Statistics and Its Interface Volume 0 (2024) 1

Development and evaluation of the updated risk prediction model involving in new candidate predictors*

Hong Xu, Ya Qi Cao†,

In biomedical research, risk prediction models play a pivotal role in forecasting various health outcomes. However, traditional evaluation metrics, such as the Area Under the Receiver Operating Characteristic Curve (AUC), exhibit notable limitations. Specifically, once the AUC reaches a high value in baseline model, it becomes increasingly difficult to achieve substantial improvements, as any further increases in the AUC necessitate unrealistically large effect sizes from newly incorporated variables.This paper proposes a comprehensive evaluation framework that incorporates, in addition to AUC, two advanced metrics: the Integrated Discrimination Improvement (IDI) and the Net Reclassification Improvement (NRI). These metrics provide a more nuanced assessment of the incremental value contributed by new predictors. We explore the theoretical properties of IDI and NRI, demonstrating that both metrics belong to the family of U-statistics and possess desirable asymptotic properties.Furthermore, we apply this evaluation framework to a study of cardiovascular disease risk in patients with polycystic ovary syndrome (PCOS). Our results demonstrate that the inclusion of IDI and NRI, alongside AUC, allows for a more comprehensive understanding of model performance and the added value of novel predictors.

Keywords and phrases: Model discrimination, Risk prediction model, IDI, NRI, U-statistics.

1. INTRODUCTION

Advances in technology and the increasing availability of new biomarkers have led to continuous refinement of risk prediction models, thereby enhancing their predictive accuracy. These models are now widely applied across various medical domains, including oncology, cardiovascular diseases (CVD) such as coronary artery disease, stroke, atrial fibrillation, and diabetes ([11, 40, 35, 6]). In the context of risk prediction, researchers typically focus on two key aspects of model performance ([8]): calibration and discrimination. Calibration assesses the accuracy of predicted probabilities for the future occurrence of an event in individuals, reflecting the alignment between predicted and actual risks. Discrimination, on the other hand, involves categorizing individuals above a certain risk threshold as high-risk, thereby enabling the model to effectively distinguish between different risk groups.

To further enhance the performance of these models, researchers are increasingly focused on incorporating new predictors. The challenge of selecting predictors from a large pool of candidates to maximize model performance has become a central concern in medical statistics. As a result, it is crucial to conduct comprehensive evaluations of risk prediction models after the inclusion of new predictors. [17] proposed a multi-stage evaluation approach for assessing the inclusion of new risk predictors. [19] utilized a risk stratification framework to evaluate the incremental impact of incorporating breast density as a predictor in breast cancer risk models. Furthermore, quantifying the contribution of new predictors to the improvement of model performance remains a vital research direction in risk prediction modeling ([32, 5, 41]).

In the field of cardiovascular disease research, the Framingham Heart Study, initiated in 1948, is widely regarded as one of the longest-running and most influential epidemiological studies. [20] proposed a multivariable risk scoring system for predicting coronary heart disease (CHD) risk, which incorporates factors such as age, gender, smoking status, systolic blood pressure, serum total cholesterol levels, electrocardiographic evidence of left ventricular hypertrophy, and impaired glucose tolerance. Building upon this, [23] introduced the estimation of left ventricular mass via echocardiography as an additional predictor, which significantly improved the accuracy of cardiovascular risk prediction compared to traditional models based solely on clinical risk factors.In the domain of cancer research, [15] developed a breast cancer risk prediction model that integrated several key variables, including age at first live birth ("Agefb"), age at menarche ("Agemen"), number of prior breast biopsies ("Nbiops"), and the number of first-degree relatives ("Numrel") diagnosed with breast cancer. This model represented a significant advancement in the ability to stratify breast cancer risk. Subsequently, [2] incorporated breast density as a biomarker in risk prediction and demonstrated that the inclusion of this factor improved the model's performance

arXiv: 00000
*we b
†Corresponding author.

in identifying high-risk women when compared to the Gail model. Additionally, a model incorporating breast density was used to predict absolute cancer risk in white women, as shown by [4].

When assessing the discriminatory ability of risk prediction models, the Area Under the Curve (AUC) is the most commonly employed metric ([25]). Conceptually, the AUC quantifies the area under the Receiver Operating Characteristic (ROC) curve, bounded by the horizontal axis. Statistically, when computed using the trapezoidal rule, the AUC is equivalent to the Mann-Whitney U statistic, which is used to compare the value distributions of two samples ([16]). This equivalence highlights the statistical properties of the AUC. Furthermore, AUC can be interpreted as the probability that a randomly selected positive sample will have a higher predicted score than a randomly selected negative sample, as classified by the model ([1]).The change in AUC (ΔAUC) serves as a key indicator of the added discriminatory power when evaluating the performance of model. For example, in a study on Community Acquired Pneumonia (CAP), the inclusion of C-Reactive Protein (CRP), Procalcitonin (ProCT), and copeptin levels in a baseline model (which included age, gender, smoking history, antibiotic pretreatment, and leukocyte count) resulted in AUC values of 0.59 for the baseline, increasing to 0.61, 0.68, and 0.75, respectively, highlighting the effectiveness of copeptin as a predictor ([28]). Similarly, the inclusion of breast density in the Gail model for breast cancer risk raised the AUC from 0.607 to 0.631 ([2]). However, comparing AUC values alone is insufficient; statistical significance must be assessed. The DeLong test, based on the generalized U statistic, is commonly used to evaluate differences in AUC between models ([9]).

It is common practice to calculate the difference in AUC between models with and without a new predictor. However, such differences are often marginal. For example, [12] showed that adding a new predictor to a standard cardiovascular disease (CVD) risk model increased the AUC from 0.76 to 0.77, a modest change of only 0.01. Numerical simulations by [33] and [39], using pancreatic cancer data, further confirmed that substantial increases in the Odds Ratio (OR) are typically required to significantly improve AUC. Additionally, [32] demonstrated that when the baseline model already possesses a high AUC, incorporating a strong predictor results in minimal improvements. As a result, statisticians have explored alternative methods for quantifying model improvement. Reclassification tables have become increasingly popular in clinical settings ([34, 7]). For instance, the addition of parental history of myocardial infarction and C-Reactive Protein (CRP) to a female CVD risk prediction model increased the AUC from 0.805 to 0.808. However, reclassification of individuals into different risk categories revealed that approximately 30% changed categories, which led to recalculated event rates, providing more clinically relevant information ([34]).

To better assess model improvement, [31] introduced two new metrics: Integrated Discrimination Improvement (IDI) and Net Reclassification Improvement (NRI). The IDI, based on sensitivity and specificity, is defined as:

$$\text{IDI} = (\text{IS}_{new} - \text{IS}_{old}) - (\text{IP}_{new} - \text{IP}_{old}). \tag{1}$$

where IS is the integral of "sensitivity" over all cutoff values, and IP is the corresponding integral of "1 minus specificity".In this equation, "new" refers to the model including the new predictor, and 'old" refers to the baseline model. The NRI, calculated using reclassification tables, is defined as:

$$\begin{aligned}\text{NRI} = {} & p(\text{up} \mid \text{events}) - p(\text{down} \mid \text{events}) \\ & + p(\text{down} \mid \text{nonevents}) - p(\text{up} \mid \text{nonevents}).\end{aligned} \tag{2}$$

where "up" and "down" refer to individuals moving up or down in risk categories. An effective new predictor increases risk for cases and decreases it for non-cases, leading to a substantial IDI ([21]). Under the assumption of multivariate normality in linear discriminant analysis, AUC, IDI, and NRI can all be expressed as functions of the squared Mahalanobis distance ([30]). The NRI value primarily reflects the association strength between the new predictor and the outcome, independent of the baseline model's performance ([31]), and is often considered a general measure of effect size ([30]). IDI, on the other hand, measures the difference in discrimination slopes between the new and baseline models ([42]). [36] established the relationship between IDI and the coefficient of determination ($\mathbf{R^2}$) in binary response models. [3] extended these metrics to survival analysis, providing empirical analyses in the Atherosclerosis Risk In Communities (ARIC) study.

When incorporating new predictors into risk prediction models, it is crucial not only to conduct statistical significance tests but also to demonstrate their contribution to improving model performance. However, much of the existing literature relies heavily on AUC as the sole measure of model improvement, which may fail to capture meaningful gains, particularly when the baseline AUC is already high. This study introduces menstrual dysfunction severity as a novel predictor for cardiometabolic risk among women with polycystic ovary syndrome (PCOS), a prevalent endocrine and metabolic disorder in women of reproductive age ([26]). Previous research has linked menstrual irregularities with hyperinsulinemia and insulin resistance (IR), highlighting the potential of menstrual history as a significant marker for metabolic dysfunction risk ([14, 13]). We utilizes the proposed evaluation framework to assess the added value of menstrual dysfunction severity in risk prediction. While these metrics have been increasingly applied, the statistical properties of IDI and NRI have been insufficiently explored in the existing literature. This work addresses this gap by proving that both IDI and NRI are U-statistics and deriving their asymptotic properties, thereby providing a more comprehensive assessment of risk prediction model.

2. PREDICTIVE PERFORMANCE MEASURES IN PREDICTION MODEL

2.1 Risk prediction model

Let Y denote the binary response variable of interest, where $Y = 1$ represents cases and $Y = 0$ represents noncases. Let $\boldsymbol{X}$ denote the conventional predictors. Subsequently, we consider the incorporation of the new predictors $\boldsymbol{Z}$. Considering the logistic model as the risk prediction model, it is thus feasible to obtain the risk prediction value for each individual based on a risk model.Let the risk value of an individual be denoted as the random variable R.Given a threshold ν, where $0 < \nu < 1$, if $R > \nu$, the individual is classified as a positive class; conversely, if $R < \nu$, the individual is classified as a negative class.

Here, we denote "model 0" as the risk model containing only conventional predictors, and "model 1" as the new model after incorporating the new predictor. **Model 0**:

(3)

$$\Pr\left(Y = 1 \mid \mathbf{X}; \boldsymbol{\theta}_0\right) \equiv p_0\left(\mathbf{X}; \boldsymbol{\theta}_0\right) = \frac{\exp\left(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right)}{1 + \exp\left(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right)},$$

where $\boldsymbol{\theta}_0 = (\beta_0, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ with superscript "T" indicating vector transpose. Let $F_X(\mathbf{x})$ denote the cumulative distribution function (CDF) for $\boldsymbol{X}$. Then the true positive rate (TPR) and false positive rate (FPR) at risk cutoff ν, written as TPR_ν^0 and FPR_ν^0 under Model 0, can be expressed as:

$$\begin{aligned}
\mathrm{TPR}_\nu^0 &= \Pr\left(p_0\left(\mathbf{X}; \boldsymbol{\theta}_0\right) > \nu \mid Y = 1\right) \\
&= \frac{\Pr\left(p_0\left(\mathbf{X}; \boldsymbol{\theta}_0\right) > \nu, Y = 1\right)}{\Pr(Y = 1)} \\
&= \frac{\int \mathbb{I}\left\{p_0\left(\mathbf{x}; \boldsymbol{\theta}_0\right) > \nu\right\} p_0\left(\mathbf{x}; \boldsymbol{\theta}_0\right) dF_X(\mathbf{x})}{\int p_0\left(\mathbf{x}; \boldsymbol{\theta}_0\right) dF_X(\mathbf{x})}, \qquad (4) \\
\mathrm{FPR}_\nu^0 &= \Pr\left(p_0\left(\mathbf{X}; \boldsymbol{\theta}_0\right) > \nu \mid Y = 0\right) \\
&= \frac{\Pr\left(p_0\left(\mathbf{X}; \boldsymbol{\theta}_0\right) > \nu, Y = 0\right)}{\Pr(Y = 0)} \\
&= \frac{\int \mathbb{I}\left\{p_0\left(\mathbf{x}; \boldsymbol{\theta}_0\right) > \nu\right\} \left(1 - p_0\left(\mathbf{x}; \boldsymbol{\theta}_0\right)\right) dF_X(\mathbf{x})}{\int \left(1 - p_0\left(\mathbf{x}; \boldsymbol{\theta}_0\right)\right) dF_X(\mathbf{x})}. \qquad (5)
\end{aligned}$$

To assess the predictive value of new factor $\boldsymbol{Z}$, we incorporated it into the existing risk prediction model, resulting in an updated model, referred to as Model 1.

Model 1:

(6)

$$\Pr\left(Y = 1 \mid \mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_1\right) \equiv p_1\left(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}_1\right) = \frac{\exp\left(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}\right)}{1 + \exp\left(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{Z}\right)},$$

where $\boldsymbol{\theta}_1 = (\beta_0, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ with superscript "T" indicating vector transpose. Let $F_{X,Z}(\mathbf{x}, \mathbf{z})$ denote the cumulative distribution function (CDF) for $(\boldsymbol{X}, \boldsymbol{Z})$. Then the true positive rate (TPR) and false positive rate (FPR) at risk cutoff ν, written as TPR_ν^1 and FPR_ν^1 under Model1, can be expressed as:

(7)

$$\mathrm{TPR}_\nu^1 = \frac{\int \mathbb{I}\left\{p_1\left(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1\right) > \nu\right\} p_1\left(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1\right) dF_{X,Z}(\mathbf{x}, \mathbf{z})}{\int p_1\left(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1\right) dF_{X,Z}(\mathbf{x}, \mathbf{z})},$$

(8)

$$\mathrm{FPR}_\nu^1 = \frac{\int \mathbb{I}\left\{p_1\left(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1\right) > \nu\right\} \left(1 - p_1\left(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1\right)\right) dF_{X,Z}(\mathbf{x}, \mathbf{z})}{\int \left(1 - p_1\left(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1\right)\right) dF_{X,Z}(\mathbf{x}, \mathbf{z})}.$$

The AUC of the baseline model and the new model can therefore be calculated as follows:

$$\begin{aligned}
\mathrm{AUC}_0 &= \int_0^1 (\mathrm{TPR}_v^0) d(\mathrm{FPR}_v^0), \\
\mathrm{AUC}_1 &= \int_0^1 (\mathrm{TPR}_v^1) d(\mathrm{FPR}_v^1),
\end{aligned}$$

The predictive performance of the new predictor can be assessed by analyzing the difference in AUC between the two models, denoted as ΔAUC:

$$\begin{aligned}
\Delta\mathrm{AUC} &= \mathrm{AUC}_1 - \mathrm{AUC}_0, \\
&= \int_0^1 (\mathrm{TPR}_v^1) d(\mathrm{FPR}_v^1) - \int_0^1 (\mathrm{TPR}_v^0) d(\mathrm{FPR}_v^0).
\end{aligned}$$

2.2 Integrated Discrimination Improvement:IDI

Integrated discrimination improvement(IDI) shows the difference between the value of mean change in predicted probability between the group of cases and noncases when a new factor of mean change in predicted probability between the group of ill and healthy women when a new factor is added to the model.

The IDI can be estimated through the following equation:

$$\begin{aligned}
\widehat{\mathrm{IDI}} &= \left(\bar{\hat{p}}_{\mathrm{new,events}} - \bar{\hat{p}}_{\mathrm{old,events}}\right) \\
&\quad - \left(\bar{\hat{p}}_{\mathrm{new,nonevents}} - \bar{\hat{p}}_{\mathrm{old,nonevents}}\right), \qquad (9)
\end{aligned}$$

where $\bar{\hat{p}}_{new,events}$ represents the average predicted risk for all case individuals based on the new model, $\bar{\hat{p}}_{old,events}$ represents the average predicted risk for all case individuals based on the old model; $\bar{\hat{p}}_{new,nonevents}$ represents the average predicted risk for all non-case individuals based on the new model, $\bar{\hat{p}}_{old,nonevents}$ represents the average predicted risk for all non-case individuals based on the old model.
By rearranging equation (9), we obtain:

$$\begin{aligned}
\widehat{\mathrm{IDI}} &= \left(\bar{\hat{p}}_{\mathrm{new,events}} - \bar{\hat{p}}_{\mathrm{new,nonevents}}\right) \\
&\quad - \left(\bar{\hat{p}}_{\mathrm{old,events}} - \bar{\hat{p}}_{\mathrm{old,nonevents}}\right). \qquad (10)
\end{aligned}$$

The integration of sensitivity and "1-specificity" over (0,1) can be regarded as the average effect of both, thus the IDI is also considered as the difference between the average "sensitivity" and the average "1-specificity" of the two models. It can also be interpreted as the integrated difference of

Youden Indices ([43]).Yates proposed that the discrimination slope is the difference in the average predicted probabilities between the case and non-case populations. Therefore, equation (10) also represents the difference in the discrimination slopes between the new and old models. The IDI is often used to measure the change in performance of a new predictor after it is added to a baseline model.

In this context, let IS denote the integral of sensitivity over the (0,1) interval for all thresholds, and IP represent the integral of "1-specificity" over the (0,1) interval for all thresholds.Based on the TPR_v and FPR_v of the baseline and new models, we derive the expression for the IDI to evaluate the improvement introduced by adding the new predictors $\boldsymbol{Z}$ to the model. Thus, we derive the IDI its integral form.

$$
\begin{aligned}
\mathrm{IDI} &= (\mathrm{IS}_{new} - \mathrm{IS}_{old}) - (\mathrm{IP}_{new} - \mathrm{IP}_{old}) \\
&= \left\{ \int_0^1 \mathrm{TPR}_\nu^1 \, d\nu - \int_0^1 \mathrm{TPR}_\nu^0 \, d\nu \right\} \\
&\quad - \left\{ \int_0^1 \mathrm{FPR}_\nu^1 \, d\nu - \int_0^1 \mathrm{FPR}_\nu^0 \, d\nu \right\} \\
&= \left\{ \int_0^1 \frac{\int \mathbb{I}\left\{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1) > \nu\right\} p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1) \, dF_{X,Z}(\mathbf{x},\mathbf{z})}{\int p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1) \, dF_{X,Z}(\mathbf{x},\mathbf{z})} \, d\nu \right. \\
&\quad \left. - \int_0^1 \frac{\int \mathbb{I}\left\{p_0(\mathbf{x};\boldsymbol{\theta}_0) > \nu\right\} p_0(\mathbf{x};\boldsymbol{\theta}_0) \, dF_X(\mathbf{x})}{\int p_0(\mathbf{x};\boldsymbol{\theta}_0) \, dF_X(\mathbf{x})} \, d\nu \right\} \\
&\quad - \left\{ \int_0^1 \frac{\int \mathbb{I}\left\{p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1) > \nu\right\} (1 - p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1)) dF_{X,Z}(\mathbf{x},\mathbf{z})}{\int (1 - p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1)) dF_{X,Z}(\mathbf{x},\mathbf{z})} \, d\nu \right. \\
&\quad \left. - \int_0^1 \frac{\int \mathbb{I}\left\{p_0(\mathbf{x};\boldsymbol{\theta}_0) > \nu\right\} (1 - p_0(\mathbf{x};\boldsymbol{\theta}_0)) dF_X(\mathbf{x})}{\int (1 - p_0(\mathbf{x};\boldsymbol{\theta}_0)) dF_X(\mathbf{x})} \, d\nu \right\}.
\end{aligned} \tag{11}
$$

2.3 Net Reclassification Improvement:NRI

The Net Reclassification Improvement (NRI) index is used to compare the predictive capabilities of two risk models, focuses on the reclassification table describing the number of women in whom an upward or downward shift in the disease probability value occurred after a new factor had been added, including the results for three outcomes. In the definition given by equation (2), an "up" classification occurs when the risk score assigned to an individual by the new model is higher than that assigned by the old model, while a "down" classification occurs when it is lower. An NRI greater than 0 indicates an improvement in the predictive ability of the new model compared to the old model; an NRI less than 0 suggests a decline in predictive performance; and an NRI equal to 0 indicates no change. For each risk prediction model, we can assign a risk score to each individual, allowing us to compare the differences in risk scores between the baseline model and the new model.

The definition of NRI in equation (2) originally relies on discrete, pre-defined risk categories, making it sensitive to the choice and number of these categories. However,in cases where no established categories exist, it is more prudent to use a version of NRI which does not require categories. A category-less or continuous NRI is a more objective and versatile measure, especially in fields where risk categories are not well-established([29]).We use "$\mathrm{NRI}^{>0}$" denote the continuous NRI,it gives an interpretation as a summary measure quantifying the correct upward versus downward movement in model based predicted probabilities for events and nonevents.

$$
\begin{aligned}
\mathrm{NRI}^{>0} = 2\big\{ &\Pr(\hat{p}_{new,events} > \hat{p}_{old,events}) \\
&- \Pr(\hat{p}_{new,nonevents} > \hat{p}_{old,nonevents}) \big\}.
\end{aligned} \tag{12}
$$

Where the $\hat{p}_{new,events}$ denotes the risk score for all cases under the new model, while $\hat{p}_{old,events}$ represents the corresponding score under the baseline model. Similarly, $\hat{p}_{new,nonevents}$ indicates the risk score for non-cases under the new model, and $\hat{p}_{old,nonevents}$ reflects the score under the baseline model. This comparison forms the basis for the reclassification scenario in the new model, from which we derive the expression for NRI as follows:

$$
\begin{aligned}
\mathrm{NRI}^{>0} &= 2\big\{ p(\mathrm{up} \mid \mathrm{events}) - p(\mathrm{up} \mid \mathrm{nonevents}) \big\} \\
&= 2\big\{ \Pr\left(p_1(\mathbf{X},\mathbf{Z};\boldsymbol{\theta}_1) > p_0(\mathbf{X};\boldsymbol{\theta}_0) \mid Y = 1\right) \\
&\quad - \Pr\left(p_1(\mathbf{X},\mathbf{Z};\boldsymbol{\theta}_1) > p_0(\mathbf{X};\boldsymbol{\theta}_0) \mid Y = 0\right) \big\} \\
&= 2 \left\{ \frac{\int \mathbb{I}\big(p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1) > p_0(\mathbf{x};\boldsymbol{\theta}_0)\big) p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1) dF_{X,Z}(\mathbf{x},\mathbf{z})}{\int p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1) dF_{X,Z}(\mathbf{x},\mathbf{z})} \right. \\
&\quad \left. - \frac{\int \mathbb{I}\big(p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1) > p_0(\mathbf{x};\boldsymbol{\theta}_0)\big) (1 - p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1)) dF_{X,Z}(\mathbf{x},\mathbf{z})}{\int (1 - p_1(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_1)) dF_{X,Z}(\mathbf{x},\mathbf{z})} \right\}.
\end{aligned} \tag{13}
$$

3. ASYMPTOTIC PROPERTY BASED ON GENERALIZED U-STATISTICS

3.1 $\widehat{\mathrm{IDI}}$ and $\widehat{\mathrm{NRI}}^{>0}$ belong to the generalized U-Statistics family

U-statistics theory can be regarded as an extension of the Central Limit Theorem (CLT) to sums of correlated variables ([22]). The CLT states that the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables, each with finite mean and variance, will tend to follow a normal distribution, regardless of the original distribution of the variables. In contrast, U-statistics are a class of statistics used to estimate population parameters, often formulated by averaging a function applied to all possible combinations of a fixed number of sample elements. Unlike the sum of i.i.d. random variables, U-statistics involve sums where the terms may be correlated, as they are functions of overlapping subsets of the data. U-statistics theory provides extensions of the CLT to handle such correlated sums, demonstrating that, under certain

conditions, U-statistics can converge to a normal distribution, even when the variables involved are correlated.

Generalized U-statistics can be constructed for samples drawn from k different distributions, where k is unrestricted ([22]). For this study, we focus on the two-sample case $k = 2$) of generalized U-statistics.

In the context of a risk prediction model, we incorporate new predictors, $\mathbf{Z}$, with the conventional factors, $\mathbf{X}$. The combined predictors are denoted as $\mathbf{W} = (\mathbf{X}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$, indicating that $\mathbf{W}$ includes both conventional and new predictors. When considering only the baseline model, $\mathbf{W}$ degenerates to include only the conventional predictors, $\mathbf{X}$. Consider the sets $\{\mathbf{W}_1^1, \mathbf{W}_2^1, \ldots, \mathbf{W}_{n_1}^1\}$ and $\{\mathbf{W}_1^0, \mathbf{W}_2^0, \ldots, \mathbf{W}_{n_0}^0\}$, which consist of n_1 and n_0 observations, respectively. These sets are drawn from two distinct distribution functions,which the $\mathbf{W}^1$ from the distribution of case population,the $\mathbf{W}^0$ from the non-case population, which the observations being independent within each sample. The generalized U-statistics is then defined as follows:

$$U_{n_1,n_0} = \binom{n_1}{c}^{-1} \binom{n_0}{d}^{-1} \sum_{(n_1,c)} \sum_{(n_0,d)} \psi\left(\mathbf{W}_{i_1}^1, \ldots, \mathbf{W}_{i_c}^1; \mathbf{W}_{j_1}^0, \ldots, \mathbf{W}_{j_d}^0\right), \tag{14}$$

where the sum $\sum_{(n_1,c)}$ is taken over all permutations $(i_1, i_2, ..., i_c)$ of $\{1, 2, ..., n_1\}$, $\sum_{(n_0,d)}$ is taken over all permutations $(j_1, j_2, ..., j_d)$ of $\{1, 2, ..., n_0\}$. The $\psi(\cdot;\cdot)$ is a real-valued kernel symmetric in each set of arguments and c and d are dimension constants,which satisfied $1 \le c \le n_1, 1 \le d \le n_0$. We consider $c = d = 1$,the symmetry condition is automatically satisfied and formula (14) simplifies to:

$$U_{n_1,n_0} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \psi\left(\mathbf{W}_i^1; \mathbf{W}_j^0\right). \tag{15}$$

In this context, under the new model,$p_1(\mathbf{W}^1)$ represents the risk value for a case, while $p_1(\mathbf{W}^0)$ represents the risk for a non-case. Similarly,under the baseline model, $p_0(\mathbf{W}^1)$ denotes the risk value for a case, and $p_0(\mathbf{W}^0)$ represents the risk for a non-case. We derive the estimators for IDI and NRI based on the generalized U-statistics framework, demonstrating that these estimators belong to the generalized U-statistics family with different kernel function. From the above analysis, we obtain the estimator for IDI as follows:

$$\begin{aligned}\widehat{\mathrm{IDI}} =& (\bar{\hat{p}}_{new,events} - \bar{\hat{p}}_{new,nonevents}) - (\bar{\hat{p}}_{old,events} - \bar{\hat{p}}_{old,nonevents}) \\ =& \frac{\sum_{i=1}^{n_1} p_1(\mathbf{W}_i^1) - p_0(\mathbf{W}_i^1)}{n_1} - \frac{\sum_{j=1}^{n_0} p_1(\mathbf{W}_j^1) - p_0(\mathbf{W}_j^0)}{n_0} \\ =& \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \left[p_1(\mathbf{W}_i^1) - p_0(\mathbf{W}_i^1)\right] - \left[p_1(\mathbf{W}_j^0) - p_0(\mathbf{W}_j^0)\right].\end{aligned} \tag{16}$$

Denote kernel as $\psi_{IDI}(\mathbf{W}_i^1, \mathbf{W}_j^0) = \left[p_1(\mathbf{W}_i^1) - p_0(\mathbf{W}_i^1)\right] - \left[p_1(\mathbf{W}_j^0) - p_0(\mathbf{W}_j^0)\right]$. It is evident that the $\widehat{\mathrm{IDI}}$ fulfills the criteria outlined in the definition of generalized two-sample U-statistics. Consequently, $\widehat{\mathrm{IDI}}$ can be classified within the generalized U-statistics family.Similarly, we derive the following form for the estimator of NRI:

$$\begin{aligned}\widehat{\mathrm{NRI}}^{>0} &= 2\left(\frac{\#\{\text{up,events}\}}{n_1} - \frac{\#\{\text{up,nonevents}\}}{n_0}\right) \\ &= \frac{2}{n_1 n_0}\left(n_0 \cdot \#\{\text{up,events}\} - n_1 \cdot \#\{\text{up,nonevents}\}\right) \\ &= \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} 2\left(\mathbb{I}\left[p_1(\mathbf{W}_i^1) > p_0(\mathbf{W}_i^1)\right]\right. \\ &\quad \left. - \mathbb{I}\left[p_1(\mathbf{W}_j^0) > p_0(\mathbf{W}_j^0)\right]\right).\end{aligned} \tag{17}$$

where "#{up,events}" and "#{up,nonevents}" denote the number of "up" in all cases and non-cases, respectively. We define the kernel as the $\psi_{NRI^{>0}}(\mathbf{W}_i^1, \mathbf{W}_j^0) = 2\left(\mathbb{I}[p_1(\mathbf{W}_i^1) > p_0(\mathbf{W}_i^1)] - \mathbb{I}[p_1(\mathbf{W}_j^0) > p_0(\mathbf{W}_j^0)]\right)$.The $\mathrm{NRI}^{>0}$ satisfies the conditions specified in the definition of generalized two-sample U-statistics, thus confirming its membership in generalized the U-statistic family.

3.2 Asymptotically normal distribution and Consistency

The generalized two-sample U-statistic can be decomposed into a sum of several components, each of which consists of sums of i.i.d. random variables. This decomposition process is referred to as the H-decomposition. The H-decomposition partitions the U-statistic into uncorrelated components that decrease in order with respect to the sample size, which is crucial for the asymptotic theory of U-statistics. By breaking down the U-statistic into these simpler, uncorrelated components, the H-decomposition facilitates both theoretical derivations and practical calculations, particularly when handling large sample sizes. Building on the generalized U-statistic and its H-decomposition, we can derive the distributions of $\widehat{\mathrm{IDI}}$ and $\widehat{\mathrm{NRI}}^{>0}$.

Lemma

$$\sqrt{\mathbf{N}}(\widehat{IDI} - \mathbb{E}[IDI]) \xrightarrow{D} N\left(0, N\sigma_{IDI}^2\right), \tag{18}$$

$$\sqrt{\mathbf{N}}(\widehat{NRI}^{>0} - \mathbb{E}[NRI^{>0}]) \xrightarrow{D} N\left(0, N\sigma_{NRI^{>0}}^2\right). \tag{19}$$

where, the variancen of the IDI and $\mathrm{NRI}^{>0}$ as follows:

$$\hat{\sigma}_{IDI}^2 = \frac{\mathrm{Var}[p_1(\mathbf{W}_i^1) - p_0(\mathbf{W}_i^1)]}{n_1} + \frac{\mathrm{Var}[p_1(\mathbf{W}_j^0) - p_0(\mathbf{W}_j^0)]}{n_0}, \tag{20}$$

$$\hat{\sigma}^2_{NRI^{>0}} = 4\Big(\frac{\mathrm{Var}[\mathbb{I}[p_1(\mathbf{W}_i^1) - p_0(\mathbf{W}_i^1)]]}{n_1} + \frac{\mathrm{Var}[\mathbb{I}[p_1(\mathbf{W}_j^0) - p_0(\mathbf{W}_j^0)]]}{n_0}\Big).$$

4. PREDICATING CVD RISK IN PCOS WOMEN

Polycystic ovary syndrome (PCOS) is a prevalent endocrine and metabolic disorder affecting women of reproductive age, with a global prevalence ranging from 5% to 20% ([27]). PCOS is associated with various cardiometabolic risk factors, including obesity, insulin resistance (IR), dyslipidemia, hypertension, and metabolic syndrome ([18, 37]), all of which contribute to an increased risk of cardiovascular events, type 2 diabetes, myocardial infarction, and stroke. Additionally, women with PCOS exhibit a higher incidence of coronary artery disease and reduced cardiovascular event-free survival, leading to an elevated risk of premature mortality ([38]). These findings underscore the importance of assessing cardiometabolic risk in clinical practice for women with PCOS.

Recent studies have highlighted a connection between the severity of menstrual irregularities and both hyperinsulinemia and insulin resistance (IR), suggesting that menstrual history could be a straightforward yet clinically important marker of metabolic dysfunction risk ([13]). Building on the established link between menstrual cycle irregularities and cardiovascular disease in PCOS patients, the focus of our work is on developing and assessing a new prediction model that incorporates menstrual cycle information to improve CVD risk prediction in women with PCOS. Our data were collected from the Reproductive Medicine Center at Shunde Hospital of Southern Medical University (The First People's Hospital of Shunde). We included 154 patients aged 18 to 40 years who were diagnosed with PCOS based on the Rotterdam criteria between July 2021 and September 2022.

In the analysis of the association between menstrual cycle patterns and cardiometabolic risk markers in women with PCOS, participants were categorized into groups based on the length of their intervals between episodes of vaginal bleeding. Women with bleeding intervals of 26–34 days were classified as eumenorrheic (Eumeno, n = 23), those with intervals of 35 days to 3 months were classified as oligomenorrheic (Oligo, n = 74), and those with intervals exceeding 3 months per year were classified as amenorrheic (Ameno, n = 57). For the analysis involving the Eumeno and Ameno groups, the menstrual cycle was treated as a binary variable (0–1) and included in a logistic regression model where IFG/IGT served as the outcome variable. The odds ratio (OR) for this model was 4.8 (95% CI: [1.45-22.25]). We developed three different models for three separate outcome variables, each incorporating menstrual cycle grouping as a new predictor. The OR values and their 95% confidence intervals for these new predictors are presented in Table 6.

In our predictive model, we initially employed the proposed comprehensive dimensions to assess the value of novel predictors. The baseline model included only the variables age and BMI. Subsequently, we considered the addition of waist circumference (WC), free androgen index (FAI), and categorized menstrual cycle data as new predictors, thereby forming an updated model. We focused on three cardiovascular disease (CVD) risk-related outcomes: insulin resistance (IR), prediabetes—defined by impaired fasting glucose (IFG) and impaired glucose tolerance (IGT)—and dyslipidemia. All three outcome variables were binary, classified according to clinical standards as detailed in [24].

Table 1 presents our risk prediction model setting. To evaluate the value of these new predictors under this scenario, we calculated the AUC, IDI, and continuous NRI. Compared to the basic prediction model (AUC_0=0.684), AUC increased most when E.A. group was added to the model which IFGIGT as the outcome (AUC_1=0.744). The values of IDI and NRI were also the highest for this model, with IDI=0.055 and NRI=0.533.

In contrast, when IR was the outcome, the addition of FAI as a new predictor resulted in only a marginal improvement (Δ AUC = 0.004, IDI = 0.005, and NRI = 0.150), indicating limited enhancement in predictive performance. However, incorporating menstrual cycle data, particularly the E.A. group, as a predictor yielded more substantial results (Δ AUC = 0.04, IDI = 0.051, and NRI = 0.474). This suggests that in women with PCOS, the inclusion of menstrual cycle data, especially amenorrhea, is more effective in predicting insulin resistance-related health outcomes.Newer prediction indicators showed that among age, BMI, WC, menstrual cycle and menstrual cycle played a much more important role in CVD prediction than other characteristics. These indicators more strongly highlighted the differences between predictors than the results of commonly used odds ratios.Further comparative results are presented in Table 2.

Moreover, we employed this comprehensive evaluation approach to assess the predictive performance of the updated model. In our final prediction model, the baseline model incorporated conventional predictors such as patient age, BMI, WC, and FAI. Menstrual cycle data, stratified into three groups—Eumenorrheic vs. Amenorrheic, Eumenorrheic vs. Oligomenorrheic, and Oligomenorrheic vs. Amenorrheic—were introduced as new predictors, forming an enhanced model. This systematic assessment of novel predictors in risk prediction models aims to assist clinicians in identifying more valuable predictors, thereby improving the precision of risk stratification and patient management.

We observed that the inclusion of a new predictor in the risk prediction model often results in trivial improvement in the AUC, even when the predictor shows a significant effect size with the outcome. Specifically, the OR for the association between menstrual cycle and IR in the Oligo vs. Ameno groups reached 1.708. However, the difference in AUC between the baseline model and the new model was 0.018 (see

Table 1. Estimates of the log odds ratio parameters and their asymptotic standard errors for the risk prediction models for cardiometabolic risk markers in PCOS women.

Menstrual group	Outcome	β (Asy.Sd)				
		Age(years)	BMI(kg/m^2)	WC(cm)	FAI(%)	Z
Eumeno VS Ameno	IR	0.069 (0.062)	0.146 (0.081)	-0.015 (0.027)	0.022 (0.043)	1.215 (0.561)
	IFGIGT	-0.016 (0.061)	0.153 (0.069)	0.030 (0.027)	-0.028 (0.040)	1.303 (0.699)
	Dyslipidemia	-0.109 (0.01)	-0.018 (0.023)	0.033 (0.034)	-0.036 (0.013)	0.535 (0.02)
Eumeno VS Oligo	IR	-0.043 (0.053)	0.297 (0.094)	-0.024 (0.026)	0.098 (0.052)	0.880 (0.564)
	IFGIGT	0.002 (0.058)	0.136 (0.078)	0.023 (0.028)	0.079 (0.044)	0.927 (0.724)
	Dyslipidemia	0.043 (0.048)	-0.190 (0.076)	0.006 (0.023)	-0.034 (0.042)	-0.108 (0.513)
Oligo VS Ameno	IR	0.9 (0.02)	0.9 (0.02)	0.54 (0.02)	3.055 (0.02)	3.055 (0.02)
	IFGIGT	-0.037 (0.046)	0.153 (0.054)	0.030 (0.019)	0.012 (0.033)	0.546 (0.404)
	Dyslipidemia	0.000 (0.041)	-0.090 (0.049)	0.004 (0.018)	-0.012 (0.031)	0.921 (0.379)

Abbreviations: OR, odds ratios; IR, insulin resistance; IFGIGT, impaired fasting glucose and impaired glucose tolerance.

Table 3). Despite this small change in AUC, the values for IDI = 0.022 and continuous NRI = 0.276 indicated significant results (see Table 3).All the predictive performance measures were obtained through 1,000 bootstrap iterations, with the mean values and standard error calculated.

These findings suggest that relying solely on AUC to evaluate the effectiveness of a new predictor may lead to an underestimation of its potential contribution. NRI and IDI provides clinically favorable interpretation. Therefore, it is essential to consider multiple metrics to comprehensively assess the impact of new predictors on model performance. This multifaceted approach is crucial for enhancing the utility of predictive models in clinical diagnostics.

5. SIMULATION STUDY AND ANALYSIS

The primary objective of this study was to evaluate the improvement in model performance following the inclusion of novel risk prediction factors. To achieve this, we employed three key evaluation metrics: Area Under the Curve (AUC), Integrated Discrimination Improvement (IDI), and Continuous Net Reclassification Improvement (NRI). Using real-world data from women diagnosed with polycystic ovary syndrome (PCOS), we generated four conventional predictors. Subsequently, we simulated three different effect sizes—large, medium, and small—across various scenarios to examine the impact of new predictive factors.

We simulated three distinct types of novel predictors: Z_1, a binary variable; Z_2, following a Poisson distribution; and Z_3, a discrete variable with values ranging from 1 to 4. Detailed parameters for these simulations are provided in Table 4.

For each combination of effect size and new predictor, we conducted 1,000 simulation runs to rigorously assess model performance. The mean AUC, IDI, and NRI were calculated for each scenario, along with both empirical and asymptotic variances. This comprehensive analysis aims to evaluate the potential improvements in predictive accuracy when novel risk factors are incorporated into models designed to predict cardiovascular outcomes in PCOS patients.

Furthermore, we aimed to simulate scenarios that reflect the complexity of real-world data to examine how AUC, IDI, and NRI, when used together, can effectively evaluate the value of a novel risk prediction factor and the overall predictive performance of an updated model. By doing so, this study offers insights into the utility of these metrics as a comprehensive evaluation framework for model performance enhancement.

Table 5 presents the results of the simulation analysis. The findings from this simulation can be summarized as follows:

When the effect size is set at 0.6 and the new predictive factor is designated as Z_2, the observed increase in AUC is 0.102, while the Integrated Discrimination Improvement

Table 2. The AUC, IDI, and NRI, and their standard errors, after adding WC, FAI, and categorized menstrual cycle data as novel predictors to the model.

Outcome	Model	AUC_0 (Std.Error)	AUC_1 (Std.Error)	IDI (Std.Error)	$NRI^{>0}$ (Std.Error)
IR	WC	0.736 (0.042)	0.735 (0.042)	0.002 (0.003)	0.073 (0.169)
	FAI	0.736 (0.042)	0.740 (0.041)	0.005 (0.005)	0.150 (0.167)
	E.O.	0.771 (0.050)	0.771 (0.049)	0.006 (0.010)	0.214 (0.181)
	E.A.	0.683 (0.065)	0.723 (0.060)	0.051 (0.028)	0.474 (0.225)
	O.A.	0.752 (0.045)	0.765 (0.045)	0.016 (0.010)	0.257 (0.182)
IFGIGT	WC	0.700 (0.046)	0.724 (0.045)	0.022 (0.013)	0.156 (0.176)
	FAI	0.700 (0.046)	0.710 (0.045)	0.003 (0.006)	0.114 (0.175)
	E.O.	0.713 (0.059)	0.719 (0.060)	0.015 (0.010)	0.280 (0.177)
	E.A.	0.684 (0.066)	0.744 (0.058)	0.055 (0.022)	0.533 (0.182)
	O.A.	0.702 (0.049)	0.719 (0.046)	0.016 (0.012)	0.332 (0.184)
Dyslipidemia	WC	0.625 (0.045)	0.616 (0.046)	0.005 (0.006)	-0.097 (0.162)
	FAI	0.625 (0.045)	0.637 (0.046)	0.003 (0.004)	0.013 (0.161)
	E.O.	0.704 (0.054)	0.702 (0.054)	0.000 (0.000)	0.047 (0.176)
	E.A.	0.628 (0.063)	0.643 (0.062)	0.019 (0.016)	0.292 (0.213)
	O.A.	0.629 (0.049)	0.665 (0.048)	0.045 (0.018)	0.401 (0.171)

Abbreviations: AUC_0, area under the receiver operating characteristic curve for baseline model; AUC_1, area under the receiver operating characteristic curve for new model; IDI, Integrated Discrimination Improvement; $NRI^{>0}$, Net Reclassification Improvement; IR, insulin resistance; IFGIGT, impaired fasting glucose and impaired glucose tolerance; WC, waist circumference; FAI, free androgen index; E.O., Eumeno vs. Oligo; E.A., Eumeno vs. Ameno; O.A., Oligo vs. Ameno.

(IDI) is 0.064 and the Net Reclassification Improvement (NRI) exceeds 0, yielding a value of 0.711. This scenario represents the optimal performance across all tested conditions, with all three metrics indicating that Z_2 exerts a significant effect within the predictive model. The results indicate that all three metrics—AUC, IDI, and NRI—can simultaneously reflect the effectiveness of the new predictor.

In the instance where the predictive factor is identified as Z_3, the increase in AUC is 0.042, with IDI measured at 0.017 and NRI at a value of 0.438. This suggests that when the baseline model already exhibits high performance, as evidenced by a high AUC, the incorporation of a new predictor may not result in a substantial enhancement of the AUC, despite the considerable effect size of the new predictor. If the evaluation is conducted solely based on the increment in AUC, the significance of this factor might appear negligible. However, a comprehensive examination that incorporates the IDI and NRI—specifically IDI = 0.017 and NRI = 0.438—indicates that the new model offers a higher risk score for cases and a lower risk score for non-cases compared to the baseline model. Consequently, this analysis affirms the relevance of this factor in enhancing the predictive capabilities of the risk assessment model.

Lastly, when the effect size of the new predictor is small (e.g., 0.2), the increase in AUC is minimal, and both IDI and NRI values are similarly low. Specifically, when the new predictor is designated as Z_1, the increase in AUC is merely 0.007, representing a trivial enhancement, while IDI equals 0.001 and NRI is 0.173. Collectively, these three metrics indicate that the predictive value of this factor is not significant, and under these conditions, the new model does not demonstrate a meaningful performance improvement.

While AUC is a widely used metric, it may sometimes lack sensitivity in specific scenarios. These results suggest that, although the AUC may not significantly improve with the addition of a new predictor when the baseline model is already strong, IDI and NRI can effectively demonstrate the contribution of the new predictor. Thus, a comprehensive evaluation of a new risk prediction model should consider multiple metrics to accurately assess its performance.

6. DISCUSSION

This study emphasizes the critical need for incorporating advanced evaluation metrics beyond the commonly used AUC when assessing risk prediction models. While AUC remains a valuable and widely accepted measure, its limitations become apparent, particularly when the baseline model exhibits high predictive performance. The introduction of metrics such as IDI and NRI serves to address these limitations by offering a more nuanced evaluation of the incremental value added by new predictors. Drawing on the theoretical framework of U-statistics, we established the theoretical properties of these novel metrics. The combined

Table 3. Differences in AUC and value of IDI, NRI under different subgroups and outcomes

	Eumeno VS Ameno		Eumeno VS Oligo		Oligo VS Ameno	
	AUC_0 (Emp.sd)	AUC_1 (Emp.sd)	AUC_0 (Emp.sd)	AUC_1 (Emp.sd)	AUC_0 (Emp.sd)	AUC_1 (Emp.sd)
IR	0.708 (0.063)	0.761 (0.061)	0.801 (0.050)	0.820 (0.048)	0.766 (0.046)	0.780 (0.045)
IFGIGT	0.742 (0.064)	0.782 (0.056)	0.769 (0.055)	0.788 (0.055)	0.748 (0.046)	0.761 (0.045)
Dyslipidemia	0.702 (0.060)	0.720 (0.059)	0.725 (0.051)	0.733 (0.049)	0.648 (0.051)	0.691 (0.046)
	IDI (Emp.sd)	$NRI^{>0}$ (Emp.sd)	IDI (Emp.sd)	$NRI^{>0}$ (Emp.sd)	IDI (Emp.sd)	$NRI^{>0}$ (Emp.sd)
IR	0.073 (0.060)	0.458 (0.238)	0.032 (0.036)	0.302 (0.222)	0.022 (0.023)	0.261 (0.177)
IFGIGT	0.052 (0.042)	0.422 (0.255)	0.031 (0.031)	0.233 (0.207)	0.020 (0.022)	0.330 (0.195)
Dyslipidemia	0.022 (0.027)	0.276 (0.220)	0.010 (0.015)	0.140 (0.133)	0.049 (0.037)	0.396 (0.174)

Abbreviations: AUC_0, area under the receiver operating characteristic curve for baseline model; AUC_1, area under the receiver operating characteristic curve for new model; IDI, Integrated Discrimination Improvement; $NRI^{>0}$, Net Reclassification Improvement; IR, insulin resistance; IFGIGT, impaired fasting glucose and impaired glucose tolerance.

Table 4. Distributions used for data generation and parameters setting in simulation study

		Conventional predictors			New predictors		
generation of variable	X_1	X_2	X_3	$log(X_4)$	Z_1	Z_2	Z_3
	$\mathcal{N}(77,4.6^2)$	$\mathcal{N}(23,7^2)$	$\mathcal{N}(77,10^2)$	$\mathcal{N}(1.6,0.5^2)$	$\mathcal{B}(1,0.4)$	$\mathcal{P}(2)$	1,2,3,4
	Sample size: 3,000						
	Effect size	Replications	β_0	β_1	β_2	β_3	β_4
Simulation setting	0.6	1,000					
	0.4	1,000	-6.6	-0.06	0.1	0.01	0.02
	0.2	1,000					

use of ΔAUC, IDI, and $NRI^{>0}$ provides complementary insights, demonstrating that, in many instances, reporting all three metrics, in conjunction with traditional performance measures, is essential for a comprehensive assessment of a model's predictive improvements.

Our findings, based on real-world data analysis, further illustrate the practical advantages of employing IDI and NRI in evaluating risk prediction models. In women with polycystic ovary syndrome (PCOS), our analysis revealed a potential association between menstrual cycle characteristics and cardiovascular disease (CVD) risk. We integrated waist circumference (WC), free androgen index (FAI), and categorized menstrual cycle data into the risk prediction model. The comprehensive evaluation system, encompassing AUC, IDI, and NRI, affirmed the predictive value of menstrual cycle data as a low-cost, informative marker. Incorporating menstrual cycle information into the baseline model enhanced its predictive accuracy, thereby confirming its practical utility in cardiometabolic risk assessment.

