Chang *et al. Journal of Translational Medicine* (2026) 24:108
https://doi.org/10.1186/s12967-025-07591-z

Journal of Translational Medicine

REVIEW Open Access

Applications of artificial intelligence in non–small cell lung cancer: from precision diagnosis to personalized prognosis and therapy

Luyuan Chang[1], Haipeng Li[2], Wenzong Wu[1], Xinyu Liu[1], Jiaqi Yan[1], Zuo Chen[1], Huan Wu[2] and Shilong Song[3*]

Abstract

Background Non-small cell lung cancer (NSCLC) carries a major global burden. The rapid growth of multimodal medical data challenges conventional methods to deliver stable, transferable and interpretable decisions across heterogeneous longitudinal high dimensional inputs.

Methods This review summarizes advances in artificial intelligence (AI) for NSCLC from 2023 to 2025 and outlines a translation-focused framework that links algorithmic progress to clinical utility. We survey thoracic imaging, digital pathology and multiomics together with evaluation practices and implementation guidance. We also adopt a critical perspective.

Results Many high performing deep models remain black boxes, and popular post hoc explanations such as Grad CAM heatmaps are rarely validated for faithfulness or stability, which undermines clinician trust and limits use in high stakes decisions. To address this gap, we propose a minimum evidence package for explainability that comprises sanity checks, quantitative faithfulness tests such as deletion or insertion, ROAR or IROF and infidelity, stability analyses, concept level validation for example TCAV with statistical testing, and prospective human factors studies that demonstrate improved decisions without automation bias. Across modalities, evaluation has expanded beyond discrimination to include calibration, uncertainty quantification (UQ) and subgroup analyses across scanners, sites and populations. However, the evidence base remains constrained by retrospective single center designs, inconsistent external or temporal validation and limited decision curve analysis (DCA). Translational priorities include a staged validation ladder from technical to clinical to prospective deployment, alignment with Software as a Medical Device frameworks, interoperable governance, fairness and economic assessment, and validated explainability coupled with uncertainty aware selective workflows.

Conclusions Looking ahead, progress will depend on multimodal foundation models, causal and temporal modeling, and regulatory qualification of computable biomarkers with verified explanations, supported by multicenter prospective studies that demonstrate durable generalizability, clinical value and clinician trust.

*Correspondence:
Shilong Song
shilongsong@bbmu.edu.cn

Full list of author information is available at the end of the article

Keywords Non–small cell lung cancer (NSCLC), Artificial intelligence (AI), Precision diagnosis, Personalized prognosis, Treatment decision support

Introduction

Lung cancer remains the leading cause of cancer death worldwide. According to GLOBOCAN 2022, about 2.48 million new cases and 1.82 million deaths occurred globally, making lung cancer both highly incident and the leading cause of cancer mortality in many regions [1]. Non-small cell lung cancer (NSCLC) accounts for 80–85% of cases and comprises adenocarcinoma, squamous cell carcinoma, and large-cell carcinoma [2]. Despite advances in low-dose computed tomography (LDCT) screening, perioperative immuno-oncology, targeted therapies, and stereotactic radiotherapy, population outcomes remain suboptimal [2–4]. Recent SEER/ACS estimates show an overall 5-year relative survival of about 32% for NSCLC, with pronounced stage gradients: localized about 67%, regional about 40%, and distant about 12% [5]. Poor outcomes reflect late presentation, substantial inter- and intratumoral heterogeneity, and primary or acquired resistance to systemic therapy [6, 7]. These challenges strain traditional diagnostic and prognostic approaches, which struggle to integrate high-dimensional, multimodal information, including imaging, whole-slide histopathology, genomic and transcriptomic profiles, longitudinal clinical data, and environmental exposures, into actionable decisions [8, 9].

Heterogeneity and evolutionary dynamics are central to the biology of NSCLC [6, 7]. Spatially distinct subclones, variable target expression, divergent microenvironmental niches—including inflamed, excluded, and desert phenotypes—and shifting therapeutic selective pressures produce complex response patterns and resistance [10–13]. Environmental exposures, most notably fine particulate matter PM2.5, contribute to lung adenocarcinoma, particularly among never-smokers, and interact with molecular drivers such as the epidermal growth factor receptor (EGFR) [14]. This clinical, biological, and environmental heterogeneity motivates methods that learn structure across scales and generate individualized predictions that update over time [8, 15].

Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), has progressed from proof of concept to multicenter and early prospective evaluations across oncology workflows [16, 17]. In thoracic imaging, DL systems trained on LDCT and diagnostic CT achieve expert level performance in nodule detection and malignancy risk estimation. They are under prospective evaluation for longitudinal risk prediction in screening cohorts, for example models that predict one to six years of lung cancer risk from a single LDCT [18, 19]. In digital pathology, self supervised foundation models and multiple instance learning (MIL) standardize histologic subtyping, quantify immunohistochemistry (IHC) such as the programmed death ligand 1 (PD-L1) tumor proportion score (TPS) with higher reproducibility than manual scoring, and increasingly infer actionable biomarkers such as EGFR directly from routine hematoxylin and eosin (H&E) slides [12, 20–22]. Parallel advances in multimodal fusion and representation learning enable integration of radiology, pathology, and omics with structured clinical data to support individualized prognosis and treatment selection [8, 23, 24]. Building on these modality-specific gains, we present a unifying framework for multimodal fusion [25]. We define multimodal fusion as the principled integration of imaging, digital pathology, genomics, and clinical data to improve diagnosis, prognosis, and treatment in NSCLC [26]. This review explains how fusion is realized across radiology, pathology, and omics, describes the alignment and aggregation of heterogeneous representations into coherent patient-level embeddings, and shows how fused models are integrated into routine clinical workflows to support decision-making [27]. Throughout, we highlight fusion's contribution to generalization, interpretability, and clinical utility, and we evaluate each application with a consistent toolkit of external validation, calibration, and decision-curve analysis to enable fair, decision-relevant comparisons [28]. At the architectural level, contemporary systems include four major families. First, Transformer backbones use self attention for spatial and contextual integration [29]. Second, temporal and frequency attention, exemplified by Fourier attention (FA) and wavelet attention (WA), captures long range periodicity and multiscale transients in longitudinal imaging and circulating tumor DNA (ctDNA) time series [30–32]. Third, graph neural networks (GNNs) encode pathway and topological constraints for multi omics integration. Fourth, generative adversarial networks support denoising, super resolution, and stain or style normalization. Together with early prospective evaluations, these trends position AI as a scalable, data-driven layer that augments radiology, pathology, genomics, and clinical decision-making throughout the NSCLC care continuum [16, 33].

A targeted search of PubMed, MEDLINE, Embase, Web of Science, Scopus, and Cochrane CENTRAL identified studies published between 1 January 2023 and 17 August 2025. Search queries combined controlled vocabulary and free-text terms for NSCLC and AI across imaging, digital pathology, multi-omics, prognosis, treatment decision support, and drug discovery. Records were independently screened by two reviewers for relevance to the

review objectives, with priority given to human NSCLC studies featuring clinically meaningful endpoints and implementation insights. As a narrative review, no pre-registered protocol, formal risk-of-bias assessment, or quantitative synthesis was employed. When necessary to contextualize developments from 2023 to 2025, seminal prior work and select pre-2023 studies were cited. Using this methodology, we review 2023 to 2025 advances across the NSCLC pathway (Fig. 1): precision diagnosis and subtyping using CT, Positron Emission Tomography-Computed Tomography (PET-CT), and Magnetic Resonance Imaging (MRI) radiomics and deep learning [18], as well as whole-slide image biomarker inference [22]; individualized prognosis with multi-omics and multimodal survival models [34]; and decision support for radiotherapy, chemotherapy, targeted therapy, and immunotherapy, together with AI-enabled drug discovery [35]. We also highlight translational requirements, including external validity, calibration, clinical net benefit, fairness, and explainability, and we outline regulatory and workflow-integration considerations [36–38]. Finally, we propose an agenda for trustworthy, interpretable, and equitable deployment [39], including a pragmatic validation ladder and postmarket change-control and recalibration plans.

