Both methods 4 and 6 incorporate new predictors, and the performance of the extended model can be evaluated using net benefit measures or decision curve.[108,109]

When adding new variables to a model, it is important to compare its performance with that of the previous model, which can be evaluated using Net Reclassification Index (NRI). Although the NRI has gained popularity in recent years, some researchers have argued that it is not always helpful in evaluating new risk models and, in some cases, can even be misleading.[110-113] Therefore, it should be used with caution. Additionally, it is not recommended to evaluate model performance based solely on *p*-values. Kattan MW reported that even if certain variables are statistically significant in multivariable analysis, this does not necessarily imply an improvement in the predictive power of the model.[114]

Meta model updating

In cases where multiple historical prediction models have been published for the same or similar endpoints and populations (e.g., breast cancer and traumatic brain injury prediction models),[98,99] meta-analysis techniques can be employed to combine them into a meta-model. By integrating these models, the resulting meta-model may exhibit enhanced performance compared to individual models and facilitate generalization across diverse populations. Although literature on meta-models remains limited, some discussions on this topic are available.[115-117]

Dynamic updating

Dynamic updating refers to the continuous updating of a prediction model, in contrast to models derived from a single time point.[100,102,104] Consequently, the coefficients of an updated model exhibit temporal variability. For periodically updated models, conventional methods such as model recalibration and revision can be employed.[118,119] For continuously updated models, suggested approaches include Bayesian dynamic modeling or dynamic model averaging.[120,121] Previous studies have demonstrated that dynamically updating a clinical prediction model leads to improved performance compared to not updating the model.[119] However, the choice of updating method and frequency depends on factors such as data availability (e.g., sample size, event rate), and model complexity.[119,122]

CONDUCTING IMPACT EVALUATION

Even if a predictive model demonstrates high discrimination and calibration, it remains uncertain whether it will ultimately influence the behavior or decisions of physicians and patients in clinical practice. Therefore, conducting an impact study becomes imperative. Impact studies aim to quantify the effect of using a prediction model on physicians' behaviors, patient outcomes, or cost-effectiveness compared to not employing such a model.[101,123]

Study design

The preferred study design for an impact study is a randomized controlled trial (RCT), typically a cluster RCT.[124] Specifically, healthcare professionals (as cluster) were randomly assigned to either employ the predictive model (risk-based group) or not (control group). Randomizing centers is more preferred, as it avoids contamination within the same center. Although impact studies are relatively scarce, some notable examples of cluster RCTs exist.[125,126] However, cluster RCTs have certain drawbacks, such as requiring larger sample sizes and difficulties in maintaining balance between risk-based and control groups due to variations in patient numbers and characteristics across different physicians. In such cases, a stepped-wedge design offers an attractive alternative. In this design, physicians or centers are randomly assigned a time period to adopt the predictive model.[127,128] Because the time intervals for transitioning from the control to the intervention group are regularly randomized, the stepped-wedge design ensures better balance between groups.

In impact evaluation, given the significant cost and time required for RCTs, we recommend initially considering more cost-effective designs, such as before-and-after studies and real-world studies, when feasible. If positive results are observed, then an RCT could be conducted. In a before-and-after study, comparisons are made between outcomes measured before and after implementing the model within the same care providers.[72,101] The main limitation of this design is its sensitivity to temporal changes in treatment methods. However, if the desired outcomes are achieved in the short term, the before-and-after study remains valid, as the level of care and patient characteristics are unlikely to undergo significant changes. Additionally, implementing real-world studies utilizing routinely collected data is a cost-effective strategy. Although real-world data are prone to confounding factors, this challenge can be effectively addressed through appropriate statistical methods, such as propensity score techniques.

Study outcomes

The outcomes of an impact study can include process variables, such as changes in physicians' decisions, or patient outcomes, such as hospital stay duration or mortality rates; and cost-effectiveness measures, such as changes in costs. It is recommended to prioritize studying changes in physicians' behavior, as patient outcomes depend on these behavioral changes.[123] In addition, evaluating changes in physicians' behavior does not require follow-up, making it a more feasible approach compared to evaluating patient outcomes. For example, in an impact study predicting the risk of perspective nausea and vomiting (PONV), both the administration of prophylactic antiemetics and the incidence of PONV were compared.[124] The study found increased administration of prophylactic antiemetics in the risk-based group, despite no significant decrease in PONV incidence. This suggests that while the intervention may not directly improve patient outcomes, it effectively influenced physician behavior.

