國立臺灣大學公共衛生學院流行病學與預防醫學研究所

博士論文

Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University

Doctoral Dissertation

預測模型效能之評估：預測曲線及其幾何摘要

Evaluating Prediction Model Performance: The Predictiveness Curve and Its Geometric Summaries

邱偉珉

Wei-Min Chiu

指導教授：李文宗 博士

Advisor: Wen-Chung Lee, M.D., Ph.D.

中華民國 114 年 8 月

August 2025

doi:10.6342/NTU202504305

Chinese Abstract

預測模型是臨床醫學與公共衛生中不可或缺的重要工具，有助於疾病風險的估計與決策的支持。傳統的評估方法，例如接收者操作特徵曲線及其下面積，主要評估模型的區分能力，但在評估族群層級的風險分層方面提供的資訊有限。最初為了評估生物標記而提出的預測曲線，能以視覺方式呈現預測風險在整體族群中的分布情形，但在多變量預測模型的應用中仍較少被探討。

本研究將預測曲線的方法延伸至多變量風險預測模型的評估中，系統性地探討其幾何特性，並導出三個互補的效能指標：Pietra 指標、Gini 指標，以及標準化的 Brier 分數。這三個指標分別量化模型在解析中間風險個案、區分預測風險層級，以及提高預測確定性方面的能力。為了確保預測風險的校正與不偏性，本研究提出了一個三步驟流程，包括交叉驗證、保序迴歸校正，以及自助法平均。

透過說明性範例及一項涉及台灣 23,839 名肺癌患者的真實應用案例，研究展示並驗證了此方法的可行性。Pietra 指標、Gini 指標，以及標準化的 Brier 分數分別捕捉到風險分層表現的不同面向。即使某些模型在其中一項指標上表現相同，其他指標仍可能呈現顯著差異，顯示這些指標具有互補性。在肺癌個案研究中，最終調整後的預測模型能有效區分死亡風險，將 25.1% 的患者歸類為極低風險（<10% 死亡率），50.1% 為高風險（>75% 死亡率），僅有 5.1% 的患者落在接近平均風險的「灰色區域」。該調整後模型的 Pietra 指標為 0.6719，Gini 指標為 0.7850，標準化的 Brier 分數為 0.5186。不同細胞類型的肺癌 (肺腺癌、鱗狀細胞癌、小細胞癌和大細胞癌) 在預測表現上差異顯著，反映出不同細胞類型的肺癌在風險分層能力上的差異。

預測曲線及其幾何效能指標提供一種強大、透明且以族群為導向的框架，用以超越傳統指標來評估多變量風險預測模型的表現。這些方法能清楚地展現模型如何進行風險分層及其影響到的族群比例，進而提升模型可解釋性，協助臨床醫師與研究人員更有效地優化與應用預測模型於臨床與公共衛生實務中。

關鍵字： 預測模型評估; 預測曲線; 風險分層; Gini 指標; Pietra 指標; 標準化的 Brier 分數

doi:10.6342/NTU202504305

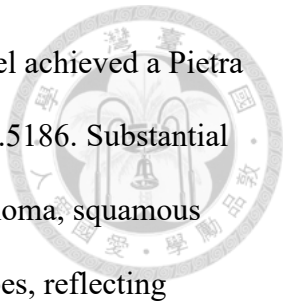
English Abstract

Prediction models are essential tools in clinical medicine and public health, facilitating disease risk estimation and supporting decision-making. Traditional evaluation methods such as the receiver operating characteristic (ROC) curve and its area under the curve (AUC) primarily assess discrimination but provide limited insight into population-level risk stratification. The predictiveness curve, originally proposed for biomarker evaluation, visually illustrates how predicted risks distribute across a population but has been underexplored in the context of multivariable prediction models.

This study extends the predictiveness curve methodology to evaluate multivariable risk prediction models, systematically exploring its geometric properties and deriving three complementary performance indices: the Pietra index, Gini index, and scaled Brier score. These indices respectively quantify a model's ability to resolve intermediate-risk cases, achieve separation among predicted risks, and enhance prediction certainty. A three-step procedure—cross-validation, isotonic regression calibration, and bootstrap averaging—was proposed to ensure calibration and unbiasedness of predicted risks. Illustrative examples and a real-world application involving 23,839 lung cancer patients from Taiwan were used to demonstrate and validate the methodology.

The Pietra index, Gini index, and scaled Brier score captured distinct dimensions of risk stratification performance. Models with identical values of one index could differ markedly in the other indices, underscoring their complementary nature. In the lung cancer case study, the final adjusted prediction model effectively stratified patients across the fatality risk spectrum, identifying 25.1% of patients as very low-risk (<10% fatality) and 50.1% as high-risk (>75% fatality), with only 5.1% of patients falling

doi:10.6342/NTU202504305



within a “gray zone” near the average fatality risk. The adjusted model achieved a Pietra index of 0.6719, a Gini index of 0.7850, and a scaled Brier score of 0.5186. Substantial variation in predictive performance was observed among adenocarcinoma, squamous cell carcinoma, small cell carcinoma, and large cell carcinoma subtypes, reflecting differential risk stratification capabilities by cell type.

The predictiveness curve and its geometric summaries—the Pietra index, Gini index, and scaled Brier score—provide a powerful, transparent, and population-oriented framework for evaluating the performance of multivariable risk prediction models beyond traditional metrics. By clearly illustrating how a model stratifies risk and for which proportion of the population, these methods enhance interpretability, supporting clinicians and researchers in refining and applying predictive models effectively in clinical and public health practice.

Keywords: prediction model evaluation; predictiveness curve; risk stratification; Gini index; Pietra index; scaled Brier score.

doi:10.6342/NTU202504305

Table of Contents

doi:10.6342/NTU202504305

doi:10.6342/NTU202504305

List of Figures

doi:10.6342/NTU202504305

List of Tables

doi:10.6342/NTU202504305

Chapter 1 Introduction

Prediction models are essential decision-support tools widely used in epidemiology, clinical medicine, and public health to estimate disease risk, inform clinical decisions, and guide population-level interventions.[1,2] By integrating diverse variables—including demographic characteristics, socioeconomic factors, lifestyle behaviors, environmental exposures, biological markers, medical histories, and clinical data—these models effectively quantify an individual's likelihood of developing specific diseases. Such risk information helps individuals better understand their health status, enables healthcare professionals to develop appropriate management strategies, and informs public health initiatives aimed at disease prevention and control.

To ensure prediction models are useful and accurate, it is essential to evaluate their calibration and discrimination performance.[3] Calibration assesses how closely predicted risks align with actual observed outcomes, while discrimination measures the model's ability to distinguish between individuals who do and do not develop the disease. The area under the receiver operating characteristic (ROC) curve (AUC), also known as the c statistic, is widely used to evaluate and compare the discrimination ability of diagnostic tests and prediction models.[4] AUC values range from 0.5 (no better than random guessing) to 1.0 (perfect discrimination). However, when multiple models share the same AUC, relying solely on this measure makes it challenging to differentiate their relative predictive performances or to clearly identify why one model may outperform others.

Huang et al. introduced the predictiveness curve as a graphical tool to illustrate how predicted risks are distributed across population percentiles.[5] Unlike the ROC curve, which depicts sensitivity and specificity at various thresholds without revealing

doi:10.6342/NTU202504305

the actual underlying distribution of predicted risks, the predictiveness curve offers deeper insights into the predictive capacity, risk stratification, and discrimination capability of a marker or diagnostic test, particularly regarding its performance in identifying distinct high- and low-risk subgroups. Another key advantage of the predictiveness curve is its use of population percentiles as a standardized scale, which facilitates consistent and interpretable comparisons across different risk distributions and populations. However, despite these strengths, the original application of the predictiveness curve primarily focused on evaluating biomarkers or diagnostic tests; its utility in the context of assessing and comparing the performance of prediction models has not been extensively explored. Expanding the use of the predictiveness curve to prediction models could provide additional insights into model performance, beyond those available through traditional metrics such as the ROC curve.

This study aims to extend the application of the predictiveness curve from its traditional use with single biomarkers and diagnostic tests to the evaluation of multivariable prediction models. Specifically, we systematically explore the geometric properties of the predictiveness curve and derive three intuitive performance indices—the Pietra index, the Gini index, and the scaled Brier score—that quantify complementary aspects of model performance. Through analytical derivation, illustrative examples, and empirical application to a large cohort of lung cancer patients, we demonstrate how these indices provide insights beyond conventional measures such as the area under the ROC curve. By focusing on risk stratification, gray-zone resolution, and certainty of prediction, this framework offers a more transparent and population-anchored evaluation of predictive effectiveness for public health researchers and clinical decision-makers.

doi:10.6342/NTU202504305

Chapter 2 Predictiveness Curve: Construction, Geometry, and Performance Indices

This paper proposes constructing the predictiveness curve for evaluating prediction models as follows: the curve is generated by plotting the predicted risk r, derived from a prediction model, on the y-axis against the cumulative proportion of the population with predicted risk less than or equal to r on the x-axis. Since both axes range from 0 to 1, the curve lies entirely within the unit square. It is monotonically non-decreasing, as higher predicted risk values correspond to greater—or at least equal—proportions of subjects with predicted risk less than or equal to that value. Figure 1 illustrates example predictiveness curves for three models applied to the same population with a disease prevalence of 0.2: an informative but imperfect model (A), a null model (B), and a perfect model (C).

The area under the predictiveness curve represents the mean predicted risk when the prediction model is applied to the population. For a well-calibrated and unbiased model—one for which approximately $100 \times r$ out of 100 individuals with a predicted risk of r are actually diseased—this area corresponds to the disease prevalence in the population, denoted by π (Appendix 1).

A horizontal line at $r = \pi$ (dotted lines in Figure 1) divides the predictiveness curve into two segments: one below and one above the line. Together with this horizontal line, the curve forms two enclosed regions—one below the line (the below-the-line, or BL, region) and one above (the above-the-line, or AL, region) (orange and red shaded regions in Figure 1). For a well-calibrated and unbiased prediction model, these two regions have equal area, denoted by A (Appendix 2).

doi:10.6342/NTU202504305

A null prediction model (Figure 1B) assigns the same risk, π, to every subject in the population. Its predictiveness curve is a horizontal line at $r = \pi$, with no BL or AL regions, meaning $A = 0$. In contrast, a perfect prediction model (Figure 1C) assigns a risk of 1 to all subjects who eventually become diseased and a risk of 0 to all subjects who remain non-diseased. Its predictiveness curve remains at $r = 0$ for 0 to $(1 - \pi)$, then jumps to $r = 1$ and stays constant from $(1 - \pi)$ to 1. Its BL region is a rectangle with a width of $(1 - \pi)$ and a height of π, while its AL region is a rectangle with a width of π and a height of $(1 - \pi)$. Thus, $A = \pi \times (1 - \pi)$.

We now demonstrate how three commonly used performance indices for prediction models—the Pietra index, the Gini index (both derived from the Lorenz curve), and the scaled Brier score—are mathematically linked to the geometric properties of the predictiveness curve. These indices are standardized: they equal 0 for a null prediction model, 1 for a perfect model, and lie between 0 and 1 for informative but imperfect models.[6,7] Notably, for well-calibrated and unbiased models, the Lorenz-based Gini and Pietra indices correspond to their ROC-based counterparts, with $\text{Gini} = 2 \times \text{AUC} - 1$, and Pietra equal to the maximum vertical distance (MVD) from the ROC curve to the diagonal line.[6,7]

Appendix 3 shows that the Pietra index equals the area A of the BL (or AL) region of the predictiveness curve, normalized by the maximum possible area from a perfect model:

$$\text{Pietra} = A/A_{\text{perfect model}} = A/[\pi \times (1 - \pi)]. \quad (1)$$

Let $(x_{\text{BL}}, y_{\text{BL}})$ and $(x_{\text{AL}}, y_{\text{AL}})$ represent the coordinates of the centers of gravity for the BL and AL region, respectively (orange and red dots in Figure 1). Appendix 4 shows that the Gini index is related to the horizontal separation between these centers of

doi:10.6342/NTU202504305

gravity, given by:

$$\text{Gini} = 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra}. \quad (2)$$

Appendix 5 demonstrates that the scaled Brier score is connected to the vertical separation between these centers of gravity, expressed as:

$$\text{scaled Brier} = 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra}. \quad (3)$$

Additionally, Appendix 6 establishes the following relationship among the indices: $0 \leq \text{scaled Brier} \leq \text{Pietra} \leq \text{Gini} \leq 1$.

We now return to the example predictiveness curves in Figure 1. For the informative but imperfect model (Figure 1A), the BL and AL regions have equal areas of 0.0519, with centers of gravity at (0.2370, 0.1503) and (0.8536, 0.2794), as detailed in Appendix 7. From these values and Equations (1), (2), and (3), we calculate $\text{Pietra} = 0.0519/[0.2 \times (1 - 0.2)] = 0.3244$, $\text{Gini} = 2 \times (0.8536 - 0.2370) \times \text{Pietra} = 0.4001$, and $\text{scaled Brier} = 2 \times (0.2794 - 0.1503) \times \text{Pietra} = 0.0838$. Note that the values of these performance indices fall within the expected range and maintain the correct order: $0 \leq 0.0838 \leq 0.3244 \leq 0.4001 \leq 1$.

For the null model (Figure 1B), no BL or AL regions are formed (i.e., A = 0), resulting in a Pietra index of zero [Equation (1)]. Consequently, the Gini index and the scaled Brier score also equal zero [Equations (2) and (3)]. For the perfect model (Figure 1C), the BL and AL regions form simple rectangles, allowing straightforward calculation of their areas and centers of gravity: $A = 0.16$, $(x_{\text{BL}}, y_{\text{BL}}) = (0.4, 0.1)$, and $(x_{\text{AL}}, y_{\text{AL}}) = (0.9, 0.6)$. From Equations (1), (2), and (3), the three indices achieve the value of 1 as expected.

doi:10.6342/NTU202504305

Chapter 3 Performance Indices and Their Role in Risk Stratification

Assuming the prediction models are well-calibrated and unbiased, we next show how the three geometric summary measures of the predictiveness curve—the Pietra index, Gini index, and scaled Brier score—each capture a distinct dimension of a model's risk stratification performance.

3.1 Gray-Zone Resolution: Pietra Index

The upper row of Figure 2 presents the predictiveness curves of three additional prediction models, I, II, and III, applied to the same population. These models share the same Gini index of 0.4000 and scaled Brier score of 0.1176, but their Pietra indices decrease in order: 0.4000 for model I, 0.3191 for model II, and 0.2381 for model III (Appendix 8). Model I classifies all individuals into either the high-risk or low-risk groups, leaving no individuals in the "gray zone" (average risk), whereas model II assigns a certain proportion (25.35%) and model III an even larger proportion (67.99%) of the population to the gray zone. This suggests that among prediction models with the same Gini index and scaled Brier score, a higher Pietra index reflects a stronger ability to resolve the gray zone, effectively reducing the number of individuals assigned near the average-risk level.

3.2 Horizontal Risk Separation: Gini Index

The middle row of Figure 2 presents the predictiveness curves of three additional prediction models, IV, V, and VI, applied to the same population. These models share the same Pietra index of 0.3244 and the same scaled Brier score of 0.0840, but their Gini indices differ: 0.3889 for model IV and 0.3244 for models V and VI (Appendix 8). Model IV classifies the population into three distinct groups: equal proportions

doi:10.6342/NTU202504305

(40.06%) in the high-risk and low-risk groups, with the remaining 19.88% in the average-risk group. In contrast, models V and VI classify individuals into only two groups—high-risk and low-risk—without leaving any in the average-risk category. This indicates that among prediction models with the same Pietra index and scaled Brier score, a higher Gini index reflects a stronger ability to separate predicted probabilities between individuals, while a lower Gini index suggests a tendency to assign similar probabilities to many individuals.

3.3 Vertical Certainty of Prediction: Scaled Brier Score

The lower row of Figure 2 illustrates the predictiveness curves of three prediction models, VII, VIII, and IX, applied to the same population with a disease prevalence of 0.2. Each model classifies individuals into high-risk ($\text{risk} > 0.2$), low-risk ($\text{risk} < 0.2$), or average-risk ($\text{risk} = 0.2$). As shown in Appendix 8, all three models have the same Gini index of 0.4000 and the same Pietra index of 0.3244, but their scaled Brier scores differ: 0.2835 for model VII, 0.0981 for model VIII, and 0.0878 for model IX. From Figure 2, model VII predicts high-risk individuals as diseased with certainty ($\text{risk} = 1$), model VIII predicts low-risk individuals as non-diseased with certainty ($\text{risk} = 0$), while model IX achieves neither—raising the risk of high-risk individuals to 0.3354 (less than 1) and lowering the risk of low-risk individuals to 0.0646 (greater than 0). This demonstrates that among prediction models with the same Gini and Pietra indices, a higher scaled Brier score reflects a stronger ability to shift high-risk individuals toward more certain diseased predictions or low-risk individuals toward more certain non-diseased predictions.

We also compared the risk stratification properties of these three indices across populations with varying disease prevalence. The results consistently confirm that the

doi:10.6342/NTU202504305

Pietra index, Gini index, and scaled Brier score respectively reflect a model's ability to resolve intermediate-risk cases, separate predicted risks, and enhance certainty in risk estimation.

doi:10.6342/NTU202504305

Chapter 4 Ensuring Calibration and Unbiasedness in Prediction Models

The preceding development assumes that the prediction model is both well-calibrated and unbiased. To ensure these conditions are met in practice, we propose a three-step adjustment procedure—cross-validation, calibration, and bootstrap averaging—as detailed below:

4.1 Cross-Validation

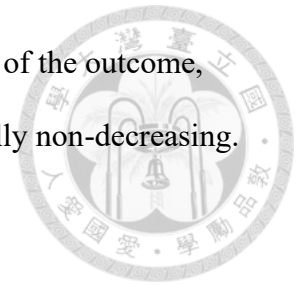
The dataset is randomly partitioned into K subsets (folds). In K-fold cross-validation, the model is trained on $K-1$ folds and tested on the remaining fold, rotating through all folds so that each data point is used for validation once. Variants include leave-one-out cross-validation (a special case where K equals the number of observations), repeated K-fold cross-validation, and stratified cross-validation[8], which maintains the outcome distribution across folds. This process ensures that the predicted risks used in evaluation are derived from models not trained on the individuals being predicted, thereby mimicking predictions for unseen individuals and supporting valid model assessment.

4.2 Calibration

To align predicted risks with observed outcomes, the cross-validated predictions are further calibrated using isotonic regression.[9] This non-parametric method assumes a monotonic increasing relationship between predicted risk and true outcomes. Individuals are first sorted by predicted risk in ascending order. The Pool-Adjacent-Violators Algorithm (PAVA) is then applied to enforce monotonicity by averaging adjacent segments where the observed outcomes violate this assumption. This process

doi:10.6342/NTU202504305

yields calibrated risk estimates that better reflect the true probability of the outcome, ensuring that the resulting predictiveness curve remains monotonically non-decreasing.



4.3 Bootstrap Averaging

When the sample size is limited, the predictiveness curve derived using the above PAVA-based calibration may appear jagged or step-like, as violations of the monotonicity assumption become more frequent, leading to more frequent averaging of adjacent segments. We propose bootstrap averaging to mitigate this problem. Specifically, multiple bootstrap samples[10] are drawn from the original dataset along with the corresponding cross-validated risks. Each bootstrap sample undergoes isotonic regression via the PAVA procedure to produce a monotonic predictiveness curve. Averaging these bootstrap predictiveness curves preserves monotonicity while smoothing out the step-like jumps inherent in individual curves, resulting in a more stable and visually interpretable final curve.

doi:10.6342/NTU202504305

Chapter 5 A Case Study: Five-Year Fatality Among Lung Cancer Patients in Taiwan

We used data from the Taiwan Cancer Registry[11, 12] to evaluate the performance of a prediction model for five-year fatality among lung cancer patients. The dataset included 23,839 patients diagnosed between 2017 and 2018: 18,885 with adenocarcinoma, 3,247 with squamous cell carcinoma, 1,660 with small cell carcinoma, and 47 with large cell carcinoma; patients with other cell types were excluded from the analysis. The dataset provided detailed information on patients' demographic characteristics, lifestyle behaviors, medical histories, and clinical data. To determine survival time, the cancer registry data were linked with mortality records, allowing calculation of time-to-death from the date of diagnosis.

We developed a five-year fatality risk prediction model using Cox proportional hazards regression, with the baseline hazard function estimated via the Breslow method to compute individual five-year fatality risks. The model included one continuous covariate (age) and twelve categorical covariates: sex, cancer histology, level of urbanization, hospital level, cancer stage, smoking status, and six treatment modalities (surgery, chemotherapy, radiotherapy, targeted therapy, immunotherapy, and palliative care).

To ensure calibration and unbiasedness of the predictions, we applied the three-step adjustment procedure described above to derive the adjusted five-year fatality risk for each patient. This involved five-fold cross-validation stratified by five-year survival status, followed by isotonic regression calibration. The final predicted risks—referred to as adjusted risks—were obtained by averaging across 50 bootstrap samples. These adjusted risks were then used to construct the predictiveness curves and compute the

doi:10.6342/NTU202504305

associated geometric summary measures: scaled Brier score, Pietra index, and Gini index. To quantify uncertainty in both the predictiveness curves and summary measures, an outer loop of 100 bootstrap resamples was performed.

Figure 3A presents the predictiveness curve for five-year fatality among lung cancer patients (combining all four cell types). This curve reveals how patients are stratified across the risk spectrum. For example, 40.8% of patients have a predicted five-year fatality risk below the population average of 0.5539 (indicated by the dotted line and reported in Table 1), while 59.2% exceed this average. The model also identifies 25.1% of patients as very low risk (risk < 0.10) and 50.1% as high risk (risk > 0.75). In contrast, only 2.2% and 5.1% of patients fall into the gray zone, with predicted risks within ±0.05 and ±0.1 of the average, respectively.

Figures 3B and 3C display the corresponding Lorenz and ROC curves for comparison. Unlike the predictiveness curve, these plots do not offer direct insights into how patients are stratified by risk; instead, they serve primarily to produce summary indices. However, the same summary indices can be obtained from the predictiveness curve: $\text{Pietra} = 0.6719$ and $\text{Gini} = 0.7850$ (Table 1), and $\text{AUC} = 0.8925$ (via $\text{Gini} = 2 \times \text{AUC} - 1$), and $\text{MVD} = 0.6719$ (noting that MVD equals Pietra). Moreover, the predictiveness curve also enables computation of the scaled Brier score (0.5186 in Table 1), a valuable performance measure not available from either the Lorenz or ROC curves.

Figure 4 presents the predictiveness curves for five-year fatality among lung cancer patients, stratified by cell type, with corresponding geometric summary measures shown in Table 1. The curves differ substantially across cell types, indicating variable risk stratification performance of the prediction model. The predictiveness curve for

doi:10.6342/NTU202504305

adenocarcinoma (Figure 4A) shows the steepest slope near the average fatality risk line, consistent with its highest Pietra index, indicating fewer patients with risks clustered around the mean and greater resolution of intermediate-risk cases. Adenocarcinoma also achieves the highest Gini index, reflecting the strongest horizontal (patient-wise) separation of fatality risks. This is visually evident in its predictiveness curve, which shows a broader spread when moving along the x-axis. The scaled Brier score is likewise highest for adenocarcinoma, corresponding to its predictiveness curve showing more patients assigned to the extreme low- or high-risk ends, indicating greater certainty in risk predictions. In contrast, the curves for squamous cell carcinoma (Figure 4B), small cell carcinoma (Figure 4C), and large cell carcinoma (Figure 4D) are flatter near the average risk line, less horizontal spread, and lower vertical certainty compared to adenocarcinoma.

doi:10.6342/NTU202504305

Chapter 6 Discussion

The ROC curve has long been used to evaluate the performance of individual biomarkers and diagnostic tests, followed by the Lorenz curve; both have since been extended to assess risk-prediction models.[6, 7,13-15] These curves primarily serve to generate summary indices—AUC and MVD for the ROC curve, and Gini and Pietra for the Lorenz curve.[6] Among them, the AUC (and equivalently, the Gini index) has dominated the evaluation of prediction models. However, as demonstrated in this paper, these indices can also be derived from the predictiveness curve.[16] Unlike ROC and Lorenz curves, the predictiveness curve offers an additional summary measure—the scaled Brier score—and, more importantly, the entire curve provides meaningful insights. It visually and quantitatively illustrates how a prediction model stratifies risk across a population, identifying what proportion falls into different risk categories. In contrast, the ROC and Lorenz curves serve only as intermediaries for computing their respective indices and offer little direct information about population-level risk stratification.

Prediction models require cross-validation to avoid overly optimistic estimates of discrimination performance; however, calibration is equally—if not more—crucial to ensure that predicted risks accurately reflect true outcome probabilities. This paper advocates integrating cross-validation and calibration into a unified development process, operationalized through a three-step adjustment procedure: cross-validation, calibration, and bootstrap averaging. Appendix 9 illustrates the predictiveness curves for the four lung cancer cell types, based on models incorporating increasing levels of adjustment: none, one step, two steps, and all three steps. For adenocarcinoma, squamous cell carcinoma, and small cell carcinoma, the curves from models with and

doi:10.6342/NTU202504305

without cross-validation are nearly identical due to their large sample sizes. In contrast, for large cell carcinoma (n = 47), cross-validation (one-step adjustment) noticeably alters the curve, reflecting the impact of limited data. Calibration applied after cross-validation (two-step adjustment) leads to substantial shifts in all four cell types, particularly at the extremes of the risk distribution. For instance, in large cell carcinoma, the uncalibrated model assigns 23.4% and 61.7% of patients predicted risks near 0 and 1, respectively—an overconfident estimate corrected by calibration. Nevertheless, the two-step procedure results in step-like, unsmoothed curves in smaller samples (e.g., squamous, small cell, and large cell carcinoma), which are effectively smoothed by applying the full three-step adjustment.

Appendix 10 presents the adjustment curves for the four lung cancer cell types. These curves plot the final adjusted five-year fatality risks—obtained through the full three-step procedure—against the cross-validated risks from the first step, which emulate the raw predicted risks for prospective new patients. The 45-degree reference lines (dotted) represent perfect agreement between predicted and observed risks; deviations from these lines indicate areas where the model tends to over- or underestimate risk and thus requires adjustment. These adjustment curves should be considered integral components of the prediction model. Alongside conventional model outputs—such as the regression coefficients from the Cox model and the baseline hazard estimated via the Breslow method—they are essential for producing accurate risk estimates in new individuals.

In conclusion, this study extends the application of the predictiveness curve from single biomarkers and diagnostic tests to multivariable risk prediction models, introducing three intuitive geometric summaries—the Pietra index, Gini index, and

doi:10.6342/NTU202504305

scaled Brier score—to comprehensively evaluate predictive performance. Demonstrated through analytical derivation, illustrative examples, and a case study using a large cohort of lung cancer patients, this framework provides deeper insights into critical aspects of risk stratification, such as gray-zone resolution, risk separation, and prediction certainty. The three-step adjustment procedure—cross-validation, calibration, and bootstrap averaging—ensures reliable and robust predictive models, complementing and enhancing conventional model evaluation methods. These tools and methods thus offer researchers and healthcare practitioners a transparent, population-oriented approach for refining and assessing prediction models to improve decision-making in clinical medicine and public health.

doi:10.6342/NTU202504305

References

1. Vogenberg FR. Predictive and prognostic models: implications for healthcare decision-making in a modern recession. *American Health & Drug Benefits*. 2009;2(6):218-222.
2. Smeden MV, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. *Journal of Clinical Epidemiology*. 2021;132:142-145.
3. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, and Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.
4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
5. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics*. 2007;63(4):1181-1188.
6. Lee WC. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Statistics in Medicine*. 1999;18(4):455-471.
7. Wu YC, Lee WC. Alternative performance measures for prediction models. *PLOS One*. 2014;9(3):e91249.
8. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Statistics Surveys.* 2010;4:40-79.
9. Jiang X, Osl M, Kim J, Machado LO. Smooth isotonic regression: a new method to calibrate predictive models. *AMIA Summits on Translational Science Proceedings*. 2011:16-20.

doi:10.6342/NTU202504305

10. Tibshirani RJ, Efron B. *An introduction to the bootstrap.* 1st ed. New York: Chapman and Hall; 1993.
11. Chiang CJ, Wang YW, Lee WC. Taiwan's nationwide cancer registry system of 40 years: past, present, and future. *Journal of the Formosan Medical Association.* 2019;118(5):856-858.
12. Kao CW, Chiang CJ, Lin LJ, Huang CW, Lee WC, Lee MY, Taiwan Society of Cancer Registry Expert Group. Accuracy of long-form data in the Taiwan cancer registry. *Journal of the Formosan Medical Association.* 2021;120(11):2037-2041.
13. Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine*. 3rd ed. New York: John Wiley & Sons; 2014.
14. Mauguen A, Begg CB. Using the Lorenz curve to characterize risk predictiveness and etiologic heterogeneity. *Epidemiology*. 2016;27(4):531-537.
15. Maiga A, Farjah F, Blume J, Deppen S, Welty VF, D'Agostino RS, Colditz GA, Kozower BD, Grogan EL. Risk prediction in clinical practice: a practical guide for cardiothoracic surgeons. *Annals of Thoracic Surgery*. 2019;108(5):1573-1582.
16. Gleiss A. Visualizing a marker's degrees of necessity and of sufficiency in the predictiveness curve. *BMC Medical Research Methodology*. 2025;25(1):107.

doi:10.6342/NTU202504305

Table Note

Table 1. Geometric summary measures of the predictiveness curves for five-year fatality prediction models among lung cancer patients, presented both overall (all four cell types combined) and stratified by cell type.

doi:10.6342/NTU202504305

Table 1.

Types	Average five-year fatality risk	Pietra [95% confidence interval]	Gini [95% confidence interval]	scaled Brier [95% confidence interval]
Overall lung cancer	0.5539	0.6719 [0.6652, 0.6843]	0.7850 [0.7777, 0.7942]	0.5186 [0.5106, 0.5314]
Adenocarcinoma	0.4970	0.6979 [0.6906, 0.7054]	0.8026 [0.7942, 0.8111]	0.5386 [0.5281, 0.5503]
Squamous cell carcinoma	0.7197	0.4871 [0.4553, 0.5180]	0.6185 [0.5801, 0.6532]	0.2862 [0.2628, 0.3212]
Small cell carcinoma	0.8753	0.3380 [0.2880, 0.4145]	0.4378 [0.3814, 0.5287]	0.1195 [0.0873, 0.1825]
Large cell carcinoma	0.7226	0.4827 [0.3017, 0.7716]	0.5729 [0.3704, 0.8614]	0.2544 [0.1112, 0.5933]

doi:10.6342/NTU202504305

Figure Note

Figure 1. Predictiveness curves and geometry: (A) informative but imperfect model, (B) null model, (C) perfect model (blue line: predictiveness curve; dotted line: disease prevalence in the population; orange shaded region: below-the-line region; red shaded region: above-the-line region; orange dot: center of gravity of the below-the-line region; red dot: center of gravity of the above-the-line region).

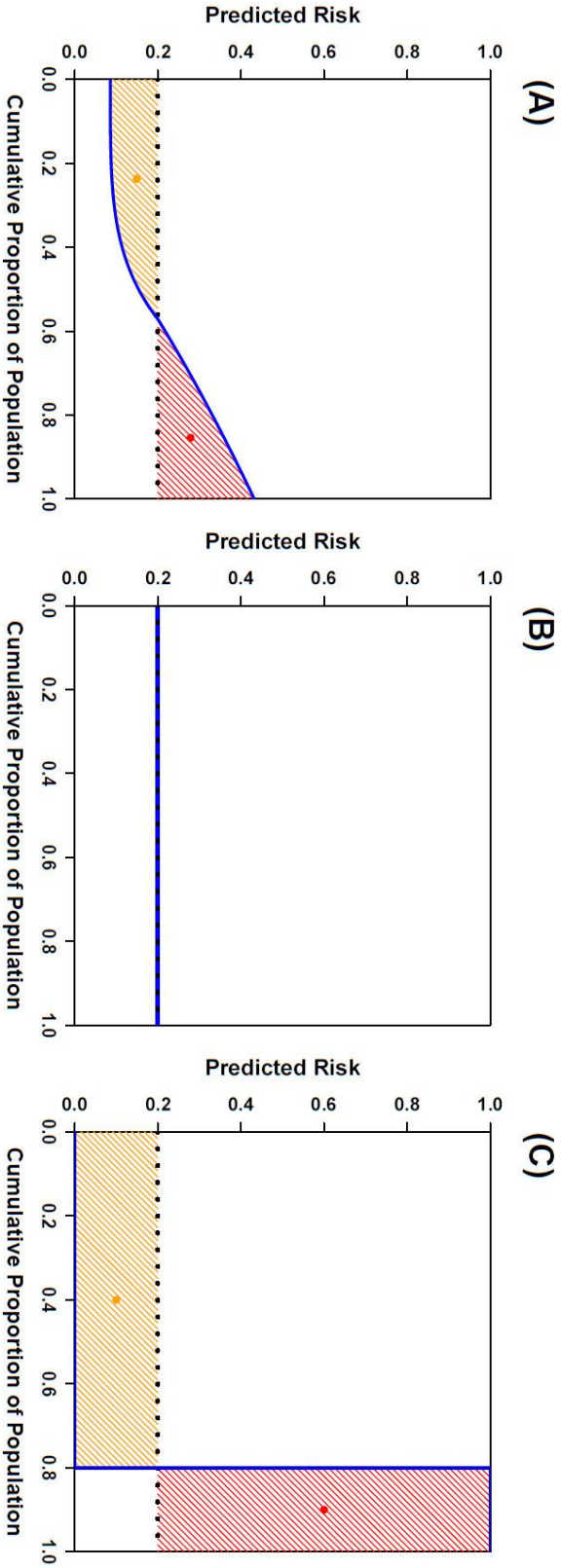
Figure 2. Predictiveness curves for nine prediction models applied to the same population with a disease prevalence of 0.2 (solid lines: predictiveness curves; dotted line: disease prevalence). Models I, II, and III differ solely in their Pietra indices; models IV, V, and VI differ solely in their Gini indices; and models VII, VIII, and IV differ solely in their scaled Brier scores.

Figure 3. Evaluation curves for five-year fatality among lung cancer patients: (A) Predictiveness curve, (B) Lorenz curve, and (C) Receiver Operating Characteristic (ROC) curve. In panel (A), the dotted line marks the average five-year fatality risk; in panels (B) and (C), it represents the diagonal reference line indicating no discriminatory power. Shaded regions denote 95% bootstrap confidence intervals.

Figure 4. Predictiveness curves for five-year fatality among lung cancer patients, stratified by cell type: (A) adenocarcinoma, (B) squamous cell carcinoma, (C) small cell carcinoma, and (D) large cell carcinoma. Dotted lines indicate the average five-year fatality risk for each subtype. Shaded regions represent 95% bootstrap confidence intervals.

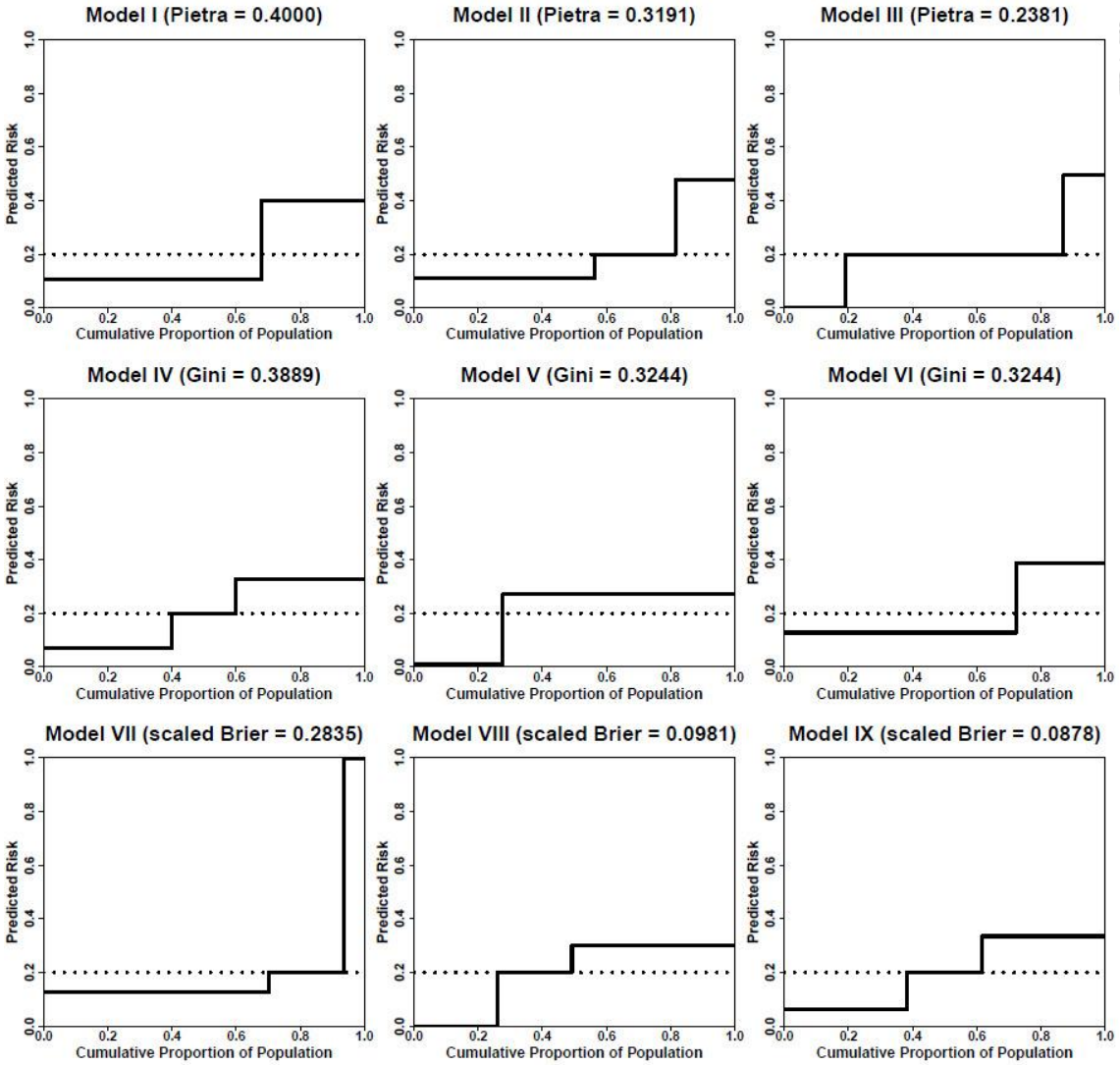
doi:10.6342/NTU202504305

Figure 1.



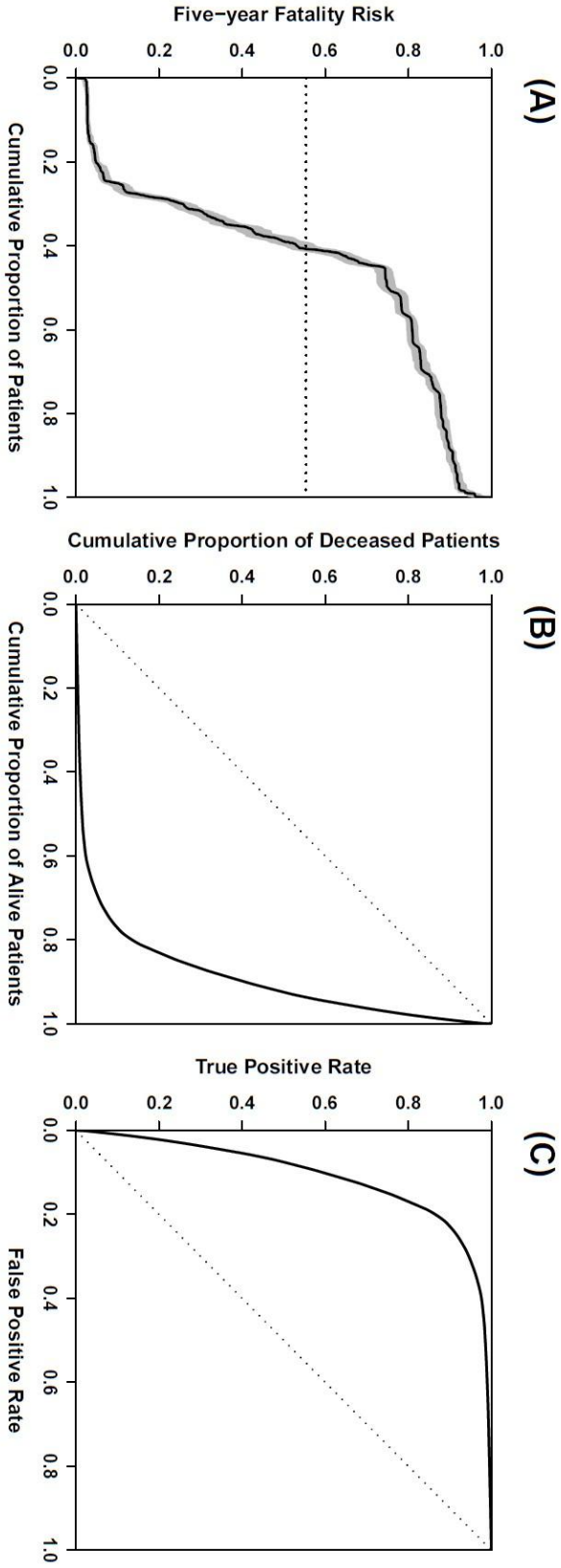
doi:10.6342/NTU202504305

Figure 2.



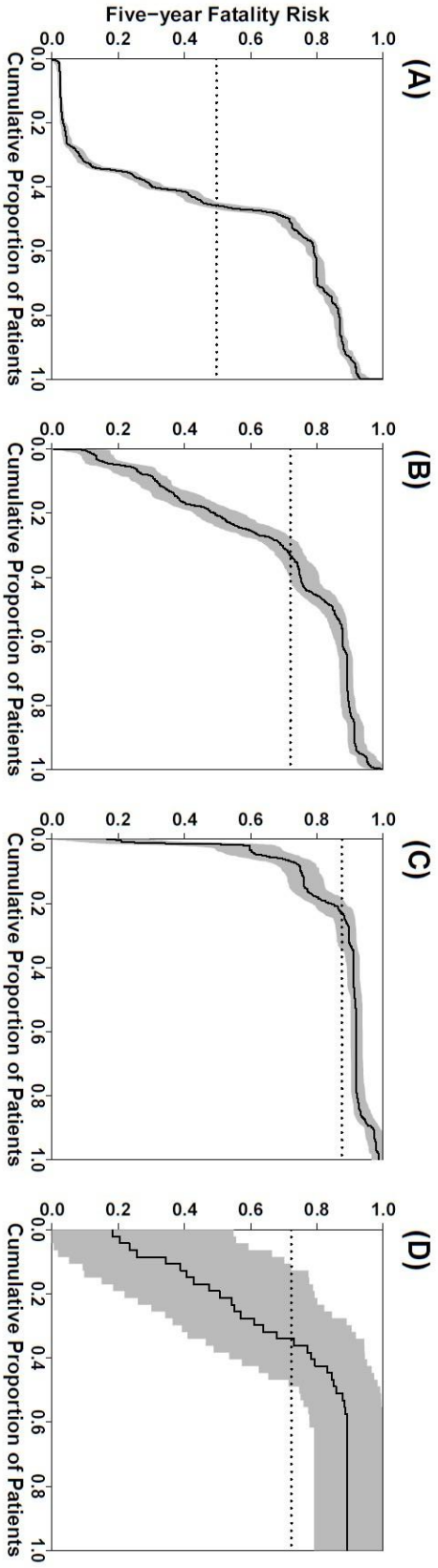
doi:10.6342/NTU202504305

Figure 3.



doi:10.6342/NTU202504305

Figure 4.



doi:10.6342/NTU202504305

Appendix Note

Appendix 1. Relationship among the area under the predictiveness curve, mean predicted risk, and disease prevalence in the population.

Appendix 2. Relationship between the areas of the below-the-line and the above-the-line regions.

Appendix 3. The Pietra index as a geometric summary of the predictiveness curve.

Appendix 4. The Gini index as a geometric summary of the predictiveness curve.

Appendix 5. The scaled Brier score as a geometric summary of the predictiveness curve.

Appendix 6. Relationships among the Pietra index, the Gini index, and the scaled Brier score.

Appendix 7. Example predictiveness curve calculations for an informative but imperfect model.

Appendix 8. Prediction Models I to IX Calculations.

Appendix 9. Predictiveness curves after stepwise application of cross-validation, calibration, and bootstrap smoothing for five-year fatality among lung cancer patients, stratified by cell type: (A) adenocarcinoma, (B) squamous cell carcinoma, (C) small cell carcinoma, and (D) large cell carcinoma. Curves are color-coded by adjustment level: blue = unadjusted, brown = cross-validation only, red = cross-validation with calibration, green = full three-step adjustment. Dotted lines indicate the average five-year fatality risk.

doi:10.6342/NTU202504305

Appendix 10. Adjustment curves for five-year fatality prediction models among lung cancer patients, stratified by cell type: (A) adenocarcinoma, (B) squamous cell carcinoma, (C) small cell carcinoma, and (D) large cell carcinoma. Each solid curve plots the final adjusted risks—obtained through cross-validation, calibration, and bootstrap averaging—against the cross-validated risks prior to calibration, which emulate the raw predicted risks for prospective new patients. Dotted 45-degree reference lines indicate perfect agreement between predicted and true risks.

doi:10.6342/NTU202504305

Appendix 1. Relationship among the area under the predictiveness curve, mean predicted risk, and disease prevalence in the population.

Let R be a random variable representing the predicted risk of disease, ranging from 0 to 1, with probability density function $f(r)$. The area under the predictiveness curve is $\int_0^1 r dF(r)$, where $F(r)$ is the cumulative distribution function of R. Since $dF(r) = f(r)dr$, this area equals $\int_0^1 r f(r)\, dr$, which is the mean predicted risk. For a well-calibrated and unbiased prediction model—where approximately $100 \times r\%$ of individuals with predicted risk r are truly diseased—this value equals the disease prevalence π in the population.

doi:10.6342/NTU202504305

Appendix 2. Relationship between the areas of the below-the-line and the above-the-line regions.

The area of the below-the-line (BL) region is $\int_0^{F(\pi)}(\pi - r)dF(r) = \pi \int_0^{F(\pi)} 1\, dF(r) - \int_0^{F(\pi)} r\, dF(r)$. The area of the above-the-line (AL) region is $\int_{F(\pi)}^1 (r - \pi)dF(r) = \int_{F(\pi)}^1 r\, dF(r) - \pi \times \int_{F(\pi)}^1 1\, dF(r)$. For a well-calibrated and unbiased prediction model, the sum of the integrals of 1 over the full domain is 1, and the total area under the predictiveness curve equals the disease prevalence, π. Thus, the difference between the areas of the AL and BL regions is $\int_0^1 r\, dF(r) - \pi \times \int_0^1 1\, dF(r) = \pi - \pi = 0,$ which shows that the two regions have equal area.

doi:10.6342/NTU202504305

Appendix 3. The Pietra index as a geometric summary of the predictiveness curve.

The Pietra index measures the average gain in information provided by the prediction model, defined as the absolute difference between the predicted risk (posterior probability) and the disease prevalence (prior probability). It is calculated as:

$$\text{Pietra} = \frac{\text{the mean gain provided by the given prediction model}}{\text{the mean gain provided by the perfect prediction model}}$$

$$= \frac{\int_0^1 |r-\pi| dF(r)}{\pi \times |1-\pi| + (1-\pi) \times |0-\pi|}$$

$$= \frac{\int_0^{F(\pi)} (\pi - r) dF(r) + \int_{F(\pi)}^1 (r-\pi) dF(r)}{2 \times \pi \times (1-\pi)}$$

$$= \frac{\text{area of BL region} + \text{area of AL region}}{2 \times \pi \times (1-\pi)} = \frac{A}{\pi \times (1-\pi)},$$

where the prediction model is assumed to be unbiased and well-calibrated such that the below-the-line (BL) and the above-the-line (AL) regions have equal area, A.

doi:10.6342/NTU202504305

Appendix 4. The Gini index as a geometric summary of the predictiveness curve.

The Gini index measures separation, defined as the absolute difference between the predicted probabilities of two randomly selected individuals. It is calculated using the formula:

$$\text{Gini} = \frac{\text{the mean separation provided by the given prediction model}}{\text{the mean separation provided by the perfect prediction model}}$$

$$= \frac{\int_0^1 \int_0^1 |r_1 - r_2| dF(r_1) dF(r_2)}{|1-1| \times \pi^2 + |1-0| \times \pi \times (1-\pi) + |0-1| \times (1-\pi) \times \pi + |0-0| \times (1-\pi)^2}$$

$$= \frac{\int_0^1 \int_0^1 |r_1 - r_2| dF(r_1) dF(r_2)}{2 \times \pi \times (1-\pi)} = \frac{\int_0^1 \int_0^1 [\max(r_1, r_2) - \min(r_1, r_2)] dF(r_1) dF(r_2)}{2 \times \pi \times (1-\pi)}$$

$$= \frac{1}{2 \times \pi \times (1-\pi)} \times \left\{ \left[\int_0^1 \int_{F(r_2)}^1 r_1 dF(r_1) dF(r_2) - \int_0^1 \int_0^{F(r_2)} r_1 dF(r_1) dF(r_2) \right] + \left[\int_0^1 \int_0^{F(r_2)} r_2 dF(r_1) dF(r_2) - \int_0^1 \int_{F(r_2)}^1 r_2 dF(r_1) dF(r_2) \right] \right\}$$

$$= \frac{1}{2 \times \pi \times (1-\pi)} \times \left\{ \left[2 \times \pi \times \int_0^1 \int_{F(r_2)}^1 1 dF(r_1) dF(r_2) - \int_0^1 \int_{F(r_2)}^1 r_2 dF(r_1) dF(r_2) + \int_0^1 \int_0^{F(r_2)} r_2 dF(r_1) dF(r_2) - \int_0^1 \int_0^1 r_2 dF(r_1) dF(r_2) \right] + \left[2 \times \pi \times \int_0^1 \int_0^{F(r_2)} 1 dF(r_1) dF(r_2) - \int_0^1 \int_0^{F(r_2)} r_1 dF(r_1) dF(r_2) + \int_0^1 \int_{F(r_2)}^1 r_1 dF(r_1) dF(r_2) - \int_0^1 \int_0^1 r_1 dF(r_1) dF(r_2) \right] \right\}$$

doi:10.6342/NTU202504305

$$= \frac{1}{\pi \times (1-\pi)}$$

$$\times \left[2 \times \pi \times \int_0^1 \int_{F(r_2)}^1 1 dF(r_1) dF(r_2) - \int_0^1 \int_{F(r_2)}^1 r_2 dF(r_1) dF(r_2) \right.$$

$$\left. + \int_0^1 \int_0^{F(r_2)} r_2 dF(r_1) dF(r_2) - \int_0^1 \int_0^1 r_2 dF(r_1) dF(r_2) \right]$$

$$= \frac{1}{\pi \times (1-\pi)}$$

$$\times \left[2 \times \pi \times \int_0^1 \int_{F(r_2)}^1 1 dF(r_1) dF(r_2) - 2 \times \int_0^1 \int_{F(r_2)}^1 r_2 dF(r_1) dF(r_2) \right]$$

$$= \frac{1}{\pi \times (1-\pi)}$$

$$\times \left\{ 2 \times \int_0^1 \left[[1 - F(r)] \times \pi \right] dF(r) - 2 \times \int_0^1 \left[[1 - F(r)] \times r \right] dF(r) \right\}$$

$$= \frac{\int_0^{F(\pi)} (\pi - r)\, dF(r) + \int_{F(\pi)}^1 (r - \pi)\, dF(r)}{\pi \times (1-\pi)}$$

$$\times \left\{ \frac{\int_{F(\pi)}^1 \left[[1 - F(r)] \times \pi \right] dF(r)}{\int_{F(\pi)}^1 (r - \pi)\, dF(r)} + \frac{\int_0^{F(\pi)} \left[[1 - F(r)] \times \pi \right] dF(r)}{\int_0^{F(\pi)} (\pi - r)\, dF(r)} \right.$$

$$\left. - \frac{\int_{F(\pi)}^1 \left[[1 - F(r)] \times r \right] dF(r)}{\int_{F(\pi)}^1 (r - \pi)\, dF(r)} - \frac{\int_0^{F(\pi)} \left[[1 - F(r)] \times r \right] dF(r)}{\int_0^{F(\pi)} (\pi - r)\, dF(r)} \right\}$$

$$= 2 \times \text{Pietra}$$

$$\times \left\{ \frac{\int_{F(\pi)}^1 (r - \pi)\, dF(r) - \int_{F(\pi)}^1\left[[1 - F(r)] \times (r - \pi) \right] dF(r)}{\int_{F(\pi)}^1 (r - \pi)\, dF(r)} \right.$$

$$\left. - \left\{ \frac{\int_0^{F(\pi)} (\pi - r)\, dF(r) - \int_0^{F(\pi)} \left[[1 - F(r)] \times (\pi - r) \right] dF(r)}{\int_0^{F(\pi)} (\pi - r)\, dF(r)} \right\} \right\}$$

doi:10.6342/NTU202504305

$$= 2 \times \left\{ \frac{\int_{F(\pi)}^{1} [F(r) \times (r - \pi)]\, dF(r)}{\int_{F(\pi)}^{1} (r - \pi)\, dF(r)} - \frac{\int_{0}^{F(\pi)} [F(r) \times (\pi - r)]\, dF(r)}{\int_{0}^{F(\pi)} (\pi - r)\, dF(r)} \right\} \times \text{Pietra}$$

$$= 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra},$$

where $x_{\text{AL}} = \frac{\int_{F(\pi)}^{1} [F(r) \times (r - \pi)]\, dF(r)}{\int_{F(\pi)}^{1} (r - \pi)\, dF(r)}$ represents the x-coordinate of the center of gravity for the above-the-line (AL) region and $x_{\text{BL}} = \frac{\int_{0}^{F(\pi)} [F(r) \times (\pi - r)]\, dF(r)}{\int_{0}^{F(\pi)} (\pi - r)\, dF(r)}$ represents the x-coordinate of the center of gravity for the below-the-line (BL) region.

doi:10.6342/NTU202504305

Appendix 5. The scaled Brier score as a geometric summary of the predictiveness curve.

The scaled Brier score measures squared gain, defined as the squared difference between the predicted probability and the disease prevalence. It is calculated using the formula:

$$\text{scaled Brier} = \frac{\text{the mean squared gain provided by the given prediction model}}{\text{the mean squared gain provided by the perfect prediction model}}$$

$$= \frac{\int_0^1 (r-\pi)^2 dF(r)}{(1-\pi)^2 \times \pi + (0-\pi)^2 \times (1-\pi)} = \frac{\int_0^1 (r-\pi)^2 dF(r)}{\pi \times (1-\pi)}$$

$$= \frac{1}{2 \times \pi \times (1-\pi)}$$

$$\times \left\{ \frac{\int_{F(\pi)}^1 (r-\pi)^2 dF(r)}{\int_{F(\pi)}^1 (r-\pi) dF(r)} \times \left[\int_0^{F(\pi)} (\pi - r) dF(r) + \int_{F(\pi)}^1 (r-\pi) dF(r) \right] \right.$$

$$\left. + \frac{\int_0^{F(\pi)} (r-\pi)^2 dF(r)}{\int_0^{F(\pi)} (\pi - r) dF(r)} \times \left[\int_0^{F(\pi)} (\pi - r) dF(r) + \int_{F(\pi)}^1 (r-\pi) dF(r) \right] \right\}$$

$$= \frac{\int_0^{F(\pi)} (\pi - r) dF(r) + \int_{F(\pi)}^1 (r-\pi) dF(r)}{2 \times \pi \times (1-\pi)}$$

$$\times \left[\frac{\int_{F(\pi)}^1 (r-\pi)^2 dF(r)}{\int_{F(\pi)}^1 (r-\pi) dF(r)} + \frac{\int_0^{F(\pi)} (r-\pi)^2 dF(r)}{\int_0^{F(\pi)} (\pi - r) dF(r)} \right]$$

$$= \text{Pietra}$$

$$\times \left\{ \frac{\int_{F(\pi)}^1 (r-\pi)^2 dF(r) + 2 \times \pi \times \int_{F(\pi)}^1 (r-\pi) dF(r)}{\int_{F(\pi)}^1 (r-\pi) dF(r)} \right.$$

$$\left. - \left[\frac{2 \times \pi \times \int_0^{F(\pi)} (\pi - r) dF(r) - \int_0^{F(\pi)} (r-\pi)^2 dF(r)}{\int_0^{F(\pi)} (\pi - r) dF(r)} \right] \right\}$$

doi:10.6342/NTU202504305

$$= 2 \times \left[\frac{\frac{1}{2} \times \int_{F(\pi)}^{1} (r^2 - \pi^2) dF(r)}{\int_{F(\pi)}^{1} (r - \pi) dF(r)} - \frac{\frac{1}{2} \times \int_{0}^{F(\pi)} (\pi^2 - r^2) dF(r)}{\int_{0}^{F(\pi)} (\pi - r) dF(r)} \right] \times \text{Pietra}$$

$$= 2 \times \left\{ \frac{\int_{F(\pi)}^{1} \left[\frac{\pi + r}{2} \times (r - \pi) \right] dF(r)}{\int_{F(\pi)}^{1} (r - \pi) dF(r)} - \frac{\int_{0}^{F(\pi)} \left[\frac{\pi + r}{2} \times (\pi - r) \right] dF(r)}{\int_{0}^{F(\pi)} (\pi - r) dF(r)} \right\}$$

$$\times \text{Pietra}$$

$$= 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra},$$

where $y_{\text{AL}} = \frac{\int_{F(\pi)}^{1} \left[\frac{\pi + r}{2} \times (r - \pi) \right] dF(r)}{\int_{F(\pi)}^{1} (r - \pi) dF(r)}$ represents the y-coordinate of the center of gravity for the above-the-line (AL) region and $y_{\text{BL}} = \frac{\int_{0}^{F(\pi)} \left[\frac{\pi + r}{2} \times (\pi - r) \right] dF(r)}{\int_{0}^{F(\pi)} (\pi - r) dF(r)}$ represents the y-coordinate of the center of gravity for the below-the-line (BL) region.

doi:10.6342/NTU202504305

Appendix 6. Relationships among the Pietra index, the Gini index, and the scaled Brier score.

Let r_{L} and r_{H} represent the lowest and highest risks on the predictiveness curve, respectively ($0 \le r_{\mathrm{L}} \le \pi \le r_{\mathrm{H}} \le 1$). The horizontal line at $r = \pi$ divides the predictiveness curve into three segments: one (r_{BL}) below the line, spanning from (0, r_{L}) to ($F_{\mathrm{BL}}(\pi)$, π), one extending along it, spanning from ($F_{\mathrm{BL}}(\pi)$, π) to ($F_{\mathrm{AL}}(\pi)$, π), and one (r_{AL}) above it, spanning from ($F_{\mathrm{AL}}(\pi)$, π) to (1, r_{H}), where $0 \le F_{\mathrm{BL}}(\pi) \le F_{\mathrm{AL}}(\pi) \le 1$.

Note that r_{BL} is a monotonically non-decreasing function for $0 \le F(r) \le F_{\mathrm{BL}}(\pi)$, which implies $0 \le (\pi - r_{\mathrm{BL}}) \le \pi - r_{\mathrm{L}}$ over this range. Similarly, r_{AL} is a monotonically non-decreasing function for $F_{\mathrm{AL}}(\pi) \le F(r) \le 1$, implying $0 \le (r_{\mathrm{AL}} - \pi) \le r_{\mathrm{H}} - \pi$. As a result, the areas of the BL and AL regions are constrained as: $0 \le A_{\mathrm{BL}} = \int_0^{F_{\mathrm{BL}}(\pi)} (\pi - r_{\mathrm{BL}}) dF(r) \le (\pi - r_{\mathrm{L}}) \times F_{\mathrm{BL}}(\pi)$, and $0 \le A_{\mathrm{AL}} = \int_{F_{\mathrm{AL}}(\pi)}^{1} (r_{\mathrm{AL}} - \pi) dF(r) \le (r_{\mathrm{H}} - \pi) \times [1 - F_{\mathrm{AL}}(\pi)]$. Since the prediction model is assumed to be well-calibrated and unbiased, A_{BL} and A_{AL} must be equal. Thus, the Pietra index is bounded as: $0 \le \text{Pietra} = \frac{A_{\mathrm{BL}}}{\pi \times (1-\pi)} = \frac{A_{\mathrm{AL}}}{\pi \times (1-\pi)} \le \frac{(\pi - r_{\mathrm{L}}) \times F_{\mathrm{BL}}(\pi)}{\pi \times (1-\pi)} = \frac{(r_{\mathrm{H}} - \pi) \times [1 - F_{\mathrm{AL}}(\pi)]}{\pi \times (1-\pi)}$. For a perfect prediction model, $r_{\mathrm{L}} = 0$, r_{H}=1, and $F_{\mathrm{BL}}(\pi) = F_{\mathrm{AL}}(\pi) = 1 - \pi$, making the upper bound equal to 1. Therefore, the Pietra index satisfies the constraint: $0 \le \text{Pietra} \le 1$.

The x-coordinate of the center of gravity for the BL region is given by $x_{\mathrm{BL}} = \frac{\int_0^{F_{\mathrm{BL}}(\pi)} [F(r) \times (\pi - r_{\mathrm{BL}})] dF(r)}{\int_0^{F_{\mathrm{BL}}(\pi)} (\pi - r_{\mathrm{BL}}) dF(r)}$, and it is constrained by $x_{\mathrm{BL}} \le \frac{F_{\mathrm{BL}}(\pi)}{2}$. Similarly, the x-coordinate of the center of gravity for the AL region is given by $x_{\mathrm{AL}} =$

doi:10.6342/NTU202504305

$\frac{\int_{F_{\mathrm{AL}}(\pi)}^{1}[F(r)\times(r_{\mathrm{AL}}-\pi)]dF(r)}{\int_{F_{\mathrm{AL}}(\pi)}^{1}(r_{\mathrm{AL}}-\pi)dF(r)}$, and it is constrained by $x_{\mathrm{AL}} \geq \frac{1+F_{\mathrm{AL}}(\pi)}{2}$. Thus, the difference between the two x-coordinates satisfies $(x_{\mathrm{AL}} - x_{\mathrm{BL}}) \geq \frac{1+F_{\mathrm{AL}}(\pi)-F_{\mathrm{BL}}(\pi)}{2} \geq \frac{1}{2}$. From Appendix 4, the Gini index is expressed as $\mathrm{Gini} = 2 \times (x_{\mathrm{AL}} - x_{\mathrm{BL}}) \times \mathrm{Pietra}$. This relationship implies that the Gini index is bounded below by the Pietra index. For a perfect prediction model, $\mathrm{Pietra} = 1$, $x_{\mathrm{BL}} = \frac{1-\pi}{2}$, and $x_{\mathrm{AL}} = 1 - \frac{\pi}{2}$. Consequently, the Gini index is bounded above by 1.

The y-coordinate of the center of gravity for the BL region is given by $y_{\mathrm{BL}} = \frac{\int_{0}^{F_{\mathrm{BL}}(\pi)}\left[\frac{\pi+r}{2}\times(\pi-r)\right]dF(r)}{\int_{0}^{F_{\mathrm{BL}}(\pi)}(\pi-r)dF(r)}$, and it is constrained by $y_{\mathrm{BL}} \geq \frac{\pi+r_{\mathrm{L}}}{2}$. Similarly, the y-coordinate of the center of gravity for the AL region is given by $y_{\mathrm{AL}} = \frac{\int_{F_{\mathrm{AL}}(\pi)}^{1}\left[\frac{\pi+r}{2}\times(r-\pi)\right]dF(r)}{\int_{F_{\mathrm{AL}}(\pi)}^{1}(r-\pi)dF(r)}$, and it is constrained by $y_{\mathrm{AL}} \leq \frac{\pi+r_{\mathrm{H}}}{2}$. Thus, the difference between the two y-coordinates satisfies $(y_{\mathrm{AL}} - y_{\mathrm{BL}}) \leq \frac{r_{\mathrm{H}}-r_{\mathrm{L}}}{2} \leq \frac{1}{2}$. From Appendix 5, the scaled Brier score is expressed as $\mathrm{scaled\ Brier} = 2 \times (y_{\mathrm{AL}} - y_{\mathrm{BL}}) \times \mathrm{Pietra}$. This relationship implies that the scaled Brier score is bounded above by the Pietra index. For a null prediction model, $\mathrm{Pietra} = 0$, and $y_{\mathrm{BL}} = y_{\mathrm{AL}} = \pi$. Consequently, the scaled Brier score is bounded below by 0.

Taken together, we obtain $0 \leq \mathrm{scaled\ Brier} \leq \mathrm{Pietra} \leq \mathrm{Gini} \leq 1$.

doi:10.6342/NTU202504305

Appendix 7. Example predictiveness curve calculations for an informative but imperfect model.

Consider a prediction model with a predictiveness curve defined by the following function:

$$r = \begin{cases} [F(r) + 0.01]^4 + 0.0859 & \text{when } 0 \le F(r) < 0.5712, \\ \sqrt{F(r) + 0.0798} - 0.6068 & \text{when } 0.5712 \le F(r) \le 1. \end{cases}$$

The area under the predictiveness curve, which represents the disease prevalence in the population, is calculated as follows:

$$\begin{aligned} \pi &= \int_0^1 r\, dF(r) \\ &= \int_0^{0.5712} \{[F(r) + 0.01]^4 + 0.0859\}\, dF(r) \\ &\qquad + \int_{0.5712}^1 \left\{\sqrt{F(r) + 0.0798} - 0.6068\right\} dF(r) \\ &= 0.2. \end{aligned}$$

The area of the BL region is calculated as follows:

$$\int_0^{F(\pi)} (\pi - r)\, dF(r) = \int_0^{0.5712} \{0.1141 - [F(r) + 0.01]^4\}\, dF(r) = 0.0519,$$

and the area of the AL region is calculated as follows:

$$\int_{F(\pi)}^1 (r - \pi)\, dF(r) = \int_{0.5712}^1 \left[\sqrt{F(r) + 0.0798} - 0.8068\right] dF(r) = 0.0519,$$

which are equal as expected.

The Pietra index is calculated as follows:

doi:10.6342/NTU202504305

$$\text{Pietra} = \frac{A}{\pi\times(1-\pi)} = \frac{0.0519}{0.2\times(1-0.2)} = 0.3244.$$

The coordinates of the center of gravity for the BL region are calculated as follows:

$$x_{\text{BL}} = \frac{\int_0^{F(\pi)}[F(r)\times(\pi-r)]dF(r)}{\int_0^{F(\pi)}(\pi-r)dF(r)}$$

$$= \frac{\int_0^{0.5712}\{F(r)\times[0.1141-[F(r)+0.01]^4]\}dF(r)}{0.0519}$$

$$= \frac{0.0123}{0.0519} = 0.2370,$$

$$y_{\text{BL}} = \frac{\int_0^{F(\pi)}\left[\frac{\pi+r}{2}\times(\pi-r)\right]dF(r)}{\int_0^{F(\pi)}(\pi-r)dF(r)}$$

$$= \frac{\frac{1}{2}\times\left\{\int_0^{0.5712}\left\{\{[F(r)+0.01]^4+0.2859\}\times\{0.1141-[F(r)+0.01]^4\}\right\}dF(r)\right\}}{0.0519}$$

$$= \frac{0.0078}{0.0519} = 0.1503.$$

The coordinates of the center of gravity for the AL region are calculated as follows:

$$x_{\text{AL}} = \frac{\int_{F(\pi)}^1[F(r)\times(r-\pi)]dF(r)}{\int_{F(\pi)}^1(r-\pi)dF(r)}$$

$$= \frac{\int_{0.5712}^1\left\{F(r)\times\left[\sqrt{F(r)+0.0798}-0.8068\right]\right\}dF(r)}{0.0519}$$

$$= \frac{0.0443}{0.0519} = 0.8536,$$

$$y_{\text{AL}} = \frac{\int_{F(\pi)}^1\left[\frac{\pi+r}{2}\times(r-\pi)\right]dF(r)}{\int_{F(\pi)}^1(r-\pi)dF(r)}$$

$$= \frac{\int_{0.5712}^1\left\{F(r)\times\left[\sqrt{F(r)+0.0798}-0.8068\right]\right\}dF(r)}{0.0519}$$

$$= \frac{0.0145}{0.0519} = 0.2794.$$

doi:10.6342/NTU202504305

The Gini index and the scaled Brier score are calculated as follows:

$$\text{Gini} = 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra} = 2 \times (0.8536 - 0.2370) \times 0.3244 = 0.4001,$$

$$\text{scaled Brier} = 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra} = 2 \times (0.2794 - 0.1503) \times 0.3244$$

$$= 0.0838.$$

doi:10.6342/NTU202504305

Appendix 8. Prediction Models I to IX Calculations

Model I

The predictiveness curve for the prediction model is defined by the following functions:

$$r = \begin{cases} 0.10588 & \text{when } \ 0.00000 \le F(r) < 0.68000, \\ 0.40000 & \text{when } \ 0.68000 \le F(r) \le 1.00000. \end{cases}$$

The Pietra index is calculated as follows:

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi \times (1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi \times (1-\pi)} = \frac{(0.2-0.10588) \times (0.68-0)}{0.2 \times (1-0.2)} = \frac{(0.4-0.2) \times (1-0.68)}{0.2 \times (1-0.2)} = 0.4.$$

The Gini index is calculated as follows:

$$\begin{aligned} \text{Gini} &= 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra} \\ &= 2 \times \left\{ \frac{2 \times [1 + 0.68]}{4} - \frac{2 \times [0.68 + 0]}{4} \right\} \times 0.4 \\ &= 0.4. \end{aligned}$$

The scaled Brier score is calculated as follows:

$$\begin{aligned} \text{sBrier} &= 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra} \\ &= 2 \times \left[\frac{2 \times (0.4 + 0.2)}{4} - \frac{2 \times (0.10588 + 0.2)}{4} \right] \times 0.4 \\ &= 0.1176. \end{aligned}$$

Model II

The predictiveness curve for the prediction model is defined by the following functions:

doi:10.6342/NTU202504305

$$r = \begin{cases} 0.10928 & \text{when } 0.00000 \le F(r) < 0.56281, \\ 0.20000 & \text{when } 0.56281 \le F(r) < 0.81631, \\ 0.47796 & \text{when } 0.81631 \le F(r) \le 1.00000. \end{cases}$$

The Pietra index is calculated as follows:

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi \times (1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi \times (1-\pi)} = \frac{(0.2-0.10928) \times (0.56281-0)}{0.2 \times (1-0.2)} = \frac{(0.47796-0.2) \times (1-0.81631)}{0.2 \times (1-0.2)} =$$

$$0.3191.$$

The Gini index is calculated as follows:

$$\begin{aligned} \text{Gini} &= 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra} \\ &= 2 \times \left\{ \frac{2 \times [1 + 0.81631]}{4} - \frac{2 \times [0.56281 + 0]}{4} \right\} \times 0.3191 \\ &= 0.4. \end{aligned}$$

The scaled Brier score is calculated as follows:

$$\begin{aligned} \text{scaled Brier} &= 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra} \\ &= 2 \times \left[\frac{2 \times (0.47796 + 0.2)}{4} - \frac{2 \times (0.10928 + 0.2)}{4} \right] \times 0.3191 \\ &= 0.1176. \end{aligned}$$

Model III

The predictiveness curve for the prediction model is defined by the following functions:

$$r = \begin{cases} 0.00013 & \text{when } 0.00000 \le F(r) < 0.19061, \\ 0.20000 & \text{when } 0.19061 \le F(r) < 0.87051, \\ 0.49422 & \text{when } 0.87051 \le F(r) \le 1.00000. \end{cases}$$

The Pietra index is calculated as follows:

doi:10.6342/NTU202504305

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi\times(1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi\times(1-\pi)} = \frac{(0.2-0.00013)\times(0.19061-0)}{0.2\times(1-0.2)} = \frac{(0.49422-0.2)\times(1-0.87051)}{0.2\times(1-0.2)} =$$

$$0.2381.$$

The Gini index is calculated as follows:

$$\begin{aligned}\text{Gini} &= 2\times(x_{\text{AL}} - x_{\text{BL}})\times\text{Pietra}\\ &= 2\times\left\{\frac{2\times[1+0.87051]}{4} - \frac{2\times[0.19061+0]}{4}\right\}\times 0.2381\\ &= 0.4.\end{aligned}$$

The scaled Brier score is calculated as follows:

$$\begin{aligned}\text{scaled Brier} &= 2\times(y_{\text{AL}} - y_{\text{BL}})\times\text{Pietra}\\ &= 2\times\left[\frac{2\times(0.49422+0.2)}{4} - \frac{2\times(0.00013+0.2)}{4}\right]\times 0.2381\\ &= 0.1176.\end{aligned}$$

Model IV

The predictiveness curve for the prediction model is defined by the following functions:

$$r = \begin{cases} 0.07045 & \text{when } 0.00000 \le F(r) < 0.40059,\\ 0.20000 & \text{when } 0.40059 \le F(r) < 0.59941,\\ 0.32955 & \text{when } 0.59941 \le F(r) \le 1.00000.\end{cases}$$

The Pietra index is calculated as follows:

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi\times(1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi\times(1-\pi)} = \frac{(0.2-0.07045)\times(0.40059-0)}{0.2\times(1-0.2)} = \frac{(0.32955-0.2)\times(1-0.59941)}{0.2\times(1-0.2)} =$$

$$0.3244.$$

The Gini index is calculated as follows:

doi:10.6342/NTU202504305

$$\text{Gini} = 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra}$$

$$= 2 \times \left\{ \frac{2 \times [1 + 0.59941]}{4} - \frac{2 \times [0.40059 + 0]}{4} \right\} \times 0.3244$$

$$= 0.3889.$$

The scaled Brier score is calculated as follows:

$$\text{scaled Brier} = 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra}$$

$$= 2 \times \left[\frac{2 \times (0.32955 + 0.2)}{4} - \frac{2 \times (0.07045 + 0.2)}{4} \right] \times 0.3244$$

$$= 0.0840.$$

Model V

The predictiveness curve for the prediction model is defined by the following functions:

$$r = \begin{cases} 0.01268 & \text{when} \quad 0.00000 \leq F(r) < 0.27705, \\ 0.27179 & \text{when} \quad 0.27705 \leq F(r) \leq 1.00000. \end{cases}$$

The Pietra index is calculated as follows:

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi \times (1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi \times (1-\pi)} = \frac{(0.2-0.01268) \times (0.27705-0)}{0.2 \times (1-0.2)} = \frac{(0.27179-0.2) \times (1-0.27705)}{0.2 \times (1-0.2)} =$$

$$0.3244.$$

The Gini index is calculated as follows:

$$\text{Gini} = 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra}$$

$$= 2 \times \left\{ \frac{2 \times [1 + 0.27705]}{4} - \frac{2 \times [0.27705 + 0]}{4} \right\} \times 0.3244$$

$$= 0.3244.$$

The scaled Brier score is calculated as follows:

doi:10.6342/NTU202504305

$$\text{scaled Brier} = 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra}$$

$$= 2 \times \left[\frac{2 \times (0.27179 + 0.2)}{4} - \frac{2 \times (0.01268 + 0.2)}{4}\right] \times 0.3244$$

$$= 0.0840.$$

Model VI

The predictiveness curve for the prediction model is defined by the following functions:

$$r = \begin{cases} 0.12821 & \text{when} \quad 0.00000 \leq F(r) < 0.72295, \\ 0.38732 & \text{when} \quad 0.72295 \leq F(r) \leq 1.00000. \end{cases}$$

The Pietra index is calculated as follows:

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi \times (1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi \times (1-\pi)} = \frac{(0.2-0.12821) \times (0.72295-0)}{0.2 \times (1-0.2)} = \frac{(0.38732-0.2) \times (1-0.72295)}{0.2 \times (1-0.2)} =$$

$$0.3244.$$

The Gini index is calculated as follows:

$$\text{Gini} = 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra}$$

$$= 2 \times \left\{\frac{2 \times [1 + 0.72295]}{4} - \frac{2 \times [0.72295 + 0]}{4}\right\} \times 0.3244$$

$$= 0.3244.$$

The scaled Brier score is calculated as follows:

$$\text{scaled Brier} = 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra}$$

$$= 2 \times \left[\frac{2 \times (0.38732 + 0.2)}{4} - \frac{2 \times (0.12821 + 0.2)}{4}\right] \times 0.3244$$

$$= 0.0840.$$

Model VII

doi:10.6342/NTU202504305

The predictiveness curve for the prediction model is defined by the following functions:

$$r = \begin{cases} 0.12606 & \text{when } 0.00000 \le F(r) < 0.70194, \\ 0.20000 & \text{when } 0.70194 \le F(r) < 0.93513, \\ 1.00000 & \text{when } 0.93513 \le F(r) \le 1.00000. \end{cases}$$

The Pietra index is calculated as follows:

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi \times (1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi \times (1-\pi)} = \frac{(0.2-0.12606) \times (0.70194-0)}{0.2 \times (1-0.2)} = \frac{(1-0.2) \times (1-0.93513)}{0.2 \times (1-0.2)} = 0.3244.$$

The Gini index is calculated as follows:

$$\begin{aligned} \text{Gini} &= 2 \times (x_{\text{AL}} - x_{\text{BL}}) \times \text{Pietra} \\ &= 2 \times \left\{ \frac{2 \times [1 + 0.93513]}{4} - \frac{2 \times [0.70194 + 0]}{4} \right\} \times 0.3244 \\ &= 0.4. \end{aligned}$$

The scaled Brier score is calculated as follows:

$$\begin{aligned} \text{scaled Brier} &= 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra} \\ &= 2 \times \left[\frac{2 \times (1 + 0.2)}{4} - \frac{2 \times (0.12606 + 0.2)}{4} \right] \times 0.3244 \\ &= 0.2835. \end{aligned}$$

Model VIII

The predictiveness curve for the prediction model is defined by the following functions:

$$r = \begin{cases} 0.00000 & \text{when } 0.00000 \le F(r) < 0.25949, \\ 0.20000 & \text{when } 0.25949 \le F(r) < 0.49268, \\ 0.30229 & \text{when } 0.49268 \le F(r) \le 1.00000. \end{cases}$$

The Pietra index is calculated as follows:

doi:10.6342/NTU202504305

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi\times(1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi\times(1-\pi)} = \frac{(0.2-0)\times(0.25949-0)}{0.2\times(1-0.2)} = \frac{(0.30229-0.2)\times(1-0.49268)}{0.2\times(1-0.2)} = 0.3244.$$

The Gini index is calculated as follows:

$$\begin{aligned}\text{Gini} &= 2\times(x_{\text{AL}} - x_{\text{BL}})\times\text{Pietra}\\ &= 2\times\left\{\frac{2\times[1+0.49268]}{4} - \frac{2\times[0.25949+0]}{4}\right\}\times 0.3244\\ &= 0.4.\end{aligned}$$

The scaled Brier score is calculated as follows:

$$\begin{aligned}\text{scaled Brier} &= 2\times(y_{\text{AL}} - y_{\text{BL}})\times\text{Pietra}\\ &= 2\times\left[\frac{2\times(0.30229+0.2)}{4} - \frac{2\times(0+0.2)}{4}\right]\times 0.3244\\ &= 0.0981.\end{aligned}$$

Model IX

The predictiveness curve for the prediction model is defined by the following functions:

$$r = \begin{cases} 0.06464 & \text{when } 0.00000 \le F(r) < 0.38340,\\ 0.20000 & \text{when } 0.38340 \le F(r) < 0.61659,\\ 0.33536 & \text{when } 0.61659 \le F(r) \le 1.00000.\end{cases}$$

The Pietra index is calculated as follows:

$$\text{Pietra} = \frac{\text{Area}_{\text{BL}}}{\pi\times(1-\pi)} = \frac{\text{Area}_{\text{AL}}}{\pi\times(1-\pi)} = \frac{(0.2-0.06464)\times(0.3834-0)}{0.2\times(1-0.2)} = \frac{(0.33536-0.2)\times(1-0.61659)}{0.2\times(1-0.2)} = 0.3244.$$

The Gini index is calculated as follows:

$$\text{Gini} = 2\times(x_{\text{AL}} - x_{\text{BL}})\times\text{Pietra}$$

doi:10.6342/NTU202504305

$$= 2 \times \left\{ \frac{2 \times [1 + 0.61659]}{4} - \frac{2 \times [0.3834 + 0]}{4} \right\} \times 0.3244$$

$$= 0.4.$$

The scaled Brier score is calculated as follows:

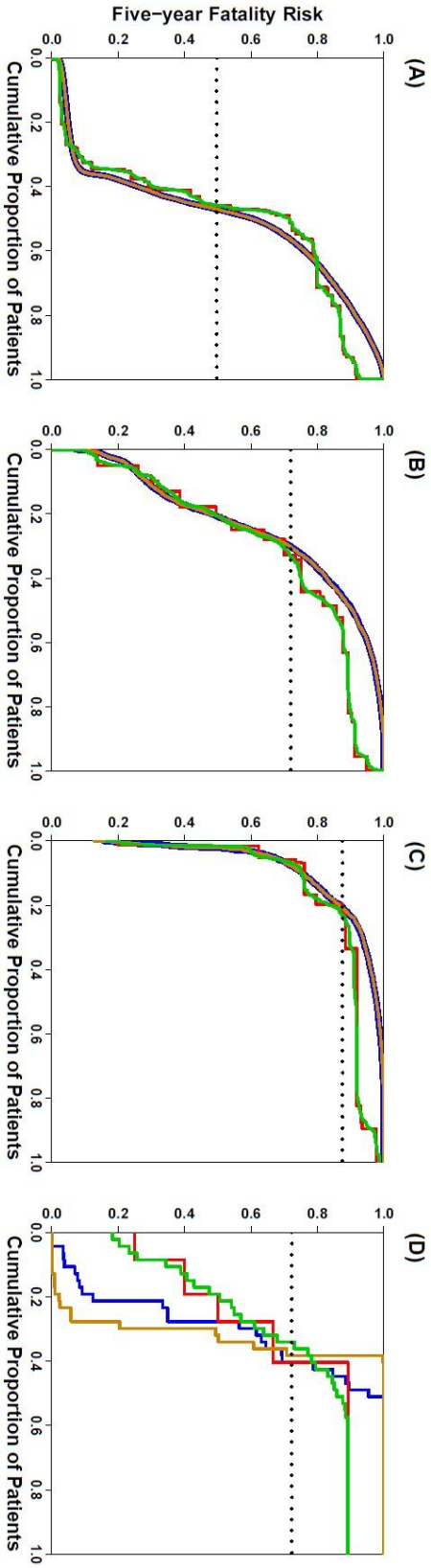
$$\text{scaled Brier} = 2 \times (y_{\text{AL}} - y_{\text{BL}}) \times \text{Pietra}$$

$$= 2 \times \left[\frac{2 \times (0.33536 + 0.2)}{4} - \frac{2 \times (0.06464 + 0.2)}{4} \right] \times 0.3244$$

$$= 0.0878.$$

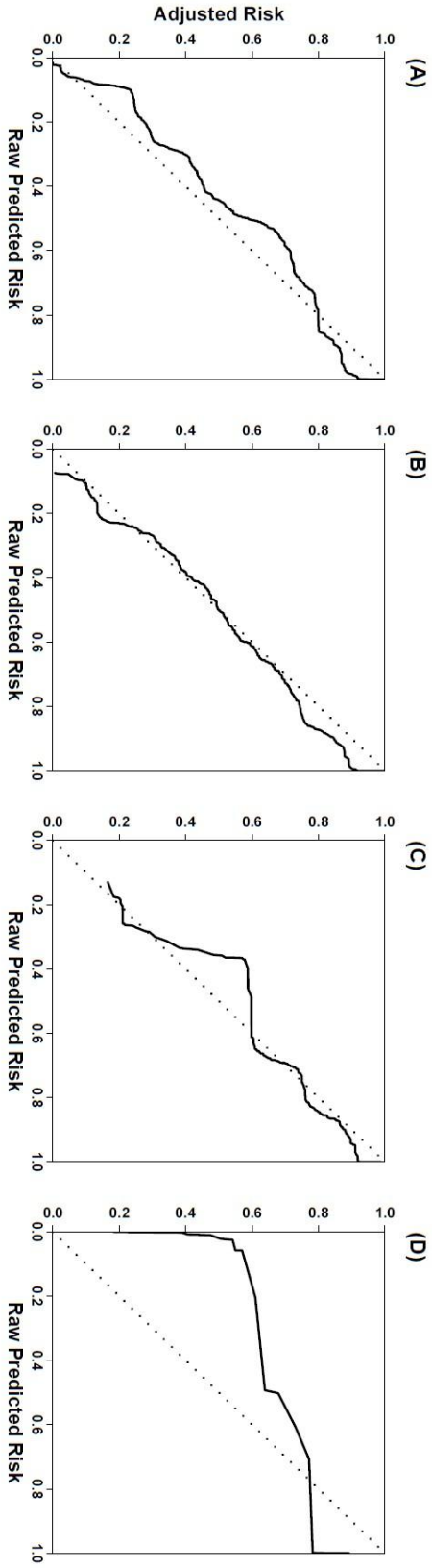
doi:10.6342/NTU202504305

Appendix 9.



doi:10.6342/NTU202504305

Appendix 10.



doi:10.6342/NTU202504305