In our simulation study, we designed more complex scenarios to evaluate the updated risk models and novel predictors using AUC, IDI, and NRI. These metrics provided a more sensitive and detailed assessment of model improvements, particularly in instances where traditional metrics, such as AUC, failed to detect meaningful changes. This further underscores the necessity of employing a multifaceted approach to model evaluation, especially when new predictors are incorporated.

Moreover, our simulation studies revealed that the asymptotic variance of IDI is highly sensitive to both the distribution and prevalence of the predictor variables. When we varied the distribution of predictor Z under the same effect size, changes in the prevalence of the predictor influenced both the calculated asymptotic and empirical variances, leading to potential biases. This phenomenon has been previously documented in the literature (e.g., [10], [21]). Consequently, IDI as an evaluation metric requires adjustments for estimated parameters, and the theoretical properties of

Table 5. Simulation results of predictive performance measures under different outcome variables and novel predictors

Effect size	New predictor	AUC_0 (Emp.sd,Asy.sd)	AUC_1 (Emp.sd,Asy.sd)	IDI (Emp.sd,Asy.sd)	$NRI^{>0}$ (Emp.sd,Asy.sd)
0.6	Z_1	0.720 (0.042,0.044)	0.737 (0.042,0.043)	0.003 (0.003,0.002)	0.166 (0.166,0.174)
	Z_2	0.698 (0.024,0.025)	0.800 (0.022,0.021)	0.064 (0.017,0.011)	0.711 (0.095,0.090)
	Z_3	0.705 (0.023,0.024)	0.747 (0.022,0.022)	0.017 (0.006,0.004)	0.438 (0.084,0.084)
0.4	Z_1	0.721 (0.044,0.046)	0.732 (0.043,0.045)	0.001 (0.002,0.001)	0.222 (0.154,0.184)
	Z_2	0.710 (0.031,0.032)	0.761 (0.029,0.030)	0.018 (0.009,0.006)	0.479 (0.127,0.125)
	Z_3	0.711 (0.030,0.030)	0.731 (0.029,0.030)	0.005 (0.004,0.002)	0.302 (0.116,0.114)
0.2	Z_1	0.723 (0.047,0.048)	0.730 (0.046,0.048)	0.001 (0.001,0.001)	0.173 (0.132,0.192)
	Z_2	0.718 (0.039,0.041)	0.734 (0.038,0.040)	0.003 (0.003,0.002)	0.237 (0.158,0.162)
	Z_3	0.718 (0.039,0.039)	0.726 (0.038,0.039)	0.001 (0.002,0.001)	0.171 (0.134,0.154)

Abbreviations: AUC_0, area under the receiver operating characteristic curve for baseline model; AUC_1, area under the receiver operating characteristic curve for new model; IDI, Integrated Discrimination Improvement; $NRI^{>0}$, Continuous Net Reclassification Improvement.

its asymptotic variance warrant further exploration.

We also propose nonparametric expressions for both the IDI and NRI metrics. In clinical practice, particularly when working with medical datasets such as electronic health records, missing data is a prevalent challenge. Thus, it is imperative to develop risk prediction models that can accommodate and adjust for missing data during the model updating and evaluation processes. Our formulas (Equations 11 and 13) offer a practical framework for calculating AUC, IDI, and NRI in the presence of missing data. We intend to further investigate this area, with a focus on model updates and evaluations in the context of incomplete data.

Future research should extend the application of IDI and NRI across various biomedical contexts to validate their robustness and broaden their integration into clinical practice. Such efforts will enhance the precision and reliability of risk prediction models, thereby improving patient outcomes and advancing the field of biomedical research.

7. ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to all the doctors, nurses, and laboratory staff employed at the Reproductive Medicine Center, Shunde Hospital, for their technical support and valuable suggestions. The research was supported by [Funding Organization].

8. APPENDIX: PROOFS OF THE MAIN RESULTS

8.1 Proof 1:

$\widehat{\text{IDI}} = (\bar{\hat{p}}_{new,events} - \bar{\hat{p}}_{new,nonevents}) - (\bar{\hat{p}}_{old,events} - \bar{\hat{p}}_{new,nonevents})$
To prove the validity of the equation, we need to demonstrate:

$$\text{IS}_{new}\widehat{-\text{IP}}_{new} = \bar{\hat{p}}_{new,events} - \bar{\hat{p}}_{new,nonevents}.$$

Therefore, the following proves:

$$\widehat{\text{IS-IP}} = \bar{\hat{p}}_{events} - \bar{\hat{p}}_{nonevents}$$

We denote $\hat{p}_{i,events}$ as the risk value obtained from the risk prediction model for each case individual, $i = 1, 2, ...N_1$, where$\hat{p}_{j,nonevents}$ as the risk value obtained from the risk prediction model for each non-case individual $j = 1, 2, ...N_0$.

$$\therefore \bar{\hat{p}}_{\text{events}} = \frac{\sum_{i=1}^{N_1} \hat{p}_{i,\text{events}}}{N_1}, \quad \bar{\hat{p}}_{\text{nonevents}} = \frac{\sum_{j=1}^{N_0} \hat{p}_{j,\text{nonevents}}}{N_0}$$

Let's assume that the values of $\hat{p}_{i,events}, i = 1, 2, ...N_1$,are sorted in descending order, so we have:

$$\begin{aligned}
\because \text{IS} &= \int_0^1 \int_\nu^1 f_R(r \mid Y = 1)\, dr\, d\nu \\
&= \int_0^1 P(R > \nu \mid Y = 1)\, d\nu \\
&\approx \int_0^1 \frac{\sum_{i=1}^{N_1} \mathbb{I}(\hat{p}_{i,\text{events}} > \nu)}{N_1}\, d\nu \\
\therefore \widehat{\text{IS}} &= (\hat{p}_{1,\text{events}} - 0) \times 1 + (\hat{p}_{2,\text{events}} - \hat{p}_{1,\text{events}}) \times \frac{N_1 - 1}{N_1} \\
&\quad + \ldots + (\hat{p}_{k+1,\text{events}} - \hat{p}_{k,\text{events}}) \times \frac{N_1 - k}{N_1} \\
&\quad + \ldots + (\hat{p}_{N_1,\text{events}} - \hat{p}_{N_1-1,\text{events}}) \times \frac{1}{N_1} \\
\widehat{\text{IS}} &= \sum_{i=1}^{N_1} \left(\frac{N_1 - (i-1)}{N_1} - \frac{N_1 - i}{N_1} \right) \times \hat{p}_{i,\text{events}} \\
&= \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{p}_{i,\text{events}} = \bar{\hat{p}}_{\text{events}}
\end{aligned}$$

Similarly, we assume that $\hat{p}_{j,nonevents}, j = 1, 2, ...N_0$, are sorted in ascending order.

$$\begin{aligned}
\because \text{IP} &= \int_0^1 \int_\nu^1 f_R(r \mid Y = 0)\, dr\, d\nu \\
&= \int_0^1 P(R > \nu \mid Y = 0)\, d\nu \\
&\approx \int_0^1 \frac{\sum_{j=1}^{N_0} \mathbb{I}(\hat{p}_{j,\text{nonevents}} > \nu)}{N_0}\, d\nu \\
\therefore \widehat{\text{IP}} &= (\hat{p}_{1,\text{nonevents}} - 0) \times 1 + (\hat{p}_{2,\text{nonevents}} - \hat{p}_{1,\text{nonevents}}) \times \frac{N_0 - 1}{N_0} \\
&\quad + \ldots + (\hat{p}_{k+1,\text{nonevents}} - \hat{p}_{k,\text{nonevents}}) \times \frac{N_0 - k}{N_0} \\
&\quad + \ldots + (\hat{p}_{N_0,\text{nonevents}} - \hat{p}_{N_0-1,\text{nonevents}}) \times \frac{1}{N_0} \\
\therefore \widehat{\text{IP}} &= \sum_{j=1}^{N_0} \left(\frac{N_0 - (j-1)}{N_0} - \frac{N_0 - j}{N_0} \right) \times \hat{p}_{j,\text{nonevents}} \\
&= \frac{1}{N_0} \sum_{j=1}^{N_0} \hat{p}_{j,\text{nonevents}} = \bar{\hat{p}}_{\text{nonevents}}
\end{aligned}$$

This completes the proof:

$$\widehat{\text{IS-IP}} = \bar{\hat{p}}_{\text{events}} - \bar{\hat{p}}_{\text{nonevents}}$$

8.2 Proof: Unbiased Estimator of IDI

To prove that $\widehat{\text{IDI}}$ is an unbiased estimator of IDI, we need to show:

$$\mathbb{E}[\widehat{\text{IDI}}] = \text{IDI}$$

where

$$\widehat{\text{IDI}} = (\bar{\hat{p}}_{\text{new,events}} - \bar{\hat{p}}_{\text{new,nonevents}}) - (\bar{\hat{p}}_{\text{old,events}} - \bar{\hat{p}}_{\text{old,nonevents}})$$

and

$$\text{IDI} = (\text{IS}_{new} - \text{IP}_{new}) - (\text{IS}_{old} - \text{IP}_{old})$$

Here:

- IS_{new} is the Integrated Sensitivity of the new model.
- IP_{new} is the Integrated Precision of the new model.
- IS_{old} is the Integrated Sensitivity of the old model.
- IP_{old} is the Integrated Precision of the old model.

Let $\hat{p}_{i,\text{events}}$ denote the risk prediction value for the i-th case individual, $i = 1, 2, \ldots, N_1$, and $\hat{p}_{j,\text{nonevents}}$ denote the risk prediction value for the j-th non-case individual, $j = 1, 2, \ldots, N_0$. The estimators for IS_{new} and IP_{new} are given by:

$$\widehat{\text{IS}}_{new} = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{p}_{i,\text{events}}^{(new)}, \widehat{\text{IP}}_{new} = \frac{1}{N_0} \sum_{j=1}^{N_0} \hat{p}_{j,\text{nonevents}}^{(new)}.$$

Similarly, the estimators for IS_{old} and IP_{old} are:

$$\widehat{\text{IS}}_{old} = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{p}_{i,\text{events}}^{(old)}, \widehat{\text{IP}}_{old} = \frac{1}{N_0} \sum_{j=1}^{N_0} \hat{p}_{j,\text{nonevents}}^{(old)}.$$

We compute:

$$\widehat{\text{IDI}} = (\bar{\hat{p}}_{\text{new,events}} - \bar{\hat{p}}_{\text{new,nonevents}}) - (\bar{\hat{p}}_{\text{old,events}} - \bar{\hat{p}}_{\text{old,nonevents}}).$$

where:

$$\begin{aligned}
\bar{\hat{p}}_{\text{new,events}} &= \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{p}_{i,\text{events}}^{(new)}, \\
\bar{\hat{p}}_{\text{new,nonevents}} &= \frac{1}{N_0} \sum_{j=1}^{N_0} \hat{p}_{j,\text{nonevents}}^{(new)}. \\
\bar{\hat{p}}_{\text{old,events}} &= \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{p}_{i,\text{events}}^{(old)}, \\
\bar{\hat{p}}_{\text{old,nonevents}} &= \frac{1}{N_0} \sum_{j=1}^{N_0} \hat{p}_{j,\text{nonevents}}^{(old)}.
\end{aligned}$$

Substituting these values into $\widehat{\text{IDI}}$:

$$\widehat{\text{IDI}} = \left(\frac{1}{N_1} \sum_{i=1}^{N_1} \hat{p}_{i,\text{events}}^{(new)} - \frac{1}{N_0} \sum_{j=1}^{N_0} \hat{p}_{j,\text{nonevents}}^{(new)} \right)$$

$$- \left(\frac{1}{N_1} \sum_{i=1}^{N_1} \hat{p}_{i,\text{events}}^{(old)} - \frac{1}{N_0} \sum_{j=1}^{N_0} \hat{p}_{j,\text{nonevents}}^{(old)} \right)$$

Rewriting this:

$$\widehat{\text{IDI}} = \widehat{\text{IS}}_{\text{new}} - \widehat{\text{IS}}_{\text{old}} - \left(\widehat{\text{IP}}_{\text{new}} - \widehat{\text{IP}}_{\text{old}} \right)$$

Since:

$$\text{IS}_{\text{new}} = \mathbb{E}[\bar{\hat{p}}_{\text{new,events}}], \quad \text{IP}_{\text{new}} = \mathbb{E}[\bar{\hat{p}}_{\text{new,nonevents}}]$$

$$\text{IS}_{\text{old}} = \mathbb{E}[\bar{\hat{p}}_{\text{old,events}}], \quad \text{IP}_{\text{old}} = \mathbb{E}[\bar{\hat{p}}_{\text{old,nonevents}}]$$

We have:

$$\mathbb{E}[\widehat{\text{IDI}}] = (\text{IS}_{\text{new}} - \text{IP}_{\text{new}}) - (\text{IS}_{\text{old}} - \text{IP}_{\text{old}}) = \text{IDI}$$

Thus, $\widehat{\text{IDI}}$ is an unbiased estimator of IDI.

8.3 Proof demonstrates that the IDI and NRI follow an asymptotic normal distribution.

$$\sqrt{\mathbf{N}}(\widehat{NRI}^{>0} - \mathbb{E}[NRI^{>0}]) \xrightarrow{D} N\left(0, N\sigma^2_{NRI^{>0}}\right), \tag{21}$$

$$\sqrt{\mathbf{N}}(\widehat{IDI} - \mathbb{E}[IDI]) \xrightarrow{D} N\left(0, N\sigma^2_{IDI}\right). \tag{22}$$

Theorem [Lee page 140]
Given a functional θ, let U_{n_0,n_1} denote a generalized U-statistic based on two samples, which serves as an unbiased estimator of θ. Define $p = \lim_{N\to\infty} \frac{n_1}{N}$, where $0 < p < 1$. Then, we have the following result:

$$\sqrt{N}\left(U_{n_1,n_0} - \theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, p^{-1}\delta^2_{1,0} + (1-p)^{-1}\delta^2_{0,1}\right)$$

as $N = n_1 + n_0 \to \infty$.

$$\delta^2_{1,0} = Var(h^{1,0}(\mathbf{w}_i^1)),$$
$$\delta^2_{0,1} = Var(h^{0,1}(\mathbf{w}_j^0))$$

where $h^{(\cdot,\cdot)}(.)$ are real-valued kernels. Formulas to calculate the kernels $h^{(\cdot,\cdot)}(.)$ are given by Lee([22]):

$$\text{h}^{(0,0)} = \iint \psi(\mathbf{w}^1;\mathbf{w}^0)\,\text{dK}(\mathbf{w}^1)\,\text{dG}(\mathbf{w}^0)$$

$$\text{h}^{(1,0)}(\mathbf{w}_i^1) = \int \psi(\mathbf{w}_i^1;\mathbf{w}^0)\,\text{dG}(\mathbf{w}^0)$$
$$- \iint \psi(\mathbf{w}^1;\mathbf{w}^0)\,\text{dK}(\mathbf{w}^1)\,\text{dG}(\mathbf{w}^0)$$

$$\text{h}^{(0,1)}\left(\mathbf{w}_j^0\right) = \int \psi\left(\mathbf{w}^1;\mathbf{w}_j^0\right)\,\text{dK}(\mathbf{w}^1)$$
$$- \iint \psi(\mathbf{w}^1;\mathbf{w}^0)\,\text{dK}(\mathbf{w}^1)\,\text{dG}(\mathbf{w}^0)$$

$$\text{h}^{(1,1)}\left(\mathbf{w}_i^1;\mathbf{w}_j^0\right) = \psi\left(\mathbf{w}_i^1;\mathbf{w}_j^0\right) - \int \psi(\mathbf{w}_i^1;\mathbf{w}^0)\,\text{dG}(\mathbf{w}^0)$$
$$- \int \psi\left(\mathbf{w}^1;\mathbf{w}_j^0\right)\,\text{dK}(\mathbf{w}^1) \quad + \iint \psi(\mathbf{w}^1;\mathbf{w}^0)\,\text{dK}(\mathbf{w}^1)\,\text{dG}(\mathbf{w}^0)$$

where $K(.)$ and $G(.)$ are cumulative distribution functions of random vectors $\mathbf{W}^1$ and $\mathbf{W}^0$, that $\mathbf{W}^1 \sim K(\cdot)$ and $\mathbf{W}^0 \sim G(\cdot)$.

8.4 Derivation of $\hat{\sigma}^2_{IDI}$

$$h^{(1,0)}(\mathbf{w}_i^1) = \int \left([p_1(\mathbf{w}_i^1) - p_0(\mathbf{w}_i^1)] - [p_1(\mathbf{w}^0) - p_0(\mathbf{w}^0)]\right) dG(\mathbf{w}^0)$$
$$- h^{(0,0)}$$

$$h^{(0,1)}(\mathbf{w}_j^0) = \int \left([p_1(\mathbf{w}^1) - p_0(\mathbf{w}^1)] - [p_1(\mathbf{w}_j^0) - p_0(\mathbf{w}_j^0)]\right) dK(\mathbf{w}^1)$$
$$- h^{(0,0)}$$

$$\delta^2_{1,0} = Var(h^{(1,0)}(\mathbf{w}_i^1)) = Var[p_1(\mathbf{w}_i^1) - p_0(\mathbf{w}_i^1)],$$
$$\delta^2_{0,1} = Var(h^{(0,1)}(\mathbf{w}_j^0)) = Var[p_1(\mathbf{w}_j^0) - p_0(\mathbf{w}_j^0)].$$

Form the theorem above we can derive the U_{n_1,n_0} has the variance as follow:

$$\hat{\sigma}^2_{U_{n_1,n_0}} = \frac{1}{N}\left(\frac{1}{p}\delta^2_{1,0} + \frac{1}{1-p}\delta^2_{0,1}\right)$$

$$\hat{\sigma}^2_{IDI} = \frac{Var[p_1(\mathbf{W}_i^1) - p_0(\mathbf{W}_i^1)]}{n_1} + \frac{Var[p_1(\mathbf{W}_j^0) - p_0(\mathbf{W}_j^0)]}{n_0} \tag{23}$$

8.5 Derivation of $\hat{\sigma}^2_{NRI}$

$$h^{(1,0)}(\mathbf{w}_i^1) = 2\int \left(\mathbb{I}[p_1(\mathbf{w}_i^1) - p_0(\mathbf{w}_i^1)] - \mathbb{I}[p_1(\mathbf{w}^0) - p_0(\mathbf{w}^0)]\right) dG(\mathbf{w}^0) -$$

$$h^{(0,1)}(\mathbf{w}_j^0) = 2\int \left(\mathbb{I}[p_1(\mathbf{w}^1) - p_0(\mathbf{w}^1)] - \mathbb{I}[p_1(\mathbf{w}_j^0) - p_0(\mathbf{w}_j^0)]\right) dK(\mathbf{w}^1) -$$

$$\delta^2_{1,0} = Var(h^{(1,0)}(\mathbf{w}_i^1)) = 4Var(\mathbb{I}[p_1(\mathbf{w}_i^1) - p_0(\mathbf{w}_i^1)]),$$
$$\delta^2_{0,1} = Var(h^{(0,1)}(\mathbf{w}_j^0)) = 4Var(\mathbb{I}[p_1(\mathbf{w}_j^0) - p_0(\mathbf{w}_j^0)]).$$

Form the analysis above we can derive the variance of $NRI^{>0}$ as follow:

$$\hat{\sigma}^2_{NRI^{>0}} = 4\Big(\frac{Var[\mathbb{I}[p_1(\mathbf{W}_i^1) - p_0(\mathbf{W}_i^1)]]}{n_1} + \frac{Var[\mathbb{I}[p_1(\mathbf{W}_j^0) - p_0(\mathbf{W}_j^0)]]}{n_0}\Big).$$

8.6 Odds ratio table

Table 6 presents the Odds ratio for association of menstrual cyclicity and cardiometabolic risk markers in PCOS women

REFERENCES

[1] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology* **12** 387–415.

[2] Barlow, W. E., White, E., Ballard-Barbash, R., Vacek, P. M., Titus-Ernstoff, L., Carney, P. A., Tice, J. A., Buist, D. S., Geller, B. M., Rosenberg, R. et al. (2006). Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute* **98** 1204–1214.

[3] Chambless, L. E., Cummiskey, C. P. and Cui, G. (2011). Several methods to assess improvement in risk prediction models: extension to survival analysis. *Statistics in medicine* **30** 22–38.

[4] Chen, J., Pee, D., Ayyagari, R., Graubard, B., Schairer, C., Byrne, C., Benichou, J. and Gail, M. H. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of the National Cancer Institute* **98** 1215–1226.

[5] Chen, W., Samuelson, F. W., Gallas, B. D., Kang, L., Sahiner, B. and Petrick, N. (2013). On the assessment of the added value of new predictive biomarkers. *BMC medical research methodology* **13** 1–9.

[6] Collins, G. S., Mallett, S., Omar, O. and Yu, L.-M. (2011). Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine* **9** 1–14.

[7] Cook, N. R., Buring, J. E. and Ridker, P. M. (2006). The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of internal medicine* **145** 21–29.

[8] D'Agostino, R., Griffith, J., Schmid, C. and Terrin, N. (1997). Measures for evaluating model performance. In *Proceedings-American Statistical Association Biometrics Section* 253–258. UNKNOWN.

[9] DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845.

[10] Demler, O. V., Pencina, M. J., Cook, N. R. and D'Agostino Sr, R. B. (2017). Asymptotic distribution of AUC, NRIs, and IDI based on theory of U-statistics. *Statistics in medicine* **36** 3334–3360.

[11] D'Agostino Sr, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M. and Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117** 743–753.

[12] ET. Wang, T. J., Gona, P., Larson, M. G., Tofler, G. H., Levy, D., Newton-Cheh, C., Jacques, P. F., Rifai, N., Selhub, J., Robins, S. J. et al. (2006). Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine* **355** 2631–2639.

[13] Ezeh, U., Pisarska, M. and Azziz, R. (2022). Association of severity of menstrual dysfunction with hyperinsulinemia and dysglycemia in polycystic ovary syndrome. *Human Reproduction* **37** 553–564.

[14] Ezeh, U., Ezeh, C., Pisarska, M. D. and Azziz, R. (2021). Menstrual dysfunction in polycystic ovary syndrome: association with dynamic state insulin resistance rather than hyperandrogenism. *Fertility and sterility* **115** 1557–1568.

[15] Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute* **81** 1879–1886.

[16] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

[17] Hlatky, M. A., Greenland, P., Arnett, D. K., Ballantyne, C. M., Criqui, M. H., Elkind, M. S., Go, A. S., Harrell Jr, F. E., Hong, Y., Howard, B. V. et al. (2009). Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation* **119** 2408–2416.

[18] Hoeger, K. M., Dokras, A. and Piltonen, T. (2021). Update on PCOS: consequences, challenges, and guiding treatment. *The Journal of Clinical Endocrinology & Metabolism* **106** e1071–e1083.

[19] Janes, H., Pepe, M. S. and Gu, W. (2008). Assessing the value of risk predictions by using risk stratification tables. *Annals of internal medicine* **149** 751–760.

[20] Kannel, W. B., McGee, D. and Gordon, T. (1976). A general cardiovascular risk profile: the Framingham Study. *The American journal of cardiology* **38** 46–51.

[21] Kerr, K. F., McClelland, R. L., Brown, E. R. and Lumley, T. (2011). Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American journal of epidemiology* **174** 364–374.

[22] Lee, A. J. (2019). *U-statistics: Theory and Practice*. Routledge.

[23] Levy, D., Garrison, R. J., Savage, D. D., Kannel, W. B. and Castelli, W. P. (1990). Prognostic implications of echocardiographically determined left ventricular mass in the Framingham Heart Study. *New England Journal of Medicine* **322** 1561–1566.

[24] Li, X., Yang, J., Blockeel, C., Lin, M., Tian, X., Wu, H., Cao, Y., Deng, L., Zhou, X., Xie, J. et al. (2024). Association of severity of menstrual dysfunction with cardiometabolic risk markers among women with polycystic ovary syndrome. *Acta Obstetricia et Gynecologica Scandinavica*.

[25] Lusted, L. B. (1971). Decision-making studies in patient management. *New England Journal of Medicine* **284** 416–424.

[26] March, W. A., Moore, V. M., Willson, K. J., Phillips, D. I., Norman, R. J. and Davies, M. J. (2010a). The prevalence of polycystic ovary syndrome in a community sample assessed under contrasting diagnostic criteria. *Human reproduction* **25** 544–551.

[27] March, W. A., Moore, V. M., Willson, K. J., Phillips, D. I., Norman, R. J. and Davies, M. J. (2010b). The prevalence of polycystic ovary syndrome in a community sample assessed under contrasting diagnostic criteria. *Human reproduction* **25** 544–551.

[28] Müller, B., Morgenthaler, N., Stolz, D., Schuetz, P., Müller, C., Bingisser, R., Bergmann, A., Tamm, M. and Christ-Crain, M. (2007). Circulating levels of copeptin, a novel biomarker, in lower respiratory tract infections. *European journal of clinical investigation* **37** 145–152.

[29] Pencina, M. J., D'Agostino Sr, R. B. and Steyerberg, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine* **30** 11–21.

[30] Pencina, M. J., D'Agostino Sr, R. B. and Demler, O. V. (2012). Novel metrics for evaluating improvement in discrim-

Table 6. Odds ratio for association of menstrual cyclicity and cardiometabolic risk markers in PCOS women.

	Eumeno VS Ameno		Eumeno VS Oligo		Oligo VS Ameno	
	OR	95% CI	OR	95% CI	OR	95% CI
No other covariates were added in model						
IR	3.055	(1.120-8.541)	1.792	(0.696-4.667)	1.704	(0.810-3.685)
IFGIGT	4.848	(1.452-22.249)	2.469	(0.742-11.277)	1.964	(0.945-4.130)
Dyslipidemia	2.018	(0.756-5.473)	0.878	(0.342-2.267)	2.298	(1.138-4.740)
Adjusted for age, BMI, and waist circumference in model						
IR	3.189	(1.106-9.641)	1.797	(0.640-5.165)	1.751	(0.771-4.103)
IFGIGT	3.819	(1.085-18.046)	2.011	(0.576-9.409)	1.737	(0.786-3.861)
Dyslipidemia	1.834	(0.652-5.207)	0.984	(0.368-2.654)	2.495	(1.202-5.328)
Adjusted for age, BMI, waist circumference, and FAI in model						
IR	3.371	(1.144-10.546)	2.411	(0.811-7.553)	1.737	(0.764-4.077)
IFGIGT	3.681	(1.035-17.479)	2.526	(0.687-12.706)	1.727	(0.781-3.842)
Dyslipidemia	1.708	(0.595-4.909)	0.898	(0.326-2.479)	2.513	(1.209-5.377)

Abbreviations: OR, odds ratios; IR, insulin resistance; IFGIGT, impaired fasting glucose and impaired glucose tolerance.

ination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Statistics in medicine* **31** 101–113.

[31] Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B. and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine* **27** 157–172.

[32] Pencina, M. J., D'agostino, R. B., Pencina, K. M., Janssens, A. C. J. and Greenland, P. (2012). Interpreting incremental value of markers added to risk prediction models. *American journal of epidemiology* **176** 473–481.

[33] Pepe, M. S., Janes, H., Longton, G., Leisenring, W. and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology* **159** 882–890.

[34] Ridker, P. M., Buring, J. E., Rifai, N. and Cook, N. R. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *Jama* **297** 611–619.

[35] Schnabel, R. B., Sullivan, L. M., Levy, D., Pencina, M. J., Massaro, J. M., D'Agostino, R. B., Newton-Cheh, C., Yamamoto, J. F., Magnani, J. W., Tadros, T. M. et al. (2009). Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *The Lancet* **373** 739–745.

[36] Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician* **63** 366–372.

[37] Van der Ham, K., Louwers, Y. V. and Laven, J. S. (2022). Cardiometabolic biomarkers in women with polycystic ovary syndrome. *Fertility and sterility* **117** 887–896.

[38] Wang, Cirillo, P. M., Vittinghoff, E., Bibbins-Domingo, K., Cohn, B. A. and Cedars, M. I. (2011). Menstrual irregularity and cardiovascular mortality. *The Journal of Clinical Endocrinology & Metabolism* **96** E114–E118.

[39] Ware, J. H. et al. (2006). The limitations of risk factors as prognostic tools. *N engl J med* **355** 2615–2617.

[40] Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H. and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* **97** 1837–1847.

[41] Xanthakis, V., Sullivan, L. M., Vasan, R. S., Benjamin, E. J., Massaro, J. M., D'Agostino Sr, R. B. and Pencina, M. J. (2014). Assessing the incremental predictive performance of novel biomarkers over standard predictors. *Statistics in medicine* **33** 2577–2584.

[42] Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance* **30** 132–156.

[43] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* **3** 32–35.

Hong Xu
School of Science
Minzu University of China
Beijing,100086
China
E-mail address: hongxu@muc.edu.cn

Ya Qi Cao
School of Science
Minzu University of China
Beijing,100086
China
E-mail address: yaqicao@muc.edu.cn