We conducted an impact study based on real-world data for the *POvaStim* model. In this study, we used the propensity score matching (PSM) method to compare physicians who voluntarily used the model with those who did not, assessing changes in physicians' dosages decisions and patient outcomes (the number of oocytes retrieved). The unpublished results of this study showed promising finding; therefore, we plan to conduct a multicenter RCT to obtain more robust evidence.

PROMOTING APPLICATION OF MODELS

Even though a model may perform excellently in external validation and be published, this does not guarantee its adoption by clinicians. In fact, only a limited number of published predictive models are actually implemented in clinical settings. Beyond demonstrating good performance, strong promotional efforts are also essential for fostering the widespread adoption of a model in clinical practice.

We propose that a model undergo five key stages from development to clinical application: first, achieving the good discrimination and calibration during internal validation; second, demonstrating the transportability and generalization through external validation; third, confirming the model's impact on clinicians' decision-making through impact evaluation; fourth, assessing the model's impact on patient outcomes through impact evaluation; and finally, promoting the model with the support of leading figures in the field.

For a well-performing predictive model, successful implementation heavily relies on support of top experts in the relevant field, who can significantly influence its promotion. For example, the *POvaStim* model has been incorporated into an expert consensus document and published in a prestigious Chinese obstetrics and gynecology journal. This endorsement has led to its adoption in dozens of hospitals and health examination centers across China. This success is attributed to our development team, which includes leading clinical researchers in reproductive health in China, thereby facilitating widespread adoption of the model in various Chinese healthcare institutions.

CONCLUSIONS

In recent years, predictive models and artificial intelligence have rapidly advanced in the field of medicine,[129,130] but their clinical applications remain challenging. We present twelve key recommendations for the development and clinical implementation of predictive models, offering not only a comprehensive framework but also detailed guidance for each stage. It is essential to recognize that a good predictive model is not merely the product of statistical software; instead, it demands thoughtful consideration at every stage, as potential pitfalls may arise throughout the process. Moreover, we emphasize that publishing a predictive model represents only a partial validation, as its effective integration into clinical practice requires addressing additional practical considerations. We believe that the recommendations outlined in this article will assist clinical researchers enhance the quality of model development and application, ultimately promoting better clinical decision-making.

REFERENCES

1. Wynants, L., Van Calster, B., Collins, G.S., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ **369**: m1328. DOI: 10.1136/bmj.m1328.
2. Liu, Y., Feng, W., Lou, J., et al. (2023). Performance of a prediabetes risk prediction model: A systematic review. Heliyon **9**: e15529. DOI: 10.1016/j.heliyon.2023.e15529.
3. Kaiser, I., Mathes, S., Pfahlberg, A.B., et al. (2022). Using the prediction model risk of bias assessment tool (PROBAST) to evaluate melanoma prediction studies. Cancers **14**: 3033. DOI: 10.3390/cancers14123033.
4. Collins, G.S., Reitsma, J.B., Altman, D.G., et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Br. J. Cancer **112**: 251–259. DOI: 10.1038/bjc.2014.639.
5. Xu, H., Feng, G., Yang, R., et al. (2023). OvaRePred: Online tool for predicting the age of fertility milestones. The Innovation **4**: 100490. DOI: 10.1016/j.xinn.2023.100490.
6. Xu, H., Feng, G., Shi, L., et al. (2023). PCOSt: A non-invasive and cost-effective screening tool for polycystic ovary syndrome. The Innovation **4**: 100407. DOI: 10.1016/j.xinn.2023.100407.
7. Xu, H., Feng, G., Han, Y., et al. (2023). POvaStim: An online tool for directing individualized FSH doses in ovarian stimulation. The Innovation **4**: 100401. DOI: 10.1016/j.xinn.2023.100401.

8. Steyerberg, E.W., and Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. Eur. Heart J. **35**: 1925–1931. DOI: 10.1093/eurheartj/ehu207.
9. Liao, Y., McGee, D.L., Cooper, R.S., et al. (1999). How generalizable are coronary risk prediction models? Comparison of Framingham and two national cohorts. Am. Heart J. **137**: 837–845. DOI: 10.1016/s0002-8703(99)70407-2.
10. Xu, H., Feng, G., Ma, C., et al. (2023). AMHconverter: An online tool for converting results between the different anti-Müllerian hormone assays of Roche Elecsys®, Beckman Access, and Kangrun. PeerJ **11**: e15301. DOI: 10.7717/peerj.15301.
11. Xu, H., Feng, G., Wang, H., et al. (2020). A novel mathematical model of true ovarian reserve assessment based on predicted probability of poor ovarian response: A retrospective cohort study. J. Assist. Reprod. Genet. **37**: 963–972. DOI: 10.1007/s10815-020-01700-1.
12. Xu, H., Shi, L., Feng, G., et al. (2020). An ovarian reserve assessment model based on anti-müllerian hormone levels, follicle-stimulating hormone levels, and age: Retrospective cohort study. J. Med. Internet Res. **22**: e19096. DOI: 10.2196/19096.
13. Han, Y., Xu, H., Feng, G., et al. (2022). An online tool for predicting ovarian reserve based on AMH level and age: A retrospective cohort study. Front. Endocrinol. **13**: 946123. DOI: 10.3389/fendo.2022.946123.
14. Xu, H., Feng, G., Alpadi, K., et al. (2022). A model for predicting polycystic ovary syndrome using serum AMH, menstrual cycle length, body mass index and serum androstenedione in Chinese reproductive aged population: A retrospective cohort study. Front. Endocrinol. **13**: 821368. DOI: 10.3389/fendo.2022.821368.
15. Zhang, X., Xu, H., Feng, G., et al. (2023). Sensitive HPLC-DMS/MS/MS method coupled with dispersive magnetic solid phase extraction followed by in situ derivatization for the simultaneous determination of multiplexing androgens and 17-hydroxyprogesterone in human serum and its application to patients with polycystic ovarian syndrome. Clin. Chim. Acta **538**: 221–230. DOI: 10.1016/j.cca.2022.11.025.
16. Xu, H., Zhang, X., Yang, R., et al. (2023). Can androgens be replaced by AMH in initial screening of Polycystic Ovary Syndrome? The Innovation Medicine **1**: 100010. DOI: 10.59717/j.xinn-med.2023.100010.
17. Vittinghoff, E., and McCulloch, C.E. (2007). Relaxing the rule of ten events per variable in logistic and cox regression. Am. J. Epidemiol. **165**: 710–718. DOI: 10.1093/aje/kwk052.
18. van Smeden, M., Moons, K.G.M., de Groot, J.A.H., et al. (2018). Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat. Methods Med. Res. **28**: 2455–2474. DOI: 10.1177/0962280218784726.
19. Steyerberg, E.W., Schemper, M., and Harrell, F.E. (2011). Logistic regression modeling and the number of events per variable: Selection bias dominates. J. Clin. Epidemiol. **64**: 1464–1465. DOI: 10.1016/j.jclinepi.2011.06.016.
20. van Smeden, M., de Groot, J.A.H., Moons, K.G.M., et al. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med. Res. Methodol. **16**: 163. DOI: 10.1186/s12874-016-0267-3.
21. Ogundimu, E.O., Altman, D.G., and Collins, G.S. (2016). Adequate sample size for developing prediction models is not simply related to events per variable. J. Clin. Epidemiol. **76**: 175–182. DOI: 10.1016/j.jclinepi.2016.02.031.
22. Austin, P.C., and Steyerberg, E.W. (2017). 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. **26**: 796–808. DOI: 10.1177/0962280214558972.
23. Wynants, L., Bouwmeester, W., Moons, K.G., et al. (2015). A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. J. Clin. Epidemiol. **68**: 1406–1414. DOI: 10.1016/j.jclinepi.2015.02.002.
24. Riley, R.D., Ensor, J., Snell, K.I.E., et al. (2020). Calculating the sample size required for developing a clinical prediction model. BMJ **368**: m441. DOI: 10.1136/bmj.m441.
25. Riley, R.D., Snell, K.I.E., Ensor, J., et al. (2018). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat. Med. **38**: 1276–1296. DOI: 10.1002/sim.7992.
26. Riley, R.D., Snell, K.I.E., Ensor, J., et al. (2019). Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. Stat. Med. **38**: 1262–1275. DOI: 10.1002/sim.7993.
27. Dhiman, P., Ma, J., Andaur Navarro, C.L., et al. (2022). Methodological conduct of prognostic prediction models developed using machine learning in oncology: A systematic review. BMC Med. Res. Methodol. **22**: 101. DOI: 10.1186/s12874-022-01577-x.
28. van der Ploeg, T., Austin, P.C., and Steyerberg, E.W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. BMC Med. Res. Methodol. **14**: 137. DOI: 10.1186/1471-2288-14-137.
29. Riley, R.D., Snell, K.I.E., Archer, L., et al. (2024). Evaluation of clinical prediction models (part 3): Calculating the sample size required for an external validation study. BMJ **384**: e074821. DOI: 10.1136/bmj-2023-074821.
30. Riley, R.D., Debray, T.P.A., Collins, G.S., et al. (2021). Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat. Med. **40**: 4230–4251. DOI: 10.1002/sim.9025.
31. Chevret, S., Seaman, S., and Resche-Rigon, M. (2015). Multiple imputation: A mature approach to dealing with missing data. Intensive Care Med. **41**: 348–350. DOI: 10.1007/s00134-014-3624-x.
32. Fletcher Mercaldo, S., and Blume, J.D. (2020). Missing data and prediction: The pattern submodel. Biostatistics **21**: 236–252. DOI: 10.1093/biostatistics/kxy040.
33. Sainani, K.L. (2015). Dealing with missing data. Pm&R **7**: 990–994. DOI: 10.1016/j.pmrj.2015.07.011.
34. Sterne, J.A., White, I.R., Carlin, J.B., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ **338**: b2393. DOI: 10.1136/bmj.b2393.
35. Steif, J., Brant, R., Sreepada, R.S., et al. (2021). Prediction model performance with different imputation strategies: A simulation study using a north American ICU registry. Pediatr. Crit. Care Med. **23**: e29–e44. DOI: 10.1097/pcc.0000000000002835.
36. Moons, K.G.M., Grobbee, D.E., Chen, Q., et al. (2009). Dealing with missing predictor values when applying clinical prediction models. Clin. Chem. **55**: 994–1001. DOI: 10.1373/clinchem.2008.115345.
37. Eekhout, I., de Boer, R.M., Twisk, J.W.R., et al. (2012). Missing data. Epidemiology **23**: 729–732. DOI: 10.1097/EDE.0b013e3182576cdb.
38. Nijman, S.W.J., Leeuwenberg, A.M., Beekers, I., et al. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. J. Clin. Epidemiol. **142**: 218–229. DOI: 10.1016/j.jclinepi.2021.11.023.
39. Steyerberg, E.W. (2019). Dealing with missing values. (ed). Clinical prediction models: A practical approach to development, validation, and updating (Springer Nature), pp: 113-128. DOI: 10.1007/978-0-387-77244-8.
40. Zeng, H., Ran, X., An, L., et al. (2021). Disparities in stage at diagnosis for five common cancers in China: A multicentre, hospital-based, observational study. Lancet Public Health **6**: e877–e887. DOI: 10.1016/s2468-2667(21)00157-2.
41. Ciccione, L., Dehaene, G., and Dehaene, S. (2023). Outlier detection and rejection in scatterplots: Do outliers influence intuitive statistical judgments. J. Exp. Psychol. Hum. Percept. Perform. **49**: 129–144. DOI: 10.1037/xhp0001065.
42. Lakra, A., Banerjee, B., and Laha, A. (2023). A data-adaptive method for outlier detection from functional data. Stat. Comput. **34**. DOI: 10.1007/s11222-023-10301-8.
43. Yang, J., Tan, X., and Rahardja, S. (2023). Outlier detection: How to select k for k-nearest-neighbors-based outlier detectors. Pattern Recogn. Lett. **174**: 112–117. DOI: 10.1016/j.patrec.2023.08.020.
44. Smiti, A. (2020). A critical overview of outlier detection methods. Comput. Sci. Rev. **38**: 100306. DOI: 10.1016/j.cosrev.2020.100306.
45. Guan, L., and Tibshirani, R. (2022). Prediction and outlier detection in classification problems. J. R. Stat. Soc. B **84**: 524–546. DOI: 10.1111/rssb.12443.
46. El-Masri, M.M., Mowbray, F.I., Fox-Wasylyshyn, S.M., et al. (2020). Multivariate outliers: A conceptual and practical overview for the nurse and health researcher. Can. J. Nurs. Res. **53**: 316–321. DOI: 10.1177/0844562120932054.
47. Sauerbrei, W., Royston, P., and Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat. Med. **26**: 5512–5528. DOI: 10.1002/sim.3148.
48. Binder, H., Sauerbrei, W., and Royston, P. (2013). Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: A simulation study with continuous response. Stat. Med. **32**: 2262–2277. DOI: 10.1002/sim.5639.
49. Nieboer, D., Vergouwe, Y., Roobol, M.J., et al. (2015). Nonlinear modeling was applied thoughtfully for risk prediction: The prostate biopsy collaborative group. J. Clin. Epidemiol. **68**: 426–434. DOI: 10.1016/j.jclinepi.2014.11.022.
50. Ma, J., Dhiman, P., Qi, C., et al. (2023). Poor handling of continuous predictors in clinical prediction models using logistic regression: A systematic review. J. Clin. Epidemiol. **161**: 140–151. DOI: 10.1016/j.jclinepi.2023.07.017.
51. Royston, P., Altman, D.G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. Stat. Med. **25**: 127–141. DOI: 10.1002/sim.2331.
52. Altman, D.G., and Royston, P. (2006). The cost of dichotomising continuous variables. BMJ **332**: 1080. DOI: 10.1136/bmj.332.7549.1080.
53. Collins, G.S., Ogundimu, E.O., Cook, J.A., et al. (2016). Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat. Med. **35**: 4124–4135. DOI: 10.1002/sim.6986.
54. Zhou, J., You, D., Bai, J., et al. (2023). Machine learning methods in real-world studies of cardiovascular disease. Cardiovascular Innovations and Applications **7**: 975. DOI: 10.15212/cvia.2023.0011.
55. Martin, S.A., Townend, F.J., Barkhof, F., et al. (2023). Interpretable machine learning for dementia: A systematic review. Alzheimer's & Dementia **19**: 2135–2149. DOI: 10.1002/alz.12948.
56. Wang, K., Tian, J., Zheng, C., et al. (2021). Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. Comput. Biol. Med. **137**: 104813. DOI: 10.1016/j.compbiomed.2021.104813.
57. Yi, F., Yang, H., Chen, D., et al. (2023). XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. BMC Med. Inform. Decis. Mak. **23**: 137. DOI: 10.1186/s12911-023-02238-9.
58. Gravesteijn, B.Y., Nieboer, D., Ercole, A., et al. (2020). Machine learning algorithms performed no better than regression models for prognostication in traumatic brain