Applications of AI in diagnosis and subtyping of NSCLC

AIassisted imaging diagnosis (CT, PETCT, MRI; radiomics and deep learning)

In thoracic radiology, AI aims to improve sensitivity, specificity, and workflow efficiency in both screening and diagnostic evaluation [34]. Deep learning CNNs trained on chest CT detect and characterize pulmonary nodules, prioritize worklists, and generate quantitative malignancy-risk estimates, which reduces interobserver variability and mitigates reader fatigue in high-volume settings [34–36]. Nodule detection and longitudinal risk stratification are related but distinct tasks [18, 35]. End-to-end 3D CNNs for LDCT screening, exemplified by the 2019 Google system in Nature Medicine, achieved an area under the receiver operating characteristic curve (AUROC) of approximately 0.94 for per-case cancer detection and reduced false positives by about 11% and false negatives by about 5% versus expert readers in retrospective testing [38]. In parallel, the Sybil model, developed and validated across multiple centers and reported in the Journal of Clinical Oncology in 2023, predicts individual one- to six-year lung-cancer risk from a single LDCT without additional clinical covariates. External AUROCs of approximately 0.75–0.81 have been reported,

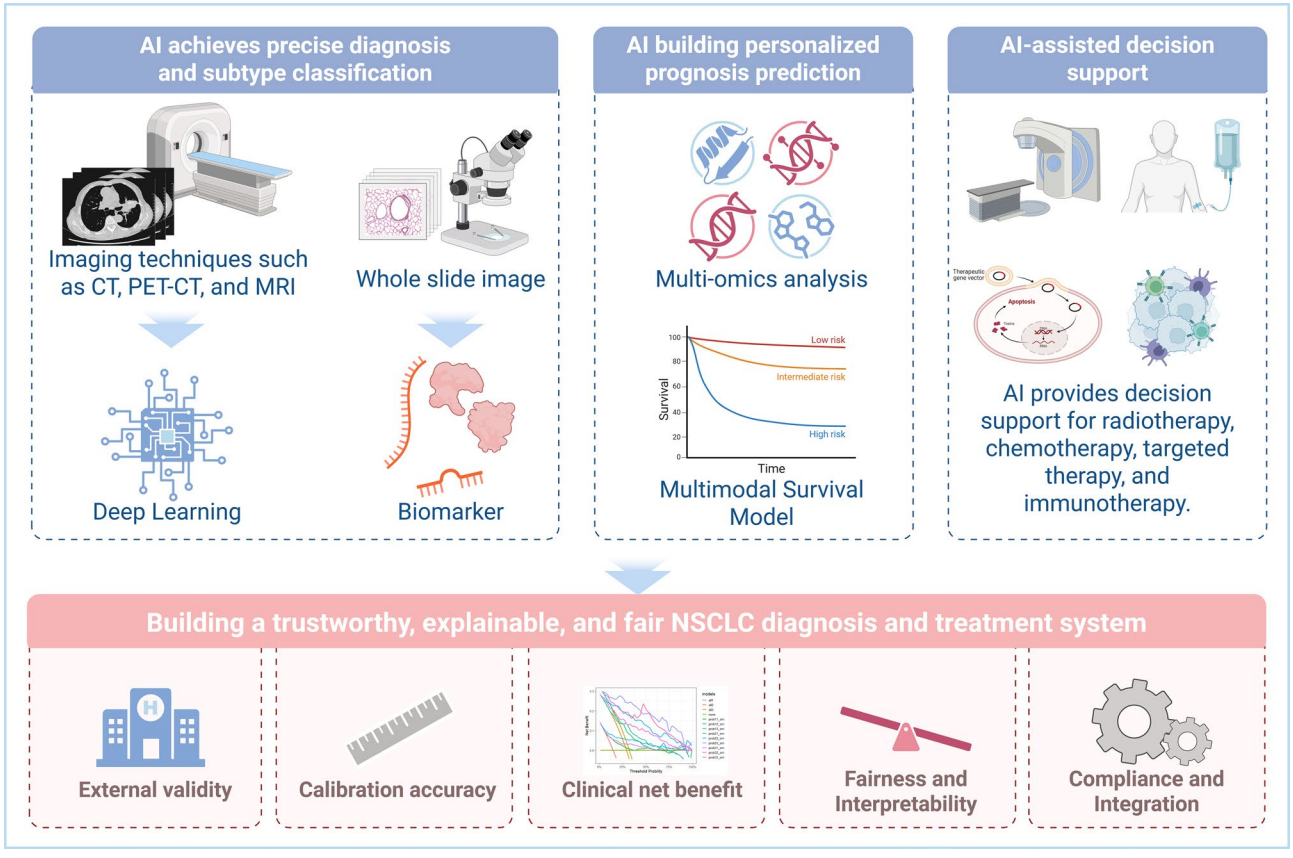

Fig. 1 AI-enabled NSCLC pathway for precise diagnosis, personalized prognosis and clinical decision support

and such models can inform interval scheduling and escalation pathways for high-risk individuals [18].

In addition to detection and risk stratification, DL enhances the segmentation of tumors and organs-at-risk, facilitating volumetry, growth-kinetics modeling, and radiotherapy planning [37, 39]. In PET-CT, AI aids in mediastinal staging, such as classifying nodal involvement with AUROCs around 0.90, and detecting occult metastases. In MRI, which is less commonly used for thoracic cancers, AI supports brain-metastasis surveillance and target-volume delineation [40–43]. Denoising and super-resolution techniques based on learning further enhance low-dose image quality, preserving quantitative imaging features [44, 45].

Methodologically, comparisons between radiomics and deep learning recur in thoracic oncology [46]. Handcrafted radiomics extracts shape, intensity, and texture features to infer phenotypes and biomarker surrogates such as histology, EGFR status, and invasiveness, yet it is sensitive to heterogeneity in slice thickness, reconstruction kernel, vendor, and to feature engineering choices [46, 47]. Accordingly, pipelines aligned with the Image Biomarker Standardisation Initiative (IBSI) should explicitly document voxel resampling with isotropic 1.0 mm, intensity discretization using a fixed bin size or number, pre filters such as Laplacian of Gaussian (LoG) radii, and the exact feature set; releasing code and parameter files enables independent reproducibility. By contrast, deep learning leverages voxel level signals and peritumoral context and often outperforms classical models when trained on diverse, harmonized datasets; however, both paradigms require external validation, calibration, and decision curve analysis (DCA) to demonstrate clinical utility [28, 48, 49] (Table 1). To carry these methods into practice, external validation and calibration must follow disciplined procedures. External validation and calibration: practical guidance. Lock the full analysis pipeline before any testing begins [52]. Select cohorts that are both geographically and temporally external and that span multiple vendors and protocols. Fit all preprocessing on development data only and apply it unchanged to the external cohort. Split at the patient level and keep sites and scanners separated to prevent leakage from acquisition signatures [53]. Report discrimination, calibration, and clinical utility together, and provide subgroup results by site, scanner, sex, age, ancestry, stage, and smoking status. See section"Validation, regulation, and workflow integration" for detailed checklists and sample size targets.

Domain shift remains a central challenge for deployment [54]. Performance on external cohorts often falls relative to internal testing because of protocol and population differences, and decreases of about 5 to 10 AUROC points are common across centers [55, 56]. Mitigation strategies include, first, multi domain training across vendors and protocols; second, strong data augmentation combined with physics informed harmonization such as ComBat and kernel aware resampling [47, 50, 57]; third, out of distribution (OOD) detection with case level uncertainty estimation to flag low reliability outputs [58, 59]; and fourth, test time adaptation or augmentation to stabilize predictions at deployment. At a minimum, reports should include internal and external validation splits, calibration curves, and error analyses stratified by scanner vendor, slice thickness, reconstruction kernel, geography, sex, age, and smoking status [60]. Rigorous external validation should use a locked pipeline [61], apply preprocessing learned on development data without modification, and adopt patient splits that are separated by site and scanner, while reporting discrimination, calibration, and decision analysis together; section"Validation, regulation, and workflow integration" provides an execution protocol that reduces optimistic bias and improves transportability [62].

Several nodule management solutions have received regulatory clearance, bridging research evidence and deployable clinical tools [63, 64]. Examples include Riverain ClearRead CT with FDA 510k clearance in the United States and Veye Lung Nodules with an EU MDR CE mark, which function as concurrent readers for detection, volumetry, and growth assessment in PACS integrated workflows [65–67]. Real world rollouts emphasize additional requirements: seamless RIS and PACS integration, stable inference latency, auditable trails, and

Table 1 Methodological comparison: radiomics vs deep learning

Topic	Radiomics (hand-crafted)	Deep Learning (end-to-end)	Key references
Features and inputs	Shape/intensity/texture features; sensitive to acquisition heterogeneity (slice thickness, reconstruction kernel, vendor)	Voxel-level signals plus peri-tumoral context; automatically learned multi-level representations	[46, 47]
Reproducibility and standardization	IBSI-aligned parameters: voxel resampling (e.g., 1.0 mm isotropic), intensity discretization, LoG radii; release code and parameter files	Document preprocessing/augmentation/architectures and weights; report differences between training and inference domains	[28, 47, 50]
Performance and generalization	Effective on homogeneous data but vulnerable to domain shift	Often outperforms traditional models on multi-domain/harmonized datasets	[46, 48]
Evidence and reporting	External validation, calibration curves, DCA, and stratified error analyses	Same requirements; additionally adhere to TRIPOD+AI for transparent reporting	[28, 49, 51]

collaboration between humans and AI such as AI triaged worklists and structured reports that state uncertainty explicitly [68, 69]. Finally, for each imaging AI task including detection, risk stratification, staging, invasiveness, and risk of metastasis, studies should prespecify outcomes and thresholds, report calibration including flexible calibration plots and DCA also called DCA, provide subgroup analyses by vendor, protocol, and key demographic and clinical factors, and release IBSI conformant parameters together with reference implementations to enable independent verification [70].

Heatmaps and saliency maps used to explain nodule detection or malignancy risk do not provide evidence of model reasoning unless validated [71]. These maps can remain stable when labels or weights are randomized, highlight scanner-specific artifacts such as kernel edges, beam-hardening, or bed/table contours, or change significantly with minor preprocessing adjustments [72]. For each imaging task, we recommend reporting the following: sanity checks where saliency should collapse with label or weight randomization, faithfulness metrics such as deletion or insertion curves or ROAR/IROF showing monotonic performance loss when removing important regions compared to matched controls, stability across seeds, augmentations, and scanners, and clinical alignment through concept-level tests such as spiculation or necrosis masks with effect sizes and confidence intervals [73]. Explanations should be accompanied by uncertainty and selective deferral, especially near decision thresholds used for escalation pathways. They must also be evaluated for automation bias in reader studies [74].

AI in pathological subtyping and biomarker inference

Digital pathology is another frontier where AI is transforming the diagnosis of NSCLC [15, 20]. Traditionally, histologic subtyping and the identification of molecular alterations require meticulous microscopy and multiple ancillary assays, such as IHC for TTF-1 and p40 and fluorescence in situ hybridization or sequencing for genomic alterations [75]. AI analyzes whole slide images (WSI) of tumor tissue to automate and augment these tasks. Deep learning models trained on labeled H&E slides accurately classify NSCLC histopathology, distinguishing adenocarcinoma from squamous carcinoma and other subtypes, with performance comparable to expert pathologists. For example, a 2020 study reported accuracy of approximately 0.95 for differentiating adenocarcinoma from squamous carcinoma on WSIs [76]. WSI pipelines use MIL with attention pooling. Tile level embeddings are aggregated via self attention or gated attention to produce slide level predictions. Recent variants apply token based Transformers to tile sequences with two dimensional positional encodings. We record the reported layer count, number of attention heads, and whether frequency aware attention such as FA or WA, or stain aware modules, was used to improve robustness. For IHC quantification, lightweight Vision Transformer heads operating on nucleus or patch tokens are increasingly used to standardize the PD-L1 TPS [29]. A particularly promising application is molecular virtual staining, namely biomarker prediction directly from routine H&E histology [20, 77]. Building on foundational work such as that by Coudray and colleagues, who reported an AUROC of approximately 0.82 for EGFR prediction from H&E, Campanella and colleagues in 2025 developed EAGLE, a refined pathology AI that was prospectively evaluated in a real world setting for EGFR prescreening [22]. In Nature Medicine, EAGLE was trained on more than 5000 digitized biopsies and achieved internal and external AUROC values of approximately 0.85 and 0.87, which translated into approximately 43% fewer reflex molecular tests while maintaining sensitivity. In a prospective deployment simulated study, it achieved an AUROC of 0.89 on new cases and substantially reduced biomarker reporting time [22]. This approach conserves tissue for comprehensive sequencing and shortens the turnaround time for initiating targeted therapy [22]. Similar AI approaches are being explored to predict other actionable alterations, such as anaplastic lymphoma kinase (ALK) or ROS proto oncogene 1 (ROS1) fusions, from morphology. However, their rarity requires larger training datasets [78] (Table 2).

AI-based analysis of immunohistochemical slides yields more consistent protein-biomarker quantification than manual scoring [21, 80]. For example, in 2022 Wu and colleagues developed a deep learning system to score PD-L1 IHC in NSCLC. The model's tumor-proportion scores showed strong agreement with pathologists, with a correlation coefficient of about 0.94, and improved inter-pathologist consistency [21]. Overall, AI in pathology enables precise subtyping of NSCLC by extracting detailed phenotypic information from routine slides [79]. These systems automate tumor classification, identify diagnostically relevant regions for review, and predict molecular markers without invasive procedures. Importantly, these tools are designed to support human experts rather than replace them [81]. With appropriate validation and regulatory approval, AI-augmented pathology can standardize diagnoses across centers and ensure accurate histologic and molecular characterization for each patient, which are foundations of personalized NSCLC care [15, 22].

Notwithstanding these advances, several limitations and sources of bias warrant emphasis.Across H&E based biomarker studies, external validation is inconsistent and subgroup calibration is seldom reported [82]. Many studies do not enforce patient level splits that separate sites and scanners at the partition stage, increasing the risk

Table 2 NSCLC digital pathology task overview

Subdomain	Input modality	External validation (centers/batches/scanners)	Clinical use/potential impact	Key references
Histologic subtyping	H&E whole-slide images (WSI)	Reported in some studies; details NR	Standardized subtyping; reduced subjective variability	[15, 20, 76, 79]
Molecular biomarker prediction	H&E WSI	Yes (multicenter, > 1 scanner; batch differences present)	~43% reduction in reflex molecular tests; tissue conservation; shorter TAT	[20, 22, 77]
Prediction of actionable fusions (ALK/ROS1) from morphology	H&E WSI	Larger-scale validation pending (NR)	Prescreening to optimize molecular testing	[78]
IHC quantification	IHC slides	Some cross-center/reader work (NR)	Standardized IHC scoring; supports immunotherapy decision-making	[21, 80]
Clinical integration and expert support	WSI/IHC plus reporting systems	Real-world deployment progressing (NR)	Provides explainable evidence to augment, not replace, experts	[28, 49, 51]

of leakage from acquisition related signatures [83]. Few reports include grayscale or stain normalized ablations, per site performance with confidence intervals, or negative region controls. Saliency maps are often presented without validity checks or quantitative evaluation. We recommend patient level splits that separate sites, explicit per site calibration, color and stain ablations, tumor only masking analyses, and stability tests across random initializations. These reporting elements help distinguish genuine morphologic associations from confounding factors [84].

In WSI pipelines, tile-level heatmaps often co-localize with staining or batch signatures, or tissue processing borders, instead of tumor morphology [85]. Without color or stain ablations, tumor-only masking, and cross-scanner stability tests, saliency can be misleading [86]. We recommend reporting the following: (i) color-space or stain normalization ablations with effect size and confidence intervals; (ii) negative-region controls, such as background and artifacts; (iii) concept-based validation, such as TCAV for gland formation, keratinization, or TIL density with bootstrap-tested significance; and (iv) slide-level counterfactuals, such as swapping stain or morphology exemplars, to test causal relevance of the explanation. Incorporate prototype-based or concept-bottleneck heads when feasible to improve faithfulness and auditability.

Role of AI in prognosis prediction and risk assessment

Integration and analysis of multiomics data

Rationale and data types

Prognosis improves when multiomics data that comprise genomics with mutations and copy number alterations, transcriptomics, methylomics, proteomics, and metabolomics are integrated with ctDNA, image-derived features from radiology and pathology such as radiomics and pathomics, and routine clinical variables [26, 87–89]. AI enables representation learning and discovery of cross modal interactions beyond linear additivity. These advances yield composite risk scores that better capture tumor biology and the host context [48].

Model classes and fusion strategies

Model classes and fusion strategies are summarized as follows. Approaches include penalized Cox models with learned embeddings [90]; deep survival models such as DeepSurv and DeepHit, Transformer based fusion for variable length longitudinal sequences [91]; and GNNs that capture pathway and interaction structure. Fusion can occur at the feature level, known as early fusion; at the decision level, known as late fusion; or at intermediate layers via cross attention [9, 91]. Pathway aware regularization improves interpretability by aligning learned representations with biological circuits [92]. To support model selection and reproduction, we summarize core architectural choices, hyperparameter sensitivity, and compute trade offs across these model classes [93]; see Table 3. Transformer based fusion requires careful specification of tokenization, sequence length and windowing, the number of layers and heads, the hidden width, and the attention variant. Training cost and inference latency scale with the token count and the network depth, which can constrain deployment in resource limited settings [94]. Practical mitigations include parameter efficient fine tuning with adapters or low rank adaptation, mixed precision training, quantization, pruning, and knowledge distillation [95]. For GNNs, the choice among convolutional families such as GCN, GraphSAGE, and GAT interacts with graph construction and neighborhood size, and excessive smoothing or excessive squashing can degrade performance [96]. For whole slide image pipelines, tile size and stride, stain normalization, and the attention pooling strategy strongly affect both accuracy and throughput. Across all classes, the most sensitive hyperparameters typically include the learning rate,

Table 3 Multimodal AI for NSCLC: from diagnosis to personalized therapy

Model class	Inputs	Objective	Limitations	Data/compute demand	Key hyperparameters and implementation notes	Key refs
Penalized Cox with learned embeddings	Hand-crafted and learned representations	Time-to-event	Linear risk composition unless embeddings capture non-linearity	Low to moderate; CPU or single GPU	Regularization strength, embedding size, feature scaling; report convergence and proportional hazards checks	[90]
Deep survival models (DeepSurv, DeepHit)	Images/WSI/omics and clinical	Hazard/risk or discrete-time event probability	Calibration and transportability require care	Moderate; single to few GPUs	Learning rate and batch size, early stopping, label discretization for discrete-time, augmentation; report time-dependent AUROC and IBS	[91]
Transformer-based fusion	Longitudinal imaging, ctDNA, labs, notes	Sequence modeling of risk over time	Data-hungry; careful masking/temporal encoding needed	High; scales with sequence length and depth; memory bound	Token size and stride, sequence length and windowing, layers and heads, hidden size, attention variant; adapters or LoRA for efficient tuning; mixed precision and quantization for deployment	[91]
GNNs	Omics graphs; patient–feature graphs	Risk via structured message passing	Graph construction choices matter; scalability	Moderate; depends on graph density	Aggregation type, neighborhood size, number of layers; avoid oversmoothing; document graph build rules	[92]
Fusion strategies	—	—	Early needs harmonized preprocessing; Late may miss interactions; Intermediate more complex	Varies	Decision-level vs cross-attention; report latency and memory impact of fusion block	[9, 91]

batch size, weight decay, data augmentation, and early stopping criteria. We recommend reporting wall clock training time, GPU type and count, peak memory footprint, and per case inference time, and releasing exact configuration files and fixed random seeds to enable deterministic reruns [97].

Recent exemplars and effect sizes

Multimodal survival models that combine CT radiomics, WSI embeddings, mutational signatures, and clinical covariates typically increase the concordance index by 0.05 to 0.12 over clinical baselines and lower the integrated Brier score. External validations show good portability with calibration drift that is usually modest and amenable to recalibration [26, 51, 98, 99]. Early stage I to II studies that combine genomics and pathomics have identified high risk subgroups that gain an absolute 5% to 10% overall survival benefit from adjuvant therapy, whereas low risk groups may be candidates for de escalation of therapy [76, 100, 101].

Dynamic and longitudinal risk

Landmarking and joint models update risk after each assessment, including ctDNA kinetics, radiographic response, and laboratory trends. These approaches outperform static baselines on time dependent area under the curve and enable earlier escalation or de escalation of therapy [102–106]. Reinforcement learning policies for adaptive sequencing are promising but require prospective oversight and clearly defined safety constraints [107–109].

Scale, privacy, and fairness

Multi institutional training is essential to capture real world variability [110]. Federated learning (FL) enables training across sites without moving raw data, while secure aggregation, client side differential privacy, and auditable logs protect confidentiality [111–113]. Fairness audits should be routine, with subgroup calibration and utility metrics such as calibration within groups and equalized odds, and any remediation, including reweighting or adversarial debiasing, should be documented [79, 114].

Reporting checklist (prognosis)

Studies should: (1) pre register analysis plans; (2) use internal and external validation across sites and time;(3) report calibration plots and metrics, including area under the curve, concordance index, integrated Brier score, and DCA, with time dependent estimates where relevant; (4) account for competing risks; (5) report subgroup performance by sex, race or ethnicity, geography, and smoking status; and (6) release code, parameter files, and data dictionaries to enable reproducibility.

Multimodal survival prediction models

AI driven prognostic models for NSCLC typically output a predicted survival time or a patient specific risk score. These outputs guide clinical decisions, such as whether to add adjuvant chemotherapy after surgery and how intensively to follow a patient [105]. Traditional models, such as tumor node metastasis staging, consider only a few variables, whereas AI models can integrate many features from imaging, pathology, and omics [26]. As a

result, AI models demonstrate improved discrimination for survival stratification [100].

For example, image based AI has identified imaging features that correlate with outcomes independent of stage [115, 116]. In a pilot study, a deep learning model applied to pre treatment CT scans predicted overall survival in NSCLC, and the model derived risk scores separated patients into distinct prognostic groups. In that study, the AUROC was approximately 0.70, outperforming a model that used only clinical factors, which achieved approximately 0.60 [117]. Similarly, AI derived pathomic features from H&E slides have prognostic value. A 2023 study reported five year survival prediction with area under the curve values between 0.64 and 0.85, exceeding models based on tumor grade or stage alone [118]. Together, these approaches enable a digital prognostic assay built from routine diagnostic data [116]. Combining radiology based and pathology based AI predictors with clinical variables yields integrated survival models. Several institutions are evaluating AI based survival nomograms for NSCLC that output individualized survival probabilities. In 2023, Song and colleagues developed a nomogram that combined a deep learning radiomics signature derived from CT with clinicopathologic variables to predict progression free survival in stage IV EGFR mutant NSCLC treated with EGFR inhibitors. The model improved one year progression risk stratification [119].

Beyond static baselines, AI models can provide dynamic prognostic updates [103]. For example, serial imaging combined with ML can assess response trajectories and adjust survival predictions; this approach is known as dynamic risk prediction [40]. In practice, dynamic risk is modeled with sequence Transformers that consume tokens indexed by event or by visit, together with temporal and frequency attention [29]. Fourier layers capture long range periodicity, and wavelet blocks capture abrupt transients induced by treatment regimens [30–32]. Key hyperparameters to report include sequence length and windowing, the time embedding scheme, which may be absolute or relative, the number of layers and attention heads, and the masking strategy for irregular sampling.Multiple time point radiomic modeling in lung cancer shows that changes in tumor texture or other features after a few therapy cycles predict long term outcomes better than baseline features alone [120]. In addition, multi omics prognostic models have been applied in specific contexts, such as early stage NSCLC after surgery [89]. One AI model integrated gene expression profiles with clinical factors to predict which patients would benefit from adjuvant chemotherapy. It identified a subset of stage I patients at high risk of recurrence who experienced significantly improved survival with chemotherapy, illustrating the potential of AI for decision making based on prognostic stratification [89, 121].

In summary, AI-based survival-prediction models—especially those built on multimodal data—are achieving higher accuracy and finer risk discrimination in NSCLC [26]. They hold promise for personalized risk assessments that inform patient counseling and rational treatment tailoring [38]. However, prospective validation remains essential [51]. Many published models still require rigorous external validation across diverse cohorts to ensure generalizