OPEN ACCESS

Evidence-Based Medicine ARTICLE

The Innovation Medicine

# A global systematic analysis and statistical evaluation of prognostic prediction models for non-small cell lung cancer

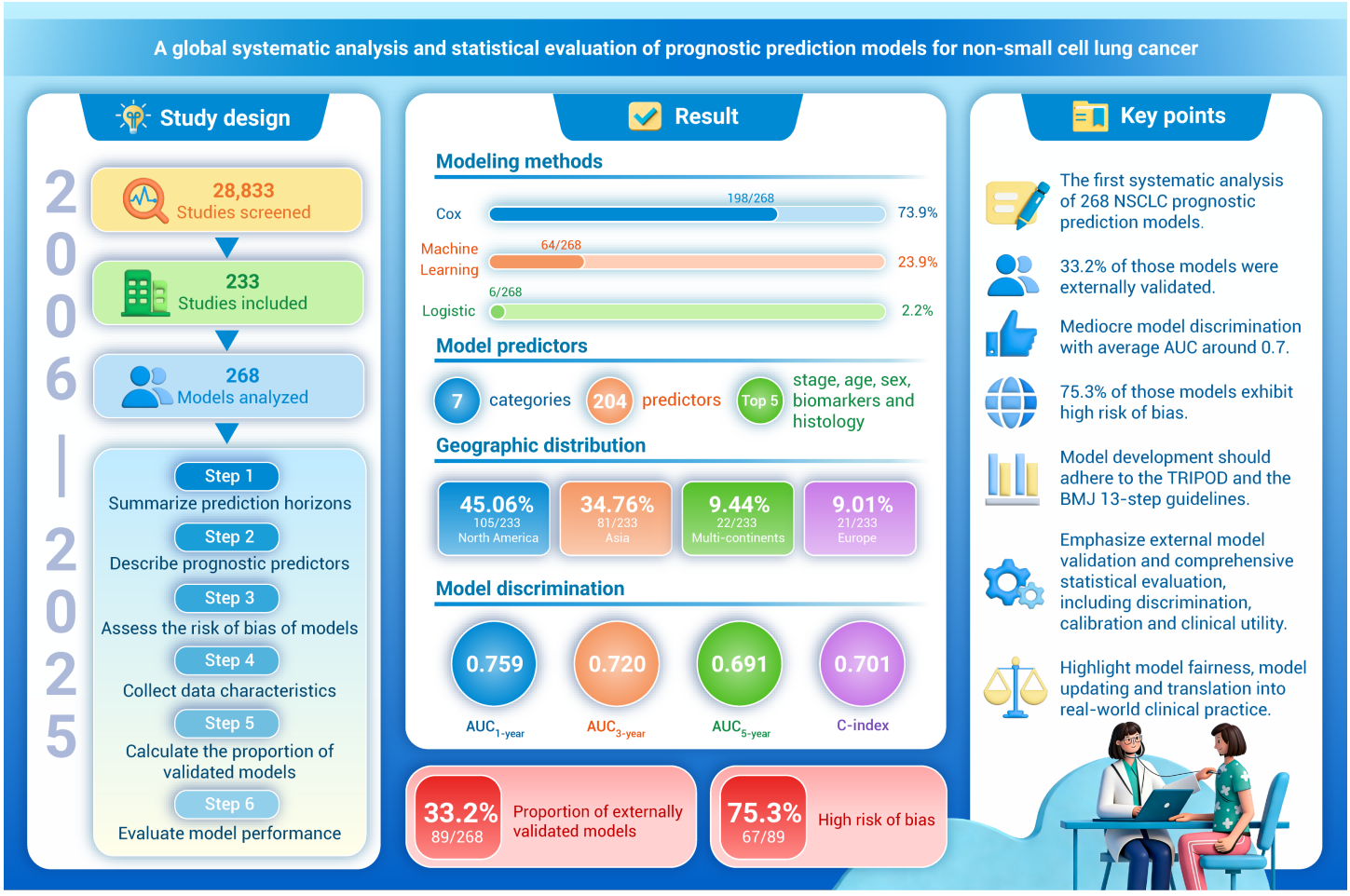
Xiaowen Xu,[1,24] Leyi Chen,[1,24] Maojie Xue,[1,24] Aoxuan Wang,[1,24] Qihan Fan,[1,24] Yuqing Yang,[1,24] Juanjuan Fan,[1] Tianhong Han,[1] Yichen Pan,[2] Shiyu Xiong,[2] Yuan Chen,[2] Zhujun Chen,[2] Zi Wang,[2] Jingtong Xu,[3] Zixing Zhang,[4] Yuchen Zhang,[5] Yu Chen Zhao,[6,7] Longyao Zhang,[1] Zhongwen Zhang,[1] Ning Xie,[1] Yanru Wang,[1] Xiang Wang,[1] Xuan Wang,[1] Manel Esteller,[8,9,10,11] Yi Li,[12] Rayjean J. Hung,[13] Christopher I. Amos,[14] David C. Christiani,[6,15,25] Feng Chen,[1,16,17,18,*] Yang Zhao,[1,16,17,18,*] Yongyue Wei,[19,20,*] and Ruyang Zhang[1,21,22,23,*]

*Correspondence: zhangruyang@njmu.edu.cn (R.Z.); ywei@pku.edu.cn (Y.W.); zhaoyang@njmu.edu.cn (Y.Z.); fengchen@njmu.edu.cn (F.C.)

Received: September 15, 2025; Accepted: January 16, 2026; Published Online: January 17, 2026; https://doi.org/10.59717/j.xinn-med.2026.100195



## GRAPHICAL ABSTRACT




## PUBLIC SUMMARY

- We performed the first global systematical analysis of 268 prognostic prediction models for NSCLC.
- Existing models exhibit mediocre discrimination with average AUC around 0.7.
- Only 33.2% of existing models were externally validated, and 75.3% of these exhibit high risk of bias.
- Adherence to TRIPOD and BMJ 13-step guidelines may improve model reliability and clinical translation.
- The community should highlight model fairness, model updating and translation into real-world clinical practice.

ARTICLE Evidence-Based Medicine

OPEN ACCESS

# A global systematic analysis and statistical evaluation of prognostic prediction models for non-small cell lung cancer

Xiaowen Xu,[1,24] Leyi Chen,[1,24] Maojie Xue,[1,24] Aoxuan Wang,[1,24] Qihan Fan,[1,24] Yuqing Yang,[1,24] Juanjuan Fan,[1] Tianhong Han,[1] Yichen Pan,[2] Shiyu Xiong,[2] Yuan Chen,[2] Zhujun Chen,[2] Zi Wang,[2] Jingtong Xu,[3] Zixing Zhang,[4] Yuchen Zhang,[5] Yu Chen Zhao,[6,7] Longyao Zhang,[1] Zhongwen Zhang,[1] Ning Xie,[1] Yanru Wang,[1] Xiang Wang,[1] Xuan Wang,[1] Manel Esteller,[8,9,10,11] Yi Li,[12] Rayjean J. Hung,[13] Christopher I. Amos,[14] David C. Christiani,[6,15,25] Feng Chen,[1,16,17,18,*] Yang Zhao,[1,16,17,18,*] Yongyue Wei,[19,20,*] and Ruyang Zhang[1,21,22,23,*]

[1]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China
[2]Department of Clinical Medicine, School of the first Clinical Medicine, Nanjing Medical University, Nanjing 211166, China
[3]Department of Oncology, The First Affiliated Hospital of Nanjing Medical University, Nanjing 210029, China
[4]Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China
[5]Department of Occupational Medicine and Environmental Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China
[6]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston MA 02115, USA
[7]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston MA 02115, USA
[8]Josep Carreras Leukaemia Research Institute, Barcelona, Catalonia 08021, Spain
[9]Centro de Investigacion Biomedica en Red Cancer, Madrid 28029, Spain
[10]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Catalonia 08010, Spain
[11]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona, Barcelona, Catalonia 08007, Spain
[12]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
[13]Lunenfeld-Tanenbaum Research Institute, Sinai Health, and University of Toronto, Toronto, ON M5G 1X5, Canada
[14]The Institute for Clinical and Translational Research, Baylor College of Medicine, Houston TX 77030, USA
[15]Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston MA 02114, USA
[16]China International Cooperation Center (CICC) for Environment and Human Health, Nanjing Medical University, Nanjing 211166, China
[17]Jiangsu Key Lab of Cancer Biomarkers, Prevention, and Treatment, Cancer Center, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China
[18]The Key Laboratory of Modern Toxicology of Ministry of Education, Nanjing Medical University, Nanjing 211166, China
[19]Center for Public Health and Epidemic Preparedness & Response, Peking University, Beijing 100191, China
[20]Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China
[21]Information Center, The Second People's Hospital of Changzhou, The Third Affiliated Hospital of Nanjing Medical University, Nanjing 213164, China
[22]Changzhou Medical Center, Nanjing Medical University, Changzhou 213164, China
[23]Laboratory for Digital Intelligence & Health Governance, Nanjing Medical University, Nanjing 211166, China
[24]These authors contributed equally
[25]The senior author who supervised the work
*Correspondence: zhangruyang@njmu.edu.cn (R.Z.); ywei@pku.edu.cn (Y.W.); zhaoyang@njmu.edu.cn (Y.Z.); fengchen@njmu.edu.cn (F.C.)
Received: September 15, 2025; Accepted: January 16, 2026; Published Online: January 17, 2026; https://doi.org/10.59717/j.xinn-med.2026.100195

Citation: Xu X., Chen L., Xue M., et al. (2026). A global systematic analysis and statistical evaluation of prognostic prediction models for non-small cell lung cancer. *The Innovation Medicine* **4**:100195.

Prognostic prediction models can aid clinical decision-making for high-risk non-small cell lung cancer (NSCLC) patients. We conducted a systematic search across multiple databases and evaluated eligible studies using PRISMA and PROBAST checklists. Of 28,833 references screened, 233 studies describing 268 models were included. Among them, 89 underwent external validation; 67 (75.28%) were classified with high risk for bias. Most models were developed in North America (45.06%). The median area under the receiver operating characteristic curve (AUC) for 1-, 3-, and 5-year predictions were 0.759, 0.720, and 0.691, respectively. The most common predictors were age (54.85%) and stage (58.21%). And, 47.01% of models were sex-specific. Recently, easily accessible radiomic features have been increasingly used in statistical modeling compared to molecular omics. Models integrating radiomic and clinical features showed potential for improved performance (1-, 3-, and 5-year AUCs: 0.931, 0.985, 0.942). Model discrimination varied across different studies, and there is a lack of model calibration and clinical utility. This review highlights advances and limitations in NSCLC prognostic models. To enhance model reliability and generalizability, it remains crucial to adhere to the TRIPOD guideline to perform prediction model study, and to emphasize comprehensive model validation and bias-reduction strategies. Following a step-by-step guide to develop and validate clinical prediction models by Efthimiou, et al. (BMJ, 2024), along with a multifaceted, multidisciplinary and multi-regional approach, is the key way to facilitate the development and clinical application of qualified models.