injury. J. Clin. Epidemiol. **122**: 95–107. DOI: 10.1016/j.jclinepi.2020.03.005.
59. Christodoulou, E., Ma, J., Collins, G.S., et al. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J. Clin. Epidemiol. **110**: 12–22. DOI: 10.1016/j.jclinepi.2019.02.004.
60. Dhiman, P., Ma, J., Andaur Navarro, C.L., et al. (2023). Overinterpretation of findings in machine learning prediction model studies in oncology: A systematic review. J. Clin. Epidemiol. **157**: 120–133. DOI: 10.1016/j.jclinepi.2023.03.012.
61. Chowdhury, M.Z.I., and Turin, T.C. (2020). Variable selection strategies and its importance in clinical prediction modelling. Fam. Med. Community Health **8**: e000262. DOI: 10.1136/fmch-2019-000262.
62. Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. Biom. J. **60**: 431–449. DOI: 10.1002/bimj.201700067.
63. Sanchez-Pinto, L.N., Venable, L.R., Fahrenbach, J., et al. (2018). Comparison of variable selection methods for clinical predictive modeling. Int. J. Med. Inform. **116**: 10–17. DOI: 10.1016/j.ijmedinf.2018.05.006.
64. Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Statist. Sci. **35**: 579-592. DOI: 10.1214/19-sts733.
65. Hanke, M., Dijkstra, L., Foraita, R., et al. (2023). Variable selection in linear regression models: Choosing the best subset is not always the best choice. Biom. J. **66**: e2200209. DOI: 10.1002/bimj.202200209.
66. Strandberg, R., Jepsen, P., and Hagström, H. (2024). Developing and validating clinical prediction models in hepatology – An overview for clinicians. J. Hepatol. **81**: 149-162. DOI: 10.1016/j.jhep.2024.03.030.
67. Alba, A.C., Agoritsas, T., Walsh, M., et al. (2017). Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. JAMA **318**: 1377-1384. DOI: 10.1001/jama.2017.12126.
68. Wessler, B.S., Lai Yh, L., Kramer, W., et al. (2015). Clinical Prediction Models for Cardiovascular Disease. Circulation: Cardiovascular Quality and Outcomes **8**: 368–375. DOI: 10.1161/circoutcomes.115.001693.
69. Carrick, R.T., Park, J.G., McGinnes, H.L., et al. (2020). Clinical Predictive Models of Sudden Cardiac Arrest: A Survey of the Current Science and Analysis of Model Performances. Journal of the American Heart Association **9**. DOI: 10.1161/jaha.119.017625.
70. Nahm, F.S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. Korean Journal of Anesthesiology **75**: 25–36. DOI: 10.4097/kja.21209.
71. Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**. DOI: 10.1186/s12864-019-6413-7.
72. Cowley, L.E., Farewell, D.M., Maguire, S., et al. (2019). Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. Diagnostic and Prognostic Research **3**. DOI: 10.1186/s41512-019-0060-y.
73. Yu, W., Xu, W., and Zhu, L. (2017). A Modified Hosmer–Lemeshow Test for Large Data Sets. Communications in Statistics - Theory and Methods **46**. DOI: 10.1080/03610926.2017.1285922.
74. Paul, P., Pennell, M.L., and Lemeshow, S. (2013). Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Stat Med **32**: 67–80. DOI: 10.1002/sim.5525.
75. Austin, P.C., Harrell, F.E., and van Klaveren, D. (2020). Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine **39**: 2714–2742. DOI: 10.1002/sim.8570.
76. Austin, P.C., and Steyerberg, E.W. (2019). The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine **38**: 4051–4065. DOI: 10.1002/sim.8281.
77. Royston, P., and Altman, D.G. (2013). External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol **13**: 33. DOI: 10.1186/1471-2288-13-33.
78. Steyerberg, E.W., Vickers, A.J., Cook, N.R., et al. (2010). Assessing the Performance of Prediction Models. Epidemiology **21**: 128–138. DOI: 10.1097/EDE.0b013e3181c30fb2.
79. Huang, Y., Li, W., Macheret, F., et al. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association **27**: 621–633. DOI: 10.1093/jamia/ocz228.
80. Rufibach, K. (2010). Use of Brier score to assess binary predictions. Journal of Clinical Epidemiology **63**: 938–939. DOI: 10.1016/j.jclinepi.2009.11.009.
81. Vickers, A.J., and Holland, F. (2021). Decision curve analysis to evaluate the clinical benefit of prediction models. The Spine Journal **21**: 1643–1648. DOI: 10.1016/j.spinee.2021.02.024.
82. Van Calster, B., Wynants, L., Verbeek, J.F.M., et al. (2018). Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. European Urology **74**: 796–804. DOI: 10.1016/j.eururo.2018.08.038.
83. Justice, A.C., Covinsky, K.E., and Berlin, J.A. (1999). Assessing the generalizability of prognostic information. Ann Intern Med **130**: 515–524. DOI: 10.7326/0003-4819-130-6-199903160-00016.
84. Ramspek, C.L., Jager, K.J., Dekker, F.W., et al. (2021). External validation of prognostic models: what, why, how, when and where. Clinical Kidney Journal **14**: 49–58. DOI: 10.1093/ckj/sfaa188.
85. Collins, G.S., Dhiman, P., Ma, J., et al. (2024). Evaluation of clinical prediction models (part 1): from development to external validation. Bmj. DOI: 10.1136/bmj-2023-074819.
86. Riley, R.D., Archer, L., Snell, K.I.E., et al. (2024). Evaluation of clinical prediction models (part 2): how to undertake an external validation study. Bmj. DOI: 10.1136/bmj-2023-074820.
87. Moons, K.G.M., Kengne, A.P., Woodward, M., et al. (2012). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart **98**: 683–690. DOI: 10.1136/heartjnl-2011-301246.
88. Staffa, S.J., and Zurakowski, D. (2021). Statistical Development and Validation of Clinical Prediction Models. Anesthesiology **135**: 396–405. DOI: 10.1097/aln.0000000000003871.
89. Steyerberg, E.W., Harrell, F.E., Jr., Borsboom, G.J., et al. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol **54**: 774–781. DOI: 10.1016/s0895-4356(01)00341-9.
90. Steyerberg, E.W., and Harrell, F.E. (2016). Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology **69**: 245–247. DOI: 10.1016/j.jclinepi.2015.04.005.
91. Steyerberg, E.W., Bleeker, S.E., Moll, H.A., et al. (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. Journal of Clinical Epidemiology **56**: 441–447. DOI: 10.1016/s0895-4356(03)00047-7.
92. Macleod, M.R., Bouwmeester, W., Zuithoff, N.P.A., et al. (2012). Reporting and Methods in Clinical Prediction Research: A Systematic Review. PLOS Medicine **9**. DOI: 10.1371/journal.pmed.1001221.
93. Mallett, S., Royston, P., Waters, R., et al. (2010). Reporting performance of prognostic models in cancer: a review. BMC Medicine **8**. DOI: 10.1186/1741-7015-8-21.
94. Steyerberg, E.W., Moons, K.G.M., van der Windt, D.A., et al. (2013). Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLOS Medicine **10**. DOI: 10.1371/journal.pmed.1001381.
95. Riley, R.D., Ensor, J., Snell, K.I.E., et al. (2016). External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. DOI: 10.1136/bmj.i3140.
96. Altman, D.G., Vergouwe, Y., Royston, P., et al. (2009). Prognosis and prognostic research: validating a prognostic model. BMJ **338**: b605–b605. DOI: 10.1136/bmj.b605.
97. Ramspek, C., Voskamp, P., van Ittersum, F., et al. (2017). Prediction models for the mortality risk in chronic dialysis patients: a systematic review and independent external validation study. Clinical Epidemiology Volume **9**: 451–464. DOI: 10.2147/clep.S139748.
98. Phung, M.T., Tin Tin, S., and Elwood, J.M. (2019). Prognostic models for breast cancer: a systematic review. BMC Cancer **19**. DOI: 10.1186/s12885-019-5442-6.
99. Perel, P., Edwards, P., Wentz, R., et al. (2006). Systematic review of prognostic models in traumatic brain injury. BMC Medical Informatics and Decision Making **6**. DOI: 10.1186/1472-6947-6-38.
100. Toll, D.B., Janssen, K.J.M., Vergouwe, Y., et al. (2008). Validation, updating and impact of clinical prediction rules: A review. Journal of Clinical Epidemiology **61**: 1085–1094. DOI: 10.1016/j.jclinepi.2008.04.008.
101. Moons, K.G.M., Kengne, A.P., Grobbee, D.E., et al. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. Heart **98**: 691–698. DOI: 10.1136/heartjnl-2011-301247.
102. Binuya, M.A.E., Engelhardt, E.G., Schats, W., et al. (2022). Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Medical Research Methodology **22**. DOI: 10.1186/s12874-022-01801-8.
103. Janssen, K.J.M., Moons, K.G.M., Kalkman, C.J., et al. (2008). Updating methods improved the performance of a clinical prediction model in new patients. Journal of Clinical Epidemiology **61**: 76–86. DOI: 10.1016/j.jclinepi.2007.04.018.
104. Su, T.L., Jaki, T., Hickey, G.L., et al. (2018). A review of statistical updating methods for clinical prediction models. Statistical Methods in Medical Research **27**: 185–197. DOI: 10.1177/0962280215626466.
105. Nieboer, D., Vergouwe, Y., Ankerst, D.P., et al. (2016). Improving prediction models with new markers: a comparison of updating strategies. BMC Medical Research Methodology **16**. DOI: 10.1186/s12874-016-0231-2.
106. van Houwelingen, H.C. (2000). Validation, calibration, revision and combination of prognostic survival models. Stat Med **19**: 3401–3415. DOI: 10.1002/1097-0258(20001230)19:24<3401::aid-sim554>3.0.co;2-2.
107. Siregar, S., Nieboer, D., Versteegh, M.I.M., et al. (2019). Methods for updating a risk prediction model for cardiac surgery: a statistical primer. Interactive CardioVascular and Thoracic Surgery **28**: 333–338. DOI: 10.1093/icvts/ivy338.
108. Pencina, M.J., D'Agostino, R.B., and Steyerberg, E.W. (2010). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine **30**: 11–21. DOI: 10.1002/sim.4085.
109. Steyerberg, E.W., Pencina, M.J., Lingsma, H.F., et al. (2011). Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. European Journal of Clinical Investigation **42**: 216–228. DOI: 10.1111/j.1365-2362.2011.02562.x.

110. Pepe, M.S., Fan, J., Feng, Z., et al. (2014). The Net Reclassification Index (NRI): A Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. Statistics in Biosciences **7**: 282–295. DOI: 10.1007/s12561-014-9118-0.
111. Kerr, K.F. (2023). Net Reclassification Index Statistics Do Not Help Assess New Risk Models. Radiology **306**. DOI: 10.1148/radiol.222343.
112. Grunkemeier, G.L., and Jin, R. (2015). Net Reclassification Index: Measuring the Incremental Value of Adding a New Risk Factor to an Existing Risk Model. The Annals of Thoracic Surgery **99**: 388–392. DOI: 10.1016/j.athoracsur.2014.10.084.
113. Burch, P.M., Glaab, W.E., Holder, D.J., et al. (2016). Net Reclassification Index and Integrated Discrimination Index Are Not Appropriate for Testing Whether a Biomarker Improves Predictive Performance. Toxicological Sciences. DOI: 10.1093/toxsci/kfw225.
114. Kattan, M.W. (2003). Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst **95**: 634–635. DOI: 10.1093/jnci/95.9.634.
115. Debray, T.P.A., Riley, R.D., Rovers, M.M., et al. (2015). Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. PLOS Medicine **12**. DOI: 10.1371/journal.pmed.1001886.
116. Debray, T.P.A., Koffijberg, H., Nieboer, D., et al. (2014). Meta - analysis and aggregation of multiple published prediction models. Statistics in Medicine **33**: 2341–2362. DOI: 10.1002/sim.6080.
117. Debray, T.P., Moons, K.G., Ahmed, I., et al. (2013). A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med **32**: 3158–3180. DOI: 10.1002/sim.5732.
118. Hickey, G.L., Grant, S.W., Caiado, C., et al. (2013). Dynamic Prediction Modeling Approaches for Cardiac Surgery. Circulation: Cardiovascular Quality and Outcomes **6**: 649–658. DOI: 10.1161/circoutcomes.111.000012.
119. Schnellinger, E.M., Yang, W., and Kimmel, S.E. (2021). Comparison of dynamic updating strategies for clinical prediction models. Diagnostic and Prognostic Research **5**. DOI: 10.1186/s41512-021-00110-w.
120. McCormick, T.H., Raftery, A.E., Madigan, D., et al. (2011). Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. Biometrics **68**: 23–30. DOI: 10.1111/j.1541-0420.2011.01645.x.
121. Jenkins, D.A., Sperrin, M., Martin, G.P., et al. (2018). Dynamic models to predict health outcomes: current status and methodological challenges. Diagnostic and Prognostic Research **2**. DOI: 10.1186/s41512-018-0045-2.
122. Siregar, S., Nieboer, D., Vergouwe, Y., et al. (2016). Improved Prediction by Dynamic Modeling. Circulation: Cardiovascular Quality and Outcomes **9**: 171–181. DOI: 10.1161/circoutcomes.114.001645.
123. Moons, K.G.M., Altman, D.G., Vergouwe, Y., et al. (2009). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ **338**: b606–b606. DOI: 10.1136/bmj.b606.
124. Kappen, T.H., van Klei, W.A., van Wolfswinkel, L., et al. (2018). Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagnostic and Prognostic Research **2**. DOI: 10.1186/s41512-018-0033-6.
125. Meyer, G., Köpke, S., Bender, R., et al. (2005). Predicting the risk of falling – efficacy of a risk assessment tool compared to nurses' judgement: a cluster-randomised controlled trial [ISRCTN37794278]. BMC Geriatrics **5**. DOI: 10.1186/1471-2318-5-14.
126. Foy, R., Penney, G.C., Grimshaw, J.M., et al. (2004). A randomised controlled trial of a tailored multifaceted strategy to promote implementation of a clinical guideline on induced abortion care. Bjog **111**: 726–733. DOI: 10.1111/j.1471-0528.2004.00168.x.
127. Grayling, M.J., Wason, J.M.S., and Mander, A.P. (2017). Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. Trials **18**. DOI: 10.1186/s13063-017-1783-0.
128. Li, F., and Wang, R. (2022). Stepped Wedge Cluster Randomized Trials: A Methodological Overview. World Neurosurgery **161**: 323–330. DOI: 10.1016/j.wneu.2021.10.136.
129. Huang, T., Xu, H., Wang, H., et al. (2023). Artificial intelligence for medicine: Progress, challenges, and perspectives. The Innovation Medicine **1**. DOI: 10.59717/j.xinn-med.2023.100030.
130. Liu, X., Zhang, S., Shao, L., et al. (2024). Improving prediction of treatment response and prognosis in colorectal cancer with AI-based medical image analysis. The Innovation Medicine **2**. DOI: 10.59717/j.xinn-med.2024.100069.

FUNDING AND ACKNOWLEDGMENTS

This study was supported by National Key Research and Development Program of China (grant number 2023YFC2705501, 2023YFC2705504]); and the Innovation & Transfer Fund of Peking University Third Hospital (Grant No. BYSYZHKC2023102, and BYSYZHZB2020102). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

G.F. wrote and supervised the manuscript. H.X., S.W., H.W., X.F., R.M., Y.H., Y.W., and H.G. edited the manuscript. All authors contributed to the article and approved the submitted version.

DECLARATION OF INTERESTS

Guoshuang Feng is Editorial Board member of The Innovation Medicine. He was blinded from reviewing or making final decisions on the manuscript. The article was subject to the journal's standard procedures, with peer review handled independently of these Editorial Board members and their research groups. The other authors declare no competing interests.

ETHICAL STATEMENT AND PATIENT CONSENT

Not applicable.

DATA AND CODE AVAILABILITY

Not applicable.