## INTRODUCTION

Lung cancer is one of the most prevalent cancers worldwide and remains the leading cause of cancer-related deaths, with an estimated 1.9 million annual deaths.[1] Over 85% of all diagnosed lung cancer cases are non-small cell lung cancer (NSCLC).[2] Improving survival and prognosis for NSCLC patients remains a substantial challenge globally. Categorizing mortality risk within lung cancer cohorts using prognostic prediction models is crucial for refining treatment strategies and advancing precision medicine.[3] Prognostic prediction models, which statistically relate multiple variables to mortality in individuals, have been the focus of extensive research over the past two decades, particularly in NSCLC patients. However, the development and validation of these models face substantial challenges, including methodological heterogeneity and concerns regarding their reliability and clinical applicability.[4] These models exhibit significant heterogeneity across various aspects, including predictors, model development methodologies, and other key factors. Despite the wealth of prognostic prediction models for NSCLC, there remains no systematic analysis and statistical evaluation of all existing models.

Currently, few reviews have partially evaluated prognostic prediction models for NSCLC. Among them, majority remains limited in the number of studies screened or have focused on outdated literature. For instance, Wang et al. focused primarily on the methodology of model development, emphasizing predictor selection and bias control. However, it is limited in scope, analyzing just nine studies with external validation.[5] Another review by Brundage et al. evaluated prognostic factors for NSCLC patients from an epidemiological perspective, restricted in studies from 1990 to 2000.[6] Though it thoroughly analyzed prognostic factors, including tumor, host, surgical, and

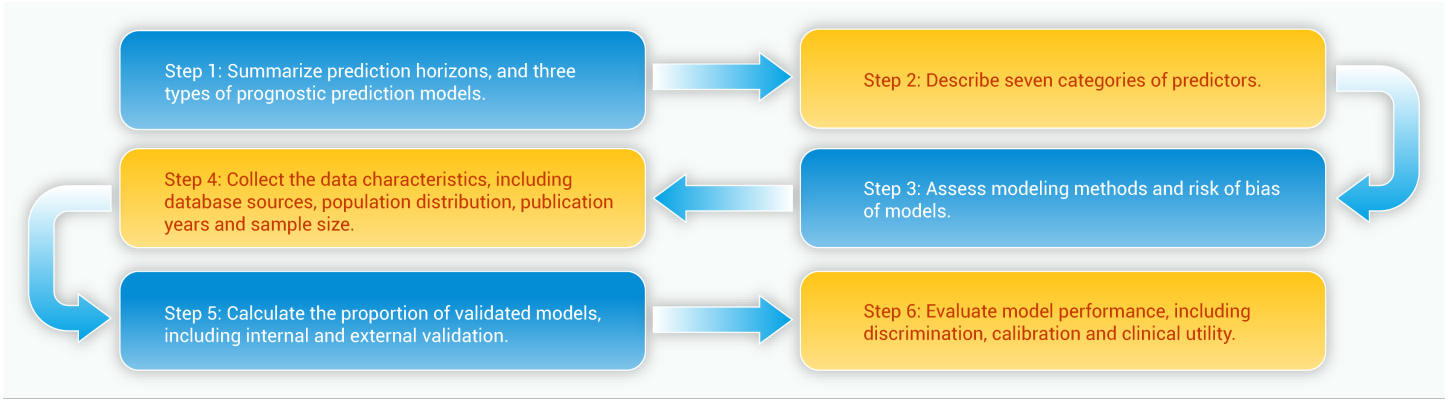



**Figure 1. The diagram outlines the structure of the systematic analysis and statistical evaluation.**

environmental factors, this review is somewhat outdated and narrow in scope, focusing solely on prognostic factors. Therefore, to the best of our knowledge, there exists no up-to-date systematic analysis and review that has evaluated the predictive factors, statistical methods, model performance, external validations, etc. in a comprehensive manner. This gap is critical, since comprehensive statistical evaluation of current models is essential to identify and address existing limitations, inform future studies, and ultimately guide the development of clinically applicable models.[7]

Our systematic analysis aims to conduct a thorough assessment of the basic characteristics of all existing NSCLC prognostic models. The structure of our study follows the latest step-by-step guide for developing clinical prediction models (BMJ 13-step guide),[8] evaluating six key aspects (Figure 1). This analysis highlights advances, limitations and challenges in the field and offers insights and recommendations to enhance future study of prognostic prediction model of NSCLC.

## MATERIALS AND METHODS

We adhered to the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies (CHARMS) in systematic review, and our study has also been registered in the International Prospective Register of Systematic Reviews (PROSPERO ID: CRD42024534850). All updates to the review will be documented in the registry. Meanwhile, our study was designed by the guide to systematic review and meta-analysis of prediction model performance,[9] with its results reported following Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) (see Table S1).[10] The evaluation of risk of bias (ROB) was critically appraised by Prediction model Risk Of Bias Assessment Tool (PROBAST).[11]

### Literature search

According to PRISMA, we systematically searched prognostic prediction models of NSCLC in five databases, including PubMed, Embase, MEDLINE, Web of Science, and The Cochrane Library, using the following search string: "((non-small cell lung cancer) or (NSCLC)) and ((prognosis) or (survival)) and ((prediction) or (AUC) or (C-index))". Literature was restricted to studies published before 5 April 2025. The full search strategy is available in Supplementary Material.

### Eligibility criteria

All studies of the development, validation and updating of NSCLC prognostic models were included: (i) studies designed to develop the new prognostic prediction models but without external validation in original studies; (ii) studies designed to develop and externally validate the same prediction models; (iii) studies designed to develop the new prognostic prediction models and compare with the previously developed comparable models; (iv) studies designed to update the previously developed models. Duplicate studies retrieved from different database were excluded, and studies meeting the following exclusion criteria were removed: (i) conference abstract, commentary, editorial or letters; (ii) clinical trials; (iii) non-prognostic prediction models; (iv) non time-to-event overall survival (OS) time outcomes; and (v) no index of model evaluation, e.g., area under the receiver operating characteristic curve (AUC) or concordance index (C-index).

Different models developed from the same data resource, e.g., The Cancer Genome Atlas (TCGA) or Sun Yat-sen University Cancer Center, in different studies were considered statistically independent of each other in the primary analysis, due to their difference in modeling strategy, variable selection, patient enrolment period, sample quality control procedure, etc. To evaluate their impacts on the pooled model discrimination, we also conducted a sensitivity analysis as the secondary analysis, by combining the model-level AUC/C-index values using their median estimates into the source-level AUC/C-index.

### Quality control and assurance

To ensure the accuracy and consistency of study selection and data extraction, we implemented a rigorous multi-step quality control procedure. Before formal screening, all involved reviewers participated in a dedicated training session to standardize the understanding and application of the inclusion and exclusion criteria. Two researchers independently conducted the literature data retrieval and extraction, and other ten researchers assessed the information, and six researchers conducted literature screening and review. Consensus was reached on disagreements that arose during the study. We extracted the first author, year of publication, study type, source of datasets and model information (e.g., sample size, model development methodology, model performance, validation status, predictors) from all included studies. Any discrepancies regarding the inclusion or exclusion of a study were first resolved through discussion between the two initial reviewers. If a consensus could not be reached, the disagreement was adjudicated by a third senior researcher whose decision was final.

For the critical phase of data extraction from the included studies, a similar dual-independent process was employed. The extracted data was then verified for consistency and accuracy by a second reviewer. Any inconsistencies were flagged and resolved through discussion, referring back to the original publication.

## RESULTS

Through our systematic search across five literature databases, we retrieved 28,833 studies. Among these, 8,673 duplicates were removed. After evaluating 20,160 studies by titles and abstracts, 873 studies were selected for full-text screening. This process led to 848 studies meeting the eligibility criteria. According to the exclusion criteria (Figure 2), 233 studies remained in this study, detailing 268 models (Table S2).

### Objectives of NSCLC prognostic models

***Outcomes and predictive horizon.*** Among the 233 NSCLC studies involving 268 models, the primary endpoint was OS time, defined from diagnosis to death from any cause or last follow-up. A substantial proportion of models (85/268, 31.72%) focused on predicting OS at 1-, 3-, or 5-year. Especially noteworthy is Xie et al.'s model, which projected OS from 1 to 10 years, offering the broadest prediction range.[12] Additionally, The C-index was the most frequently used metric for assessing model discrimination, applied in both models with (62/89, 69.66%) and without (121/179, 67.60%) external validation (Table S4).

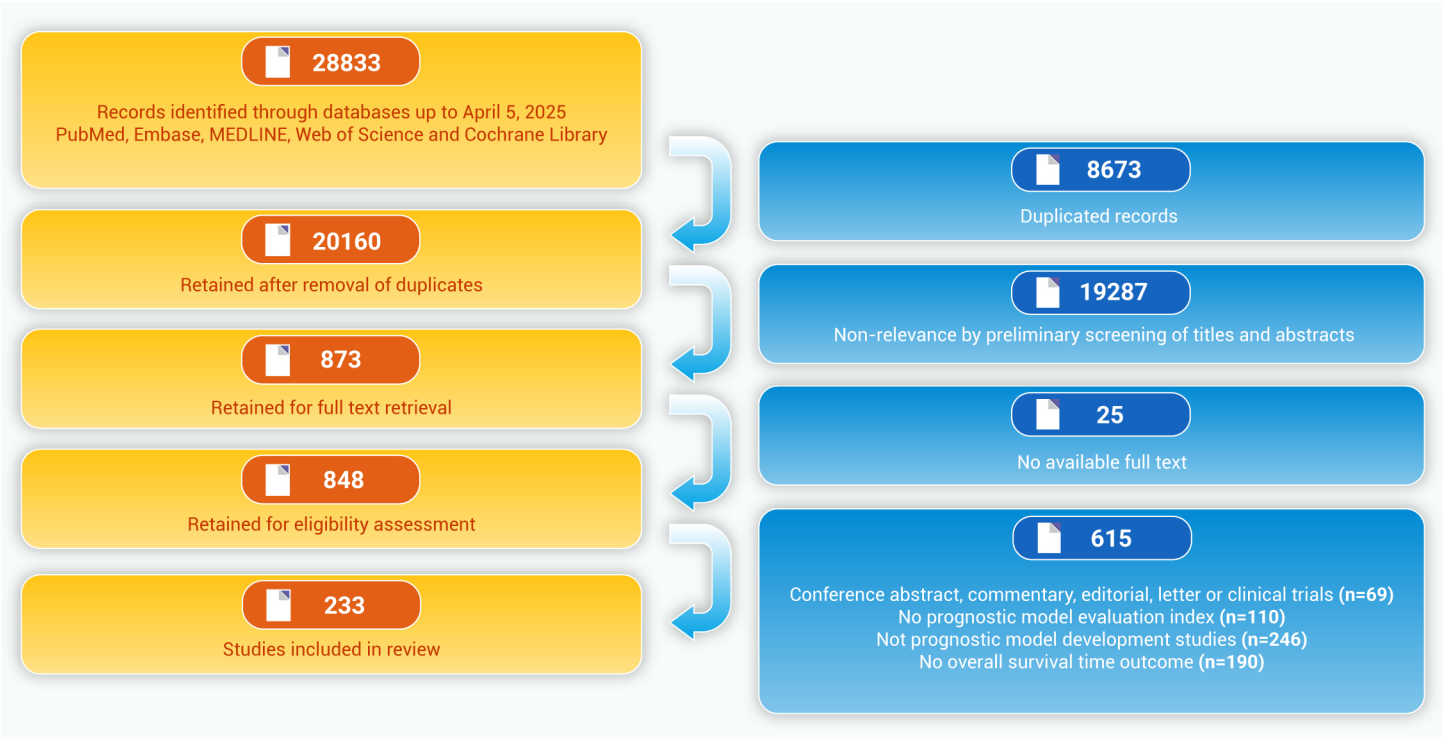



**Figure 2. PRISMA flow diagram of our study.**

***Model types.*** Models were categorized into three types: (i) newly developed without validation (71/268, 26.49%), (ii) newly developed with internal or external validation (152/268, 56.72%), and (iii) comparing or improving existing models (45/268, 16.79%). Importantly, among the 89 models with external validation, 66 (74.16%) developed new models with validation.

## Predictors of NSCLC prognostic models

***Category of predictors.*** The number of predictors in these prognostic models ranged from 1 to 13. A total of 204 predictors were categorized into seven groups: (1) demographics and family history; (2) laboratory measurements; (3) clinical information; (4) biomarkers; (5) surgical and pathological characteristics (SPC); (6) therapy and (7) lifestyle (Figure 3). The five most frequently used factors were stage (156/268, 58.21%), age (147/268, 54.85%), sex (126/268, 47.01%), biomarkers (112/268, 41.79%), and histology (67/268, 25%). Predictors such as gene expression, tumor size and smoking were prevalent in non-validated models (Figure S1).

Our study analyzed trends in predictors over time (Figure 4 and Table S3). In earlier prognostic models, especially before 2018, predictors were primarily demographics and family history, clinical information, and SPC. Over the past five years, the inclusion of predictors has broadened, now frequently incorporating biomarkers, therapy, and lifestyle factors, reflecting a shift towards less parsimonious models that account for treatment and biological mechanisms. Overall, laboratory measurements have historically been infrequent, and their proportion in models remain low.

***Demographic, family historical and clinical predictors.*** Most predictors in NSCLC prognostic models were demographic factors, which were key modifiable predictors in lung cancer.[5,13] In externally validated models, 56.18% (50/89) incorporated age as a predictor. Only a few models included insurance status and cancer history. Clinical predictors such as stage, histological type, and grade, were observed in 57.30%, 17.98%, and 13.48% of externally validated models, respectively. Wang et al. developed a prognostic model using representative clinical predictors like histological type and metastases.[14] This model developed and externally validated showed the best performance, with $AUC_{1\text{-year}}$ of 0.830 and a C-index of 0.717. Exploring combinations of clinical predictors strongly associated with lung cancer prognosis can enhance the performance of prognostic models.

***Molecular biomarkers.*** Apart from traditional prognostic predictors, recent advances in high-throughput technologies have introduced molecular biomarkers into lung cancer prognosis.[15-18] These biomarkers include DNA sequences, DNA methylation, gene expression, and protein abundance. We assessed 112 models developed using biomarkers, with 51 (45.54%) being externally validated. The inclusion of biomarkers and clinical predictors has shown potential in improving the prognostic accuracy of NSCLC models, as evidenced by their performance in various validation cohorts, e.g., Xie's model added gene expression has satisfactory AUC of 0.822 and 0.755 for 1 and 10-year, respectively, in an external validation.[12] DNA methylation also enhance prognostic prediction. For example, Zhang et al. demonstrated in a multi-omics study that incorporating epigenetic and transcriptional biomarkers with either main effects or gene-gene interactions substantially improved model performance, with $AUC_{3\text{-year}}$ and $AUC_{5\text{-year}}$ of 0.880 and 0.890, respectively.[19] Additionally, we summarized that 37 studies totally used 214 transcriptional predictors, with 11 genes (*EGFR*, *SLC2A1*, *TPX2*, *RNASE7*, *MBNL2*, *LINC01116*, *SNHG30*, *LGR4*, *BTK*, *CXCL5* and *ALCAM*) used more than twice across all studies (Figure S2).

***Imaging features.*** We identified 34 studies using radiomic features as predictors, categorized into four types based on different approaches and techniques employed in model construction: (i) radiomics-only models (10/34, 29.41%), (ii) models generating risk scores from computed tomography (CT) radiomics (2/34, 5.88%), (iii) integrated models combining radiomics with clinical factors (18/34, 52.95%), (iv) and probability-weighted models that enhance prediction accuracy by integrating radiomic and clinical features (4/34, 11.76%) (Table S2). Among the four categories of radiomics-based prognostic models, the third type of models demonstrated the highest performance in training datasets, with a median C-index of 0.742 (range: 0.6 to 0.9). In contrast, relying solely on radiomic features, exhibited the best performance in internally validated models with a median C-index of 0.82 (range: 0.604 to 0.93). Besides, the fourth type of models has not undergone external validation, though it demonstrated strong performance with 1-, 3-, and 5-year AUCs of 0.931, 0.985, and 0.942, respectively.[20]

***Publication years.*** Figure 5A illustrates the rapid growth in prognostic prediction model publications, particularly from 2019 to 2025, which accounts for 80.69% (188/233) of the total number of studies. This growth includes notable peaks in 2024. Additionally, recent years have shown a stronger emphasis on external validation, contributing to the clear upward trend in the number of publications and the validation studies for these models.

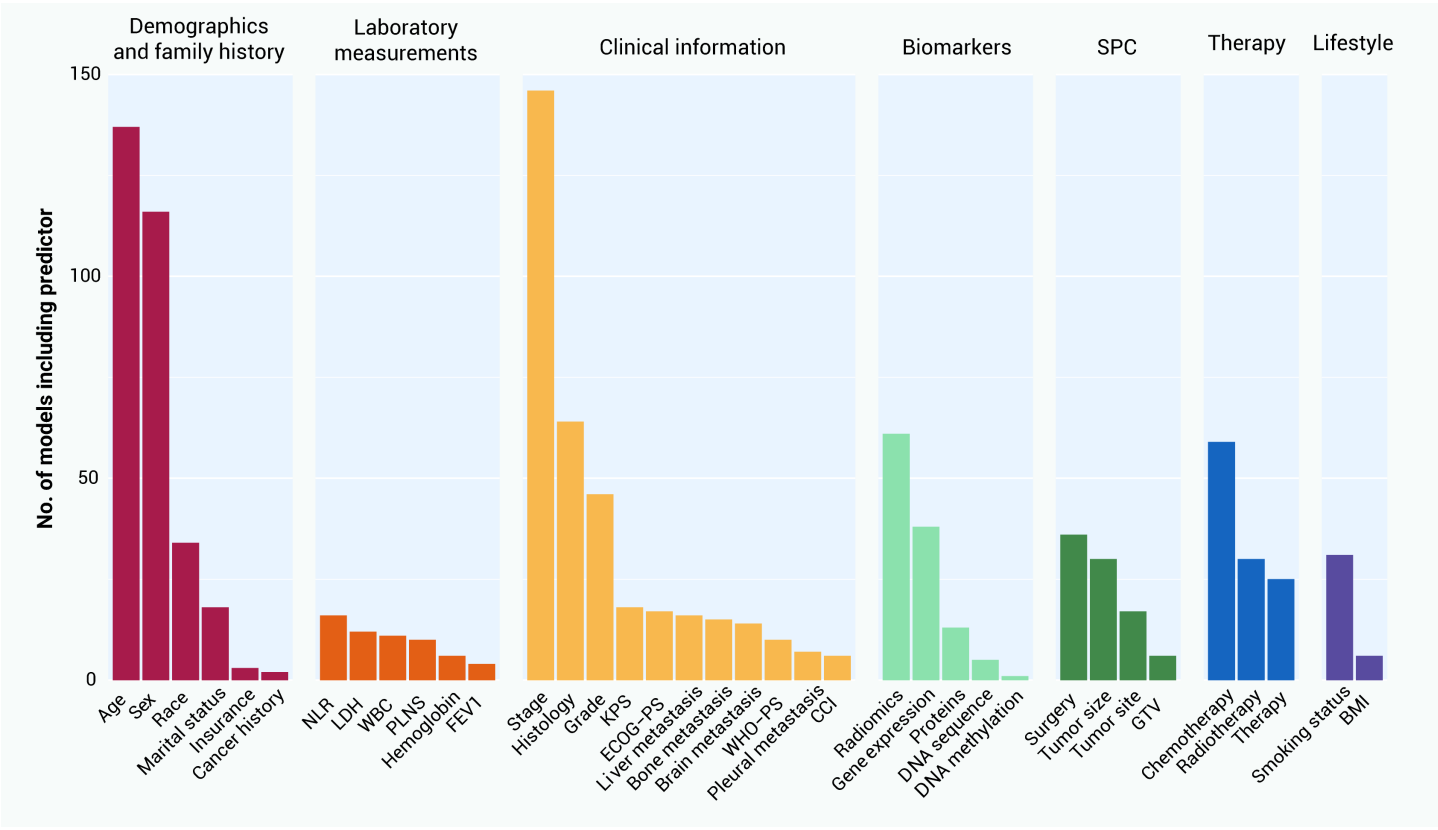



**Figure 3. The distribution of predictors in 268 prognostic prediction models of NSCLC** SPC, surgical and pathological characteristics; PLNS, positive lymph nodes ratio; NLR, neutrophil-to-lymphocyte ratio; LDH, lactate dehydrogenase; WBC, white blood cell; FEV1, forced expiratory volume in one second; WHO-PS, WHO performance status; ECOG-PS, eastern cooperative oncology group performance status; KPS, Karnofsky performance status; CCI, Charlson comorbidity index; GTV, gross tumour volume; BMI, body mass index.

## Statistical methods and ROB of models

***Statistical methods.*** Among the 268 models, the majority were developed using three primary techniques: Cox proportional hazards model (198/268, 73.88%), machine learning methods (64/268, 23.88%, e.g., random forest, convolutional neural network), and logistic regression (6/268, 2.24%), with the Cox model being the most prevalent (Figure 5B). This trend persisted regardless of external validation status. Over time, there was a noticeable increase in the use of machine learning for prognostic models. Among these, convolutional neural network was the most frequently used method (9/64, 14.06%). A detailed analysis of the development of modeling methods is in Figure S3.

***Evaluation for ROB.*** Based on the assessment of the 89 externally validated models using PROBAST checklist, 67 (75.28%) were classified as having a 'high ROB', 20 (22.47%) were assessed as 'low ROB', and the remaining two (2.25%) models was deemed 'unclear' as shown in Figure 6A. The primary source of ROB stemmed from the 'analysis' domain, with 51 (57.30%) models identified as 'high ROB' in this area. The 'participants' domain was another significant contributor to ROB, by which 28 (31.46%) models rated as 'high ROB' (Table S5). To further examine the impact of ROB on model reliability, we additionally investigated the relationship between overall ROB and model discrimination using Spearman correlation analysis. As shown in Figure S4, high ROB was correlated with high model discrimination in the training set (Spearman $\rho$ = 0.31, $P$ = 0.006), but correlated with low model discrimination in validation set (Spearman $\rho$ = -0.47, $P$ = 0.040), indicating ROB controlling benefits model generalization ability.

## Sample characteristics

***Source of database and participants.*** Out of the 233 studies analyzed, 140 used private data sources, while 93 relied on public data sources, most originated from Surveillance, Epidemiology, and End Results Program (SEER) and The Cancer Genome Atlas (TCGA). Figure 5C shows that most of (105/233, 45.06%) models utilize North American sources (primarily SEER from the US). In external validation studies, most of models were developed using data from North America (56/89, 62.92%), Europe (12/89, 13.48%) and Asia (16/89, 17.98%) (Table 1). Additionally, 9.44% (22/233) of studies utilized multi-continental databases (all participants came from multiple continents, regardless of training or validation sets), commonly combining North American public data, with external validation dataset from Chinese private hospitals. Notably, several studies, including those from Sun Yat-sen University Cancer Center in Guangzhou, were used multiple times for model development, yielding 12 models in 12 studies.[21-30]

***Sample size.*** For the 233 studies, the total number of participants per study, ranged from 51 to 98,640, with a median of 590. Sample sizes for training and testing sets varied widely, with medians of 482 and 210 respectively. In studies without a validation set (60/233, 25.75%), the median sample size was 388. Studies with validations (173/233, 74.25%), especially those with multiple validation sets, generally had larger total sample sizes (median 1,211.5) (Table S6).

## Model validation

Of the 268 models, 42.16% (113/268) underwent internal validation, primarily using random data splitting, bootstrapping, or cross-validation. A total of 89 models (89/268, 33.21%) were externally validated. There are three studies have the maximum number (four) of validation.[31-33] Among the 89 externally validated models, the most common scenario (12/89, 13.48%) entailed utilizing the SEER database with validations performed on datasets from Chinese hospitals.

## Model performance of externally validated models

Model performance was evaluated in terms of discrimination using the 1-, 3-, and 5-year AUCs and the C-index in models validated by external datasets. Only 8.99% (8/89) models reported numerical data on model calibration, while 15.73% (14/89) models included decision curve analysis (DCA) (Table S7). However, only one provided exact values of E:O ratio or net benefit. The overall performance of external validation models, as indicated by AUC and C-index values, exhibited a broad distribution across different time points, with both metrics consistently exceeding 0.7. A slight decline in AUC

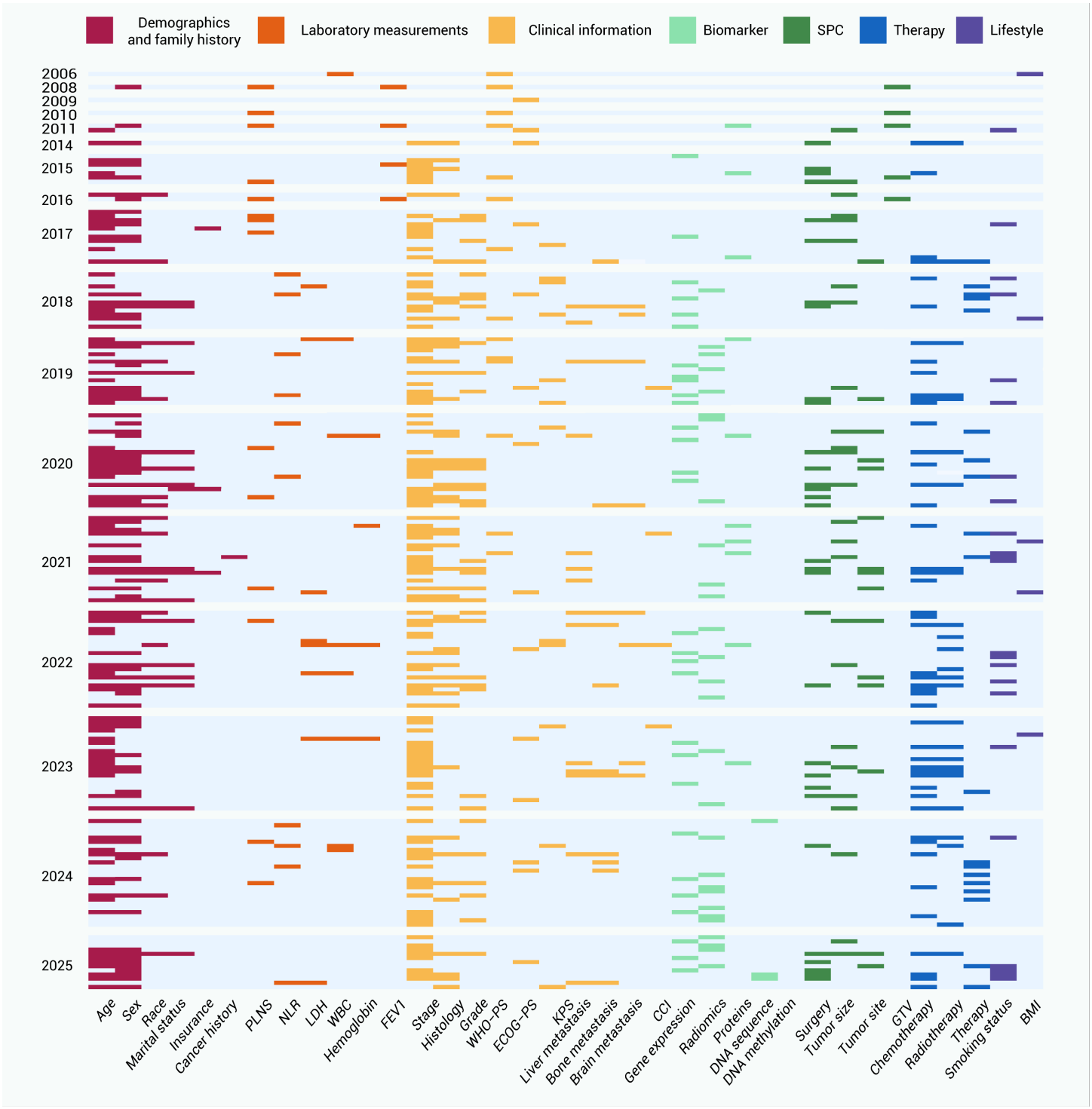



**Figure 4. Heatmap of predictors in prognostic prediction models of NSCLC over time** SPC, surgical and pathological characteristics; PLNS, positive lymph nodes ratio; NLR, neutrophil-to-lymphocyte ratio; LDH, lactate dehydrogenase; WBC, white blood cell; FEV1, forced expiratory volume in one second; WHO-PS, WHO performance status; ECOG-PS, eastern cooperative oncology group performance status; KPS, Karnofsky performance status; CCI, Charlson comorbidity index; GTV, gross tumour volume; BMI, body mass index.

values over time was observed, while C-index values remained relatively low with a wide distribution (Figure 6B). The result of sensitivity analysis for source-level model performance (Figure S5), remained highly consistent with that in the primary analysis, indicating that data overlap had minimal impact on the evaluation of average model discrimination. Only one model focused on long-term 10-year prognosis prediction, with $AUC_{10\text{-year}}$ of 0.776 in the training set and 0.755 in the external validation set.

Models developed and validated across different continents exhibited variations in performance (Figure S5). Overall, models from North America and Europe showed more stable performance in both training and validation datasets, particularly in terms of 1-year AUC and C-index, with a median validation C-index of 0.659 (95% CI: 0.573-0.862, derived via bootstrapping) for North America and 0.670 (95% CI: 0.590-0.750) for Europe, respectively (Table S8). In contrast, models based on data from Asia and multiple continents displayed fluctuations in C-index. Across both training and validation datasets, the 1-year AUC was generally higher than the 3-year and 5-year AUCs, indicating better short-term prediction accuracy. However, model performance in the validation sets was slightly lower than that in the training sets, suggesting mild overfitting during model training and reflecting heterogeneity in model performance across different populations (Figure 6B).

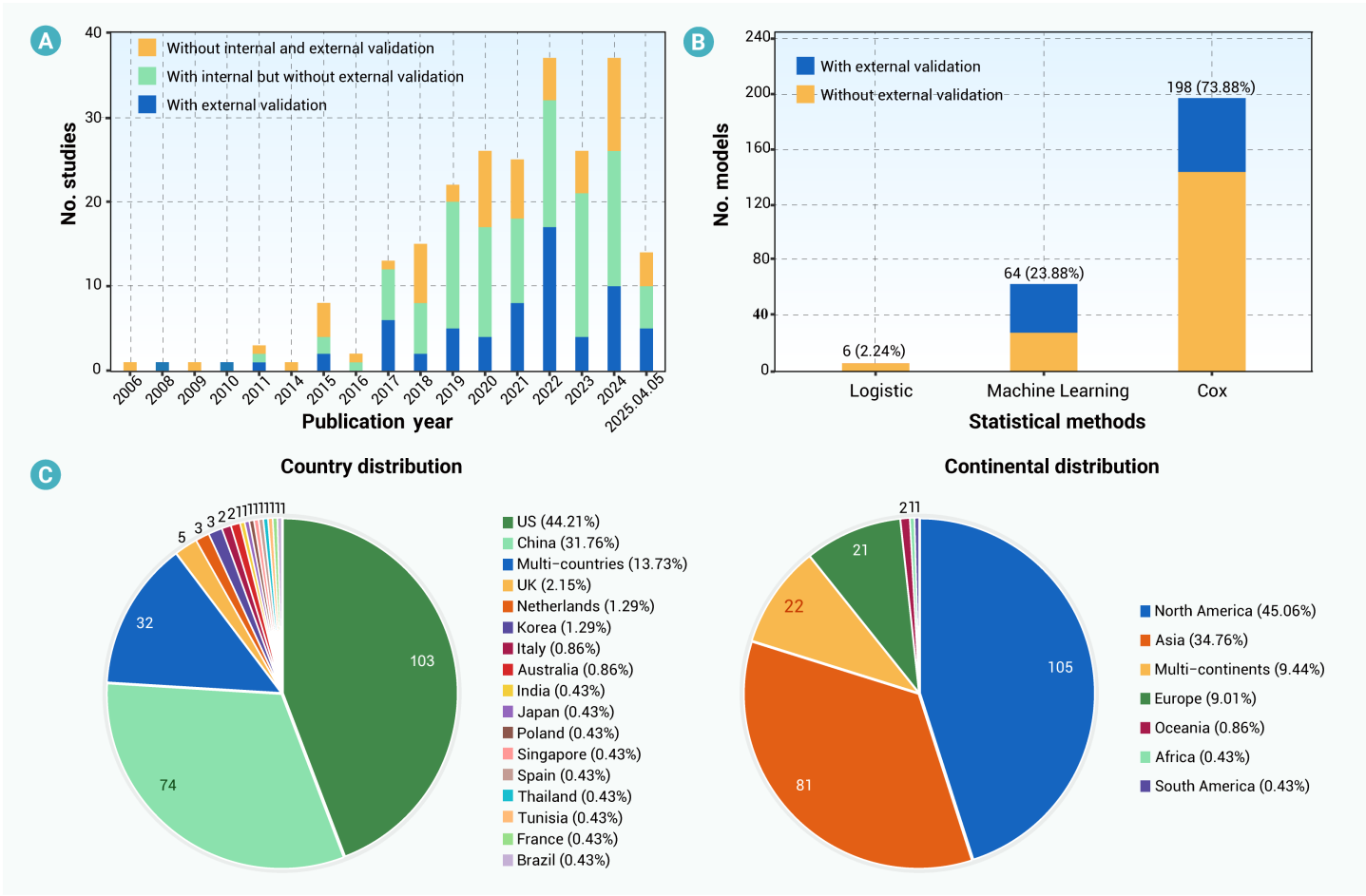



**Figure 5. Characteristics of prognostic prediction models of NSCLC** (A) The bar chart of validated and unvalidated models per year in 268 included models. (B) The bar chart of modeling methods in 268 included models. (C) The pie graph of geographical distribution. US, The United States of America; UK, The United Kingdom of Great Britain and Northern Ireland. Multi-countries or multi-continents, if the training and testing populations originate from multiple countries or continents, the model is classified as multi-countries or multi-continents.

Models involving data from multiple continents exhibited significant C-index variation in validation sets, suggesting that cross-regional model validation may be less consistent.

## DISCUSSION

This review provides the most comprehensive examination of prognostic models for NSCLC to date. Following PROBAST and the BMJ 13-step guidelines, we analyzed 268 models from 233 studies, highlighting key aspects of model development, predictive factors, NSCLC outcome prediction, and external validation status, etc. Our findings offer a valuable reference for clinicians, public health professionals, and policymakers.

### Geographical disparities

There are notable regional disparities in the development and validation of NSCLC prognostic models, with most of models originating from North America, and Asia. Model development regions, such as Africa and South America, are insufficiently represented. Although no review has yet emphasized geographical disparities in NSCLC risk prediction and prognostic models, similar phenomenon have been noted in a review of cardiovascular disease (CVD) prediction models.[34] Additionally, the overrepresentation of European ancestry in large biobanks complicates the development and evaluation of risk assessment tools for underrepresented populations,[35] particularly in low- and middle-income countries where lung cancer imposes a significant healthcare burden.[36] Therefore, considering the impact of racial heterogeneity on prediction models, it is essential to develop locally applicable NSCLC prognostic models for underrepresented regions. Transfer learning is effectively used in text recognition and image classification. The core concept of transfer learning is to leverage knowledge from models trained on data-rich regions (source domains) and adapt them to underrepresented populations (target domains),[37] which could leverage large public databases to build predictive models for regions with limited local data.[38] Taking the study by Theodoris et al. as an example, they developed the pre-trained deep learning model, Geneformer, using transfer learning, which learns gene network patterns from large-scale single-cell transcriptomic data and can be fine-tuned on limited patient-specific data to identify candidate therapeutic targets in diseases such as cardiomyopathy.[39]

### Modeling methods

Our research indicates that the Cox proportional hazards model is the most commonly used method for constructing prognostic models in existing studies. This model is appropriate and provides high interpretability for low-dimensional data. However, it is poorly suited for high-dimensional data (e.g., transcriptomic or radiomic data), where the number of predictors ($p$) far exceeds the sample size ($n$). The conventional Cox model is not designed for this "$p >> n$" (number of predictors greater than that of samples) scenario, inevitably leading to overfitting, spurious findings, and subsequent reduced accuracy in model validation,[40] though it can be solved by cross-validation or elasticated net. Besides, it is not convenient to include predictors with complex effects (e.g., non-linear effect and high-order interaction effect). These limitations have spurred the adoption of machine learning (ML) methods (e.g., XGBoost), which are better equipped to capture complex signals in high-dimensional data. However, ML models also introduce their own challenges, primarily centered on overfitting and underfitting. The frequent and substantial performance drop observed in the external validation of many ML-based models strongly indicates overfitting, often stemming from inadequate feature selection and hyperparameter tuning processes, that are

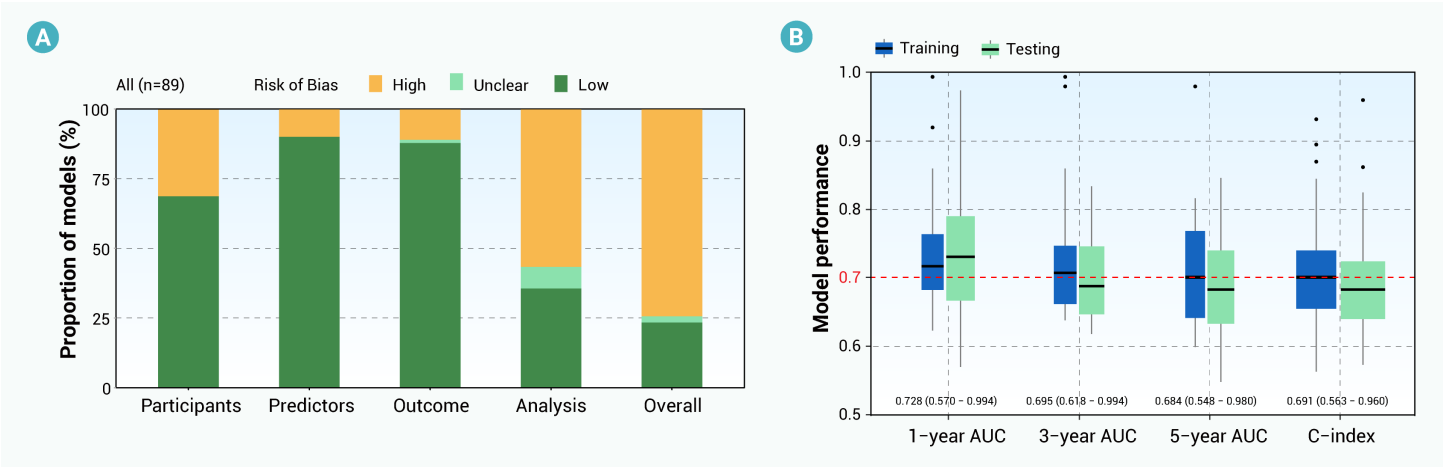



**Figure 6. Risk of bias assessment and predictive performance of externally validated NSCLC prognostic models** (A) Risk of bias evaluated with PROBAST (Prediction model Risk Of Bias Assessment Tool). The figure represents the percentage of studies scoring a low (green), high (red), or unclear (yellow) risk of bias for each of the four risk of bias assessment domains (Participants, Predictors, Outcome and Analysis) and the overall risk of bias. (B) Model performance in the training and validation sets for externally validated models.

frequently under-reported, compromising reproducibility and reliability.

To address these issues, future work must prioritize methodological rigor tailored to the data complexity. For high-dimensional data, employing robust feature selection methods (e.g., Boruta) is non-negotiable to reduce dimensionality. Furthermore, the use of nested cross-validation is essential for providing unbiased performance estimates and for proper hyperparameter optimization. For scenarios where traditional Cox models are applied to high-dimensional or extremely imbalanced data (e.g., extremely low event rate), advanced methods like the Firth correction[41] should be adopted to mitigate inflated type I error.[42–44] Finally, utilizing explainable techniques (e.g., Shapley value) is crucial to bridge the gap between the complexity of high-performance models and the need of clinical interpretability.[45]

## Development and validation

In developing prediction models, it is crucial to consider both the intended application and the quality of their development and validation. Our review indicates that 86.94% (223/268) of NSCLC prognostic models focused on developing new models, with a substantial lack of validation studies, particularly external validation. Several factors may explain this gap. Many models were built using institution-specific or geographically restricted datasets, which limited access to independent cohorts suitable for external evaluation. In addition, a considerable proportion of studies incorporated predictors (e.g., omics or radiomic features) that are not routinely collected across external datasets, making independent validation impractical. Furthermore, reliance on proprietary hospital data and the absence of harmonized, multi-center datasets constrained opportunities for validating models beyond the original development setting. As a result, many models undergo, at most, internal validation and are rarely evaluated in truly independent populations.

Although external validation is critical for establishing reliable clinical guidelines, many newly identified predictors and prediction models lack independent external validation, limiting their generalizability.[46] Moreover, while developing models across multiple settings and ensuring rigorous external validation with proper calibration are essential to address heterogeneity, many of the reviewed models have been validated only within narrowly defined populations, underscoring the inherent difficulty in achieving broad generalizability. When applied to different regions, variations in patient characteristics, measurement methods, and healthcare systems can lead to poor model performance.[47] Therefore, to address these issues, we recommend prioritizing external validation and encourage multi-center studies, particularly international collaborations, to establish shared databases.

However, the execution of multi-center validation presents profound challenges that extend beyond data collection simply. Key hurdles include: (1) population heterogeneity: variations in age, ethnicity, comorbidity profiles, and disease subtypes across centers can significantly alter the baseline risk, affecting model performance. Indeed, the model should be validated in the homogeneous population as that where the model developed. (2) measurement inconsistencies: for models incorporating features like radiomics, differences in imaging protocols (e.g., CT scanner types, acquisition parameters), segmentation methods, and feature extraction algorithms can introduce variance that severely degrades model performance. (3) healthcare system disparities: differences in local treatment guidelines, diagnostic criteria, follow-up schedules, and data recording practices create a "domain shift", that is the data used for model training is different from other data used for model validation. (4) data accessibility: stringent data privacy regulations, e.g., General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA), ethical approvals, and institutional policies often create formidable logistical and legal barriers to share individual-level data across institutions or borders.

To overcome these barriers and realize the promise of truly generalizable models, the field must adopt more sophisticated strategies. We recommend: (i) the pre-planned standardization of data collection and processing protocols in prospective, multi-center studies; (ii) the adoption of privacy-preserving distributed learning techniques, such as federated learning, which allows models to be trained across multiple institutions without sharing raw data, instead sharing only model parameters and their updates; and (iii) continued advocacy for and investment in curated, open-access data platforms that follow FAIR (Findable, Accessible, Interoperable and Reusable) principles for specific, well-defined clinical questions.

## Predictors

Most NSCLC prognostic models are built on common factors like age, stage and histological type. Despite the inclusion of over a hundred predictive factors in fewer than two models, each research is focused on discovering new predictors rather than validating existing predictors. High-quality models may benefit from combining various predictors, including surgical, pathological, and biomarkers of gene expressions and radiomic features.

Among the 37 studies featuring gene expression, 213 different genes were used, with only 11 appearing more than twice. *SLC2A1* (solute carrier family 2 member 1), involved in glucose transport,[48] was the most frequently used, appearing in three studies. Its expression is linked to tumor progression, drug resistance, and poor prognosis in lung cancer, making it a promising biomarker for NSCLC prognostic models.[49] *EGFR* (epidermal growth factor receptor), also recognized as a somatic mutation, has been considered a key factor influencing the prognosis of NSCLC.[50] However, many studies exhibit significant issues, including poorly study design to inadequate analysis (e.g., non-linear effect analysis or interaction analysis) and biased reporting.[51] Our review reveals similar issues. Despite exhibiting strong predictive performance in internal validation, these gene signatures frequently fail to achieve comparable results in external validation. This inconsistency may be related to heterogeneity in data preprocessing procedures, differences between

Table 1. Description of population and characteristics for 89 externally validated models across 66 studies.

| Characteristics | Training set (*n*=89) | One testing set [a] (*n*=67) | Multiple testing sets [a] (*n*=22) |
|---|---|---|---|
| Participants location, *n* (%) [b] | | | |
| North America | 56 (62.92) | 31 (46.27) | 11 (50.00) |
| Europe | 12 (13.48) | 12 (17.91) | 4 (18.18) |
| Asia | 16 (17.98) | 24 (35.82) | 0 (0) |
| Multi-continents[c] | 3 (3.37) | 0 (0) | 6 (27.27) |
| Oceania | 2 (2.25) | 0 (0) | 1 (4.55) |
| Publication years, Median (range) | | | |
| 2008-2020 | 524 (106-10268), *n*=24 | 125 (52-1266), *n*=14 | 152 (33-2148), *n*=10 |
| 2021-2025 | 479 (92-22617), *n*=65 | 156 (41-792), *n*=53 | 229 (36-43767), *n*=12 |
| Statistical method, Median (range) | | | |
| Cox | 519.5 (92-18805), *n*=54 | 156.5 (41-1266), *n*=40 | 248 (36-43767), *n*=14 |
| Machine learning | 349 (95-22617), *n*=35 | 132 (41-400), *n*=27 | 65 (33-229), *n*=8 |
| AUC, Median (range) | | | |
| 1-year | 0.717 (0.623-0.994) | 0.759 (0.570-0.974) | 0.728 (0.600-0.881) |
| 3-year | 0.708 (0.638-0.994) | 0.737 (0.640-0.834) | 0.680 (0.618-0.782) |
| 5-year | 0.701 (0.599-0.980) | 0.728 (0.596-0.846) | 0.652 (0.548-0.775) |
| C-index, Median (range) | 0.701 (0.563-0.932) | 0.684 (0.570-0.960) | 0.670 (0.580-0.760) |

a : Samples from internal validation are excluded in the analysis of study population. b : If populations in training and validation sets are from multiple continents, the model is classified to 'multi-continents' group. c : Datasets (either training validation) that are from multiple, distinct regions across different continents.

sequencing and microarray platforms, batch effects, and variations in tumor expression profiles. Such technical and biological discrepancies can influence measured gene expression levels and may contribute to reduced robustness of the signatures in independent datasets.[52-54] Future models would be benefited from using gene expression features validated in multiple studies or across various models to improve prognostic predictions.

## Radiomic features

***Overall role of radiomics in NSCLC prognosis.*** Radiomics provides a wide range of imaging biomarkers, offering potential improvements in cancer detection, prognosis, and treatment response prediction.[55] Current radiomic researches have significantly enhanced NSCLC prognostic models by incorporating a wide range of radiomic features. We summarized several factors influencing the accuracy of radiomics-based models, including image quality, feature extraction methods. Given these factors, we categorized the 34 studies with radiomics into four types. Each type of these model presents its own set of advantages and limitations.

***Lack of standardization in radiomic workflows.*** However, beyond the model architecture, a critical challenge for the entire field lies in the profound lack of standardization in radiomic workflows, which severely threatens the reproducibility and generalizability of these promising models. The performance of radiomics-based models is highly sensitive to variations in multiple steps of the pipeline. Key sources of heterogeneity include: (1) image acquisition: different CT scanner vendors, acquisition protocols (e.g., slice thickness, reconstruction kernels, dose levels), and magnetic resonance imaging (MRI) or positron emission tomography (PET)-CT parameters can lead to significant inconsistencies in the extracted features. (2) tumor segmentation: process of delineating the tumor region of interest (ROI), whether manual, semi-automatic, or fully automatic, introduces inter-observer and inter-algorithm variability. (3) feature extraction and pre-processing: different software platforms and feature definitions (e.g., from PyRadiomics versus in-house tools) without standardized pre-processing steps (e.g., image resampling, discretization) can yield non-comparable results for the same biological characteristic.

Therefore, the impact of these non-standardized workflows is that a model developed on data from one institution may fail to perform adequately when applied to data from another institution, even the underlying clinical question is identical.

***Limitations of different types of radiomics-based models.*** The first two types of radiomics-based models have distinct drawbacks. They often exclude key clinical factors like age, gender, and stage, which are crucial for prognosis. Additionally, traditional radiomics extraction methods are less effective compared to deep learning-based approaches.[56] The feature selection process also requires optimization. According to a study by Aerts et al., radiomic features are typically grouped into categories such as original, tumor intensity, first-order, and wavelet.[57] While this method identifies representative features from each group, it may miss other robust features, potentially leading to model overfitting, especially when using high-dimensional data.[58] This underscores the imperative for developing and implementing more sophisticated feature selection strategies in the construction of radiomics-based risk score models.

***Integrated models and future directions.*** Despite the third type of model integrating radiomic and clinical factors and performing best in terms of quantity and overall performance, it is often limited by its development within one single institution, leading to potential biases and limitations in generalizability due to small sample size. Thus, balancing radiomics with clinical factors is key for accurate predictions, as achieved by the fourth type of models using a probability-weighted approach. The fourth type of models address this by integrating clinical factors like age, histological type, and tumor staging through a probability-weighted approach, enhancing predictive capabilities beyond what could be achieved by considering each set in isolation. However, we emphasize the validity of this approach and recommend its broader implementation in future research to enhance the construction of radiomics models.

In summary, despite existing limitations, radiomics shows considerable promise for disease diagnosis and prognosis prediction. As the field evolves, the integration of radiomics into clinical models is expected to have a

substantial improvement of prognostic prediction of NSCLC.

### PROBAST

Our findings indicate that most studies with high ROB mainly encounter issues in the analysis and participant domains. These problems often arise from inadequate attention to participant inclusion or exclusion criteria, leading to unrepresentative participants. Additionally, model training often neglects overfitting or underfitting concerns, resulting in poor generalizability on independent datasets. Besides, by growing use of machine learning for prognostic models, PROBAST's items 2.3, 4.1, and 4.9 may need adjustments, and items 4.6 to 4.8 should be reconsidered to better align with machine learning's characteristics—high-dimensionality, complexity, and black-box models.[59] To further evaluate this effect, we examined the association between ROB scores and the self-reported model performance. We observed that studies with higher ROB tended to report markedly inflated discrimination in the training set, but substantially reduced performance in validation set. This pattern indicates that greater bias is linked to overoptimistic internal estimates and poor generalizability. In real-world clinical settings, such bias-driven overoptimism may result in unstable risk stratification across institutions and patient subgroups, thereby limiting the reliability and adoption of these models in routine clinical practice. Moreover, PROBAST should incorporate guidelines for overfitting control, algorithm audits, and sample size calculations in machine learning contexts. For models using imaging data, ensuring image quality control is essential, and adherence to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement is recommended to improve the reporting quality of prediction model study.[60]

### The lack of histology-specific prognostic prediction model

Our review was designed to provide a comprehensive overview of prognostic prediction models for NSCLC. However, in response to a pertinent critique, our analysis revealed a critical limitation not only in our initial approach but, more importantly, in the field itself: the notable scarcity of prognostic models developed and validated for specific histological subtypes. Upon detailed re-examination of the 89 externally validated models, we found that only 6 models were specifically designed for lung adenocarcinoma (LUAD), and none existed for lung squamous cell carcinoma (LUSC).

This lack of histology-specific models is a significant oversight. LUAD and LUSC are distinct diseases with different genetic drivers, tumor microenvironments, and treatment responses, which logically translate to divergent prognostic factors and outcomes.[61] The prevailing practice of developing models on mixed NSCLC cohorts inherently assumes a uniform prognostic structure across subtypes, an assumption that is biologically untenable and likely obscures histology-specific risk associations.[62] Consequently, the generalizability and clinical applicability of most existing models to individual patients with a specific subtype remain uncertain. Therefore, we identify this as a paramount direction for future research. There is an urgent need to move beyond NSCLC models. Future studies should prioritize the development, validation, and reporting of histology-stratified performance for LUAD, LUSC, and other relevant subtypes. This is essential to build prognostication tools that truly reflect the underlying biological diversity of NSCLC and can deliver more accurate, personalized risk assessments in clinical practice.

### The lack of comprehensive statistical evaluation of prediction model

Statistical evaluation of prognostic model performance commonly consists of three aspects: (1) ROC and AUC for model discrimination, (2) E:O value for model calibration, and (3) DCA for clinical utility. However, we observed a severe and systematic neglect of model calibration and DCA. Only 8.99% (8/89) and 15.7% (14/89) of externally validated models reported results of calibration and DCA, respectively. This omission is critical because a model with good discrimination but poor calibration can mislead clinical practice; it may overestimate or underestimate a patient's risk, potentially leading to overtreatment or missed intervention. Two possible reasons likely account for this widespread issue. (1) There is a pervasive overemphasis on discrimination metrics like the C-index and AUC within the research community, which a good discrimination is often mistakenly equated with a good model. (2) Obtaining a reliable calibration assessment requires a sufficient sample size with an enough number of events, a requirement that many single center studies fail to meet, prompting researchers to omit this analysis altogether.

To address this critical gap, the future study must elevate the importance of comprehensive model evaluation. We propose three specific pathways forward: (1) strengthen standardized reporting: journals should require all index of model performance, including both graphical plots and key numerical metrics, which is outlined in the TRIPOD statement. (2) promote transparent reporting: researchers are encouraged to share full model parameters and codes on an open platform, allowing the scientific community to independently verify and evaluate model performance on original dataset and other diverse datasets. (3) develop adaptive frameworks: for clinical implementation, the development of interactive platforms that allow for model recalibration or updating using local data could dynamically optimize model performance and ensure its ongoing accuracy across different populations and time points.

### Barriers to clinical translation and pathways forward

A critical finding of our review is the stark disparity between the volume of developed prognostic models and their subsequent clinical application. Despite the emerging of amount of models, few have been integrated into routine clinical decision-making for NSCLC. Based on our analysis, we identify several key barriers to translation and propose specific pathways to overcome them.

As highlighted by our PROBAST assessment, the high risk of bias, particularly in the analysis domain, and the lack of rigorous external validation are fundamental barriers. Models developed from single institution, retrospective data with inadequate handling of overfitting are unlikely to perform reliably in new populations.

(1) To address this, we recommend a strict adherence to the step-by-step guide for clinical prediction models. This includes pre-specifying analysis plans, using appropriate methods for high-dimensional data, and most importantly, prioritizing prospective, multi-regional external validation before any clinical implementation can be considered. International consortia should be established to create shared, diverse validation cohorts.

(2) Many models remain as complex statistical formulas or algorithms buried in supplementary materials, making them inaccessible to clinicians. A model's clinical value is null if it cannot be easily used in real practice. Following the principles outlined by Moons et al., successful translation requires presenting models in user-friendly formats, such as nomograms, web calculators, or, ideally, as integrated components within the electronic health record (EHR) system.[59] The model development process should involve clinicians from the very beginning to ensure the model answers a pressing clinical question and well fits within existing workflow constraints.

Taking OvaRePred as a successful commercialization of scientific achievements, it is a comprehensive tool for not only assessing the current ovarian reserve of individuals but also for predicting important fertility milestones,[63,64] which has been incorporated into a paid medical testing service. In terms of prognostic prediction of NSCLC, Zhang et al. also released a free and user-friendly online tool, Prophet (http://bigdata.njmu.edu.cn/Prophet/).[19]

Indeed, the ultimate evaluation of a prognostic model is not merely statistical model performance, but whether its use improves patient outcomes or optimizes resource allocation. As outlined by Moons et al.,[59] impact study of a prediction model represents the essential final step in the translation pathway. Thus, we strongly advocate for the impact study of a prediction model, where the model is implemented in a clinical setting and its effect on clinician behaviors, patient satisfaction, treatment decisions, and ultimately, survival and quality of life, is rigorously evaluated.

Taking the Optimal Vasopressin Initiation in Septic Shock (OVISS) study as an example, which shows that model-recommended actions outperform clinician decision-making for in-hospital mortality of septic shock patients.[65] However, it is missing in the field of prognostic perdition model currently.

### Strengths, limitations and comparison with other studies

The strengths of our review lie in its comprehensive assessment of NSCLC prognostic models, encompassing all publications from five databases since their inception. It details model development, predictive factors, outcome prediction, model development, and external validation. We identified,

extracted, and compared each model's performance, and discussed geographic disparities, prognostic factors, modeling methods, and bias control. However, our review still has limitations. First, non-English publications, those without full texts, or studies in formats like conferences or seminars were excluded, which is a common practice in reviews.[66] Second, manual inclusion and exclusion of published studies might introduce potential bias despite we made efforts to minimize errors.[67] Finally, based on the BMJ 13-step guideline, steps like evaluating methods for handling missing data and the predictive power of individual factors couldn't be fully assessed due to the limitations in the available study data.

### Challenges and opportunities

Despite advancements in lung cancer prognosis research and the development of various prognostic models, significant challenges still remain. Many studies face methodological limitations, such as overfitting in Cox models due to an excess of predictors. Furthermore, inconsistent approaches to handling missing data, stemming from data constraints, result in most models being classified as high-risk, highlighting the need for more rigorous analysis in model development. Additionally, there is great opportunities to enhance risk prediction, given the increasing integration of big data from diagnosis, treatment, physical examinations and multi-omics profiling.

## CONCLUSION

This study provides the first and the largest systematic analysis and statistical evaluation of existing prognostic models for NSCLC, highlighting current advancements and limitations. Despite the progress, challenges persist, notably the lack of independent external validation and widespread risk of bias. Integrating various factors, including surgical, pathological predictors, medical history, and biomarkers like gene expression and radiomics, offers opportunity for developing potentially better models. However, ensuring the reliability and generalizability of these prognostic models, rigorous validation and bias-reduction strategies are essential.

To achieve genuine clinical translation and impact, a concerted, multi-stakeholder effort is imperative. For researchers, this entails prioritizing robust external validation through prospective, multi-regional studies to establish generalizability, while rigorously adhering to reporting guidelines (e.g., TRIPOD and BMJ 13-step guidelines) and enhancing practical usability via open-source tools like web calculators. A critical focus must be placed on comprehensive model evaluation (i.e., model discrimination, model calibration, clinical utility and clinical impact) complemented by fairness assessments to ensure equitable performance across demographic subgroups. For clinicians and hospital administrators, they should champion the real-world evaluation of promising models within local workflows and multi-center pilots, drive integration of model into clinical decision support systems (CDSS) to provide actionable insights in real practice. For policymakers and healthcare systems, the mandate is to foster secure federated data platforms that enable large-scale validation while safeguarding privacy, and to establish clear regulatory frameworks for the registration, monitoring, and fairness auditing of these models. A dedicated commitment to equity is essential, requiring targeted funding and resources to support validation in underrepresented and low-resource settings, thereby preventing the exacerbation of global health disparities and ensuring the development of effective prognostic prediction tools for all populations.

## LIST OF ABBREVIATIONS

AUC, area under the receiver operating characteristic curve; BMI, body mass index; CCI, Charlson comorbidity index; CDSS, Clinical Decision Support Systems; CHARMS, CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies; CT, computed tomography; CVD, cardiovascular disease; C-index, concordance index; DAG, directed acyclic graph; DCA, decision curve analysis; ECOG-PS, Eastern Cooperative Oncology Group performance status; EHR, electronic health record; FAIR, Findable, Accessible, Interoperable, and Reusable; FEV1, forced expiratory volume in one second; GDPR, General Data Protection Regulation; GTV, gross tumor volume; HIPAA, Health Insurance Portability and Accountability Act; KPS, Karnofsky performance status; LDH, lactate dehydrogenase; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; ML, machine learning; MRI, magnetic resonance imaging; NLR, neutrophil-to-lymphocyte ratio; NSCLC, non-small cell lung cancer; OS, overall survival; PET, positron emission tomography; PLNS, positive lymph nodes ratio; PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses; PROBAST, Prediction model Risk Of Bias Assessment Tool; ROB, risk of bias; ROI, region of interest; SEER, Surveillance, Epidemiology, and End Results Program; SPA, saddlepoint approximation; SPC, surgical and pathological characteristics; TCGA, The Cancer Genome Atlas; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; WBC, white blood cell; WHO-PS, World Health Organization performance status.

## REFERENCES


1. Bray F., Laversanne M., Sung H., et al. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**:229–263. DOI:10.3322/caac.21834
2. Chen P., Liu Y., Wen Y., et al. (2022). Non-small cell lung cancer in China. *Cancer Commun. (Lond)* **42**:937–970. DOI:10.1002/cac2.12359
3. Lai Y. H., Chen W. N., Hsu T. C., et al. (2020). Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Sci. Rep.* **10**:4679. DOI:10.1038/s41598-020-61588-w
4. Liu Z., Zhang J., Liu J., et al. (2023). Combining network pharmacology, molecular docking and preliminary experiments to explore the mechanism of action of FZKA formula on non-small cell lung cancer. *Protein Pept. Lett.* **30**:1038–1047. DOI:10.2174/0109298665268153231024111622
5. Wang Y., Lin X. and Sun D. (2021). A narrative review of prognosis prediction models for non-small cell lung cancer: What kind of predictors should be selected and how to improve models. *Ann. Transl. Med.* **9**:1597. DOI:10.21037/atm-21-4733
6. Brundage M. D., Davies D. and Mackillop W. J. (2002). Prognostic factors in non-small cell lung cancer: A decade of progress. *Chest* **122**:1037–1057. DOI:10.1378/chest.122.3.1037
7. Cai Y., Sheng Z., Dong Z., et al. (2023). EGFR inhibitor CL-387785 suppresses the progression of lung adenocarcinoma. *Curr. Mol. Pharmacol.* **16**:211–216. DOI:10.2174/1874467215666220329212300
8. Efthimiou O., Seo M., Chalkou K., et al. (2024). Developing clinical prediction models: A step-by-step guide. *BMJ* **386**:e078276. DOI:10.1136/bmj-2023-078276
9. Debray T. P., Damen J. A., Snell K. I., et al. (2017). A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**:i6460. DOI:10.1136/bmj.i6460
10. Moher D., Liberati A., Tetzlaff J., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ* **339**:b2535. DOI:10.1136/bmj.b2535
11. Moons K. G. M., Wolff R. F., Riley R. D., et al. (2019). PROBAST: A tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and elaboration. *Ann. Intern. Med.* **170**:W1–W33. DOI:10.7326/M18-1377
12. Xie H. and Xie C. (2019). A six-gene signature predicts survival of adenocarcinoma type of non-small-cell lung cancer patients: A comprehensive study based on integrated analysis and weighted gene coexpression network. *Biomed. Res. Int.* **2019**:4250613. DOI:10.1155/2019/4250613
13. Paesmans M. (2012). Prognostic and predictive factors for lung cancer. *Breathe* **9**:112–121. DOI:10.1183/20734735.006911
14. Wang T., Lu R., Lai S., et al. (2019). Development and validation of a nomogram prognostic model for patients with advanced non-small-cell lung cancer. *Cancer Inform.* **18**:1176935119837547. DOI:10.1177/1176935119837547.
15. Wang H., Wang X., Xu L., et al. (2020). High expression levels of pyrimidine metabolic rate-limiting enzymes are adverse prognostic factors in lung adenocarcinoma: A study based on The Cancer Genome Atlas and Gene Expression Omnibus datasets. *Purinergic Signal.* **16**:347–366. DOI:10.1007/s11302-020-09711-4
16. Nguyen T. T., Lee H. S., Burt B. M., et al. (2022). A lepidic gene signature predicts patient prognosis and sensitivity to immunotherapy in lung adenocarcinoma. *Genome Med.* **14**:5. DOI:10.1186/s13073-021-01010-w
17. Li F., Bing Z., Chen W., et al. (2021). Prognosis biomarker and potential therapeutic target CRIP2 associated with radiosensitivity in NSCLC cells. *Biochem. Biophys. Res. Commun.* **584**:73–79. DOI:10.1016/j.bbrc.2021.11.002
18. Chen J., Wu R., Xuan Y., et al. (2020). Bioinformatics analysis and experimental validation of TTK as a biomarker for prognosis in non-small cell lung cancer. *Biosci. Rep.* **40**:BSR20202711. DOI:10.1042/BSR20202711
19. Zhang R., Chen C., Dong X., et al. (2020). Independent validation of early-stage non-small cell lung cancer prognostic scores incorporating epigenetic and transcriptional biomarkers with gene-gene interactions and main effects. *Chest* **158**:808–819. DOI:10.1016/j.chest.2020.01.048
20. Tang F. H., Fong Y. W., Yung S. H., et al. (2023). Radiomics-clinical AI model with probability weighted strategy for prognosis prediction in non-small cell lung cancer. *Biomedicines* **11**:2093. DOI:10.3390/biomedicines11082093
21. Wang X., Guo Z., Wu X., et al. (2023). Predictive nomogram for hyperprogressive disease during anti-PD-1/PD-L1 treatment in patients with advanced non-small cell

lung cancer. *Immunotargets Ther.* **12**:1–16. DOI:10.2147/ITT.S373866
22. Li X., Gu W., Liu Y., et al. (2022). A novel quantitative prognostic model for initially diagnosed non-small cell lung cancer with brain metastases. *Cancer Cell Int.* **22**:251. DOI:10.1186/s12935-022-02671-2
23. Huang Z., Xing S., Zhu Y., et al. (2020). Establishment and validation of nomogram model integrated with inflammation-based factors for the prognosis of advanced non-small cell lung cancer. *Technol. Cancer Res. Treat.* **19**:1533033820971605. DOI:10.1177/1533033820971605
24. Huang H., Chen Y., Weng X., et al. (2022). Development and validation of a nomogram for evaluating the prognosis of immunotherapy plus antiangiogenic therapy in non-small cell lung cancer. *Cancer Cell Int.* **22**:261. DOI:10.1186/s12935-022-02675-y
25. He L. N., Chen T., Fu S., et al. (2022). Reducing number of target lesions for RECIST1.1 to predict survivals in patients with advanced non-small-cell lung cancer undergoing anti-PD1/PD-L1 monotherapy. *Lung Cancer* **165**:10-17. DOI:10.1016/j.lungcan.2021.12.015
26. Dinglin X. X., Ma S. X., Wang F., et al. (2017). Establishment of an adjusted prognosis analysis model for initially diagnosed non-small-cell lung cancer with brain metastases from Sun Yat-Sen University Cancer Center. *Clin. Lung Cancer* **18**:e179–e186. DOI:10.1016/j.cllc.2016.12.016
27. Chen S., Li X., Lv H., et al. (2018). Prognostic dynamic nomogram integrated with inflammation-based factors for non-small cell lung cancer patients with chronic hepatitis B viral infection. *Int. J. Biol. Sci.* **14**:1813–1821. DOI:10.7150/ijbs.27260
28. Chen S., Lai Y., He Z., et al. (2018). Establishment and validation of a predictive nomogram model for non-small cell lung cancer patients with chronic hepatitis B viral infection. *J. Transl. Med.* **16**:116. DOI:10.1186/s12967-018-1496-5
29. Chen S., Huang H., Liu Y., et al. (2020). A multi-parametric prognostic model based on clinical features and serological markers predicts overall survival in non-small cell lung cancer patients with chronic hepatitis B viral infection. *Cancer Cell Int.* **20**:555. DOI:10.1186/s12935-020-01635-8
30. Cao X., Zheng Y. Z., Liao H. Y., et al. (2020). A clinical nomogram and heat map for assessing survival in patients with stage I non-small cell lung cancer after complete resection. *Ther. Adv. Med. Oncol.* **12**:1758835920970063. DOI:10.1177/1758835920970063
31. Xiao W., Geng W., Xu J., et al. (2023). Construction and validation of a nomogram based on N6-Methylandenosine-related lncRNAs for predicting the prognosis of non-small cell lung cancer patients. *Cancer Med.* **12**:2058–2074. DOI:10.1002/cam4.4961
32. Miao T. W., Chen F. Y., Du L. Y., et al. (2022). Signature based on RNA-binding protein-related genes for predicting prognosis and guiding therapy in non-small cell lung cancer. *Front. Genet.* **13**:930826. DOI:10.3389/fgene.2022.930826
33. He R. and Zuo S. (2019). A robust 8-gene prognostic signature for early-stage non-small cell lung cancer. *Front. Oncol.* **9**:693. DOI:10.3389/fonc.2019.00693
34. Siontis G. C., Tzoulaki I., Siontis K. C., et al. (2012). Comparisons of established risk prediction models for cardiovascular disease: Systematic review. *BMJ* **344**:e3318. DOI:10.1136/bmj.e3318
35. Allemani C., Matsuda T., Di Carlo V., et al. (2018). Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): Analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet* **391**:1023–1075. DOI:10.1016/S0140-6736(17)33326-3
36. Goss P. E., Strasser-Weippl K., Lee-Bychkovsky B. L., et al. (2014). Challenges to effective cancer control in China, India, and Russia. *Lancet Oncol.* **15**:489–538. DOI:10.1016/S1470-2045(14)70029-4
37. Weiss K., Khoshgoftaar T. M. and Wang D. (2016). A survey of transfer learning. *J. Big Data* **3**:9. DOI:10.1186/s40537-016-0043-6
38. Gu T., Han Y. and Duan R. (2023). A transfer learning approach based on random forest with application to breast cancer prediction in underrepresented populations. *Pac. Symp. Biocomput.* **28**:186–197.
39. Theodoris C. V., Xiao L., Chopra A., et al. (2023). Transfer learning enables predictions in network biology. *Nature* **618**:616–624. DOI:10.1038/s41586-023-06139-9
40. Spooner A., Chen E., Sowmya A., et al. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci. Rep.* **10**:20410. DOI:10.1038/s41598-020-77220-w
41. Mbatchou J., Barnard L., Backman J., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**:1097–1103. DOI:10.1038/s41588-021-00870-7
42. Zhou W., Nielsen J. B., Fritsche L. G., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**:1335–1341. DOI:10.1038/s41588-018-0184-y
43. Dey R., Zhou W., Kiiskinen T., et al. (2022). Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nat. Commun.* **13**:5437. DOI:10.1038/s41467-022-32885-x
44. Denny J. C., Bastarache L., Ritchie M. D., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**:1102–1110. DOI:10.1038/nbt.2749
45. Akazawa M. and Hashimoto K. (2021). Artificial intelligence in gynecologic cancers: Current status and future challenges - A systematic review. *Artif. Intell. Med.* **120**:102164. DOI:10.1016/j.artmed.2021.102164
46. Altman D. G., Vergouwe Y., Royston P., et al. (2009). Prognosis and prognostic research: Validating a prognostic model. *BMJ* **338**:b605. DOI:10.1136/bmj.b605
47. Justice A. C., Covinsky K. E. and Berlin J. A. (1999). Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**:515–524. DOI:10.7326/0003-4819-130-6-199903160-00016
48. Liu W. T., Wang Y., Zhang J., et al. (2018). A novel strategy of integrated microarray analysis identifies CENPA, CDK1 and CDC20 as a cluster of diagnostic biomarkers in lung adenocarcinoma. *Cancer Lett.* **425**:43–53. DOI:10.1016/j.canlet.2018.03.043
49. Commander R., Wei C., Sharma A., et al. (2020). Subpopulation targeting of pyruvate dehydrogenase and GLUT1 decouples metabolic heterogeneity during collective cancer cell invasion. *Nat. Commun.* **11**:1533. DOI:10.1038/s41467-020-15219-7
50. Wood D. E., Kazerooni E. A., Baum S. L., et al. (2018). Lung cancer screening, version 3.2018, NCCN clinical practice guidelines in oncology. *J. Natl. Compr. Canc. Netw.* **16**:412-441. DOI:10.6004/jnccn.2018.0020
51. Subramanian J. and Simon R. (2010). Gene expression-based prognostic signatures in lung cancer: Ready for clinical use. *J. Natl. Cancer Inst.* **102**:464–474. DOI:10.1093/jnci/djq025
52. Sauerbrei W., Taube S. E., McShane L. M., et al. (2018). Reporting recommendations for tumor marker prognostic studies (REMARK): An abridged explanation and elaboration. *J. Natl. Cancer Inst.* **110**:803–811. DOI:10.1093/jnci/djy088
53. Dagogo-Jack I. and Shaw A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**:81–94. DOI:10.1038/nrclinonc.2017.166
54. Gerlinger M., Rowan A. J., Horswell S., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**:883–892. DOI:10.1056/NEJMoa1113205
55. Royston P., Moons K. G., Altman D. G., et al. (2009). Prognosis and prognostic research: Developing a prognostic model. *BMJ* **338**:b604. DOI:10.1136/bmj.b604
56. Vial A., Stirling D., Field M., et al. (2018). The role of deep learning and radiomic feature extraction in cancer-specific predictive modeling: A review. *Transl. Cancer Res.* **7**:803–816. DOI:10.21037/tcr.2018.05.02
57. Aerts H. J., Velazquez E. R., Leijenaar R. T., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**:4006. DOI:10.1038/ncomms5006
58. Leger S., Zwanenburg A., Pilz K., et al. (2017). A comparative study of machine learning methods for time-to-event survival data for radiomics risk modeling. *Sci. Rep.* **7**:13206. DOI:10.1038/s41598-017-13448-3
59. Cai Y., Cai Y. Q., Tang L. Y., et al. (2024). Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review. *BMC Med.* **22**:56. DOI:10.1186/s12916-024-03273-7
60. Collins G. S., Reitsma J. B., Altman D. G., et al. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ* **350**:g7594. DOI:10.1136/bmj.g7594
61. Kris M. G., Johnson B. E., Berry L. D., et al. (2014). Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. *JAMA* **311**:1998–2006. DOI:10.1001/jama.2014.3741
62. Riihimaki M., Hemminki A., Fallah M., et al. (2014). Metastatic sites and survival in lung cancer. *Lung Cancer* **86**:78–84. DOI:10.1016/j.lungcan.2014.07.020
63. Xu H., Feng G., Yang R., et al. (2025). Reaffirming the role of ovarian reserve in fertility assessment: Insights from OvaRePred. *Innov. Med.* **3**:100135. DOI:10.59717/j.xinn-med.2025.100135
64. Xu H., Feng G., Yang R., et al. (2023). OvaRePred: Online tool for predicting the age of fertility milestones. *The Innovation* **4**:100490. DOI:10.1016/j.xinn.2023.100490
65. Kalimouttou A., Kennedy J. N., Feng J., et al. (2025). Optimal vasopressin initiation in septic shock: The OVISS reinforcement learning study. *JAMA* **333**:1688–1698. DOI:10.1001/jama.2025.3046
66. Moons K. G., Altman D. G., Vergouwe Y., et al. (2009). Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ* **338**:b606. DOI:10.1136/bmj.b606
67. Moons K. G., Kengne A. P., Grobbee D. E., et al. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**:691–698. DOI:10.1136/heartjnl-2011-301247

## FUNDING AND ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (82220108002 to F.C., 82273737 to R.Z., 82473728 to Y.W., 82173620 and 82373690 to Y.Z.), the US National Institutes of Health (CA209414, CA249096, CA092824, and ES000002 to D.C.C., CA249096 and CA209414 to Y.L.), Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). R.Z. was partially supported by the Outstanding Young Teachers Training Program of Nanjing Medical University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## AUTHOR CONTRIBUTIONS

R.Z. and F.C. designed the study. X.X., L.C., M.X., A.W., Q.F., Y.Y., J.F., T.H., Y.P., S.X., Y.C., Z.C., Z.W., J.X., Z.Z., and Y.Z. selected articles or extracted data. X.X., L.C., F.C., and R.Z. analyzed and interpreted the data. X.X., L.C., R.Z. wrote the first draft of the

manuscript, which was revised by all authors. X.X., L.C., Y.Z., L.Z., F.C., and R.Z. critically revised the manuscript. All authors approved the final version of the submitted manuscript. We thank Y.L., R.J.H., C.I.A., and D.C.C. for their valuable insights and expertise, which greatly contributed to the success of this project. All authors have full access to all of the data (including statistical reports and tables) in the study, take responsibility for the integrity of the data, and the accuracy of the data analysis. R.Z. is the guarantor. The corresponding authors confirm that all authors meet the authorship criteria, and no individuals who meet these criteria have been omitted.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## ETHICAL STATEMENT AND PATIENT CONSENT

Not Applicable.

## DATA AND CODE AVAILABILITY

The authors adhered to the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies (CHARMS) in systematic review, and this study has also been registered in the International Prospective Register of Systematic Reviews (PROSPERO ID: CRD42024534850). All updates to the review will be documented in the registry.

## SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.59717/j.xinn-med.2026.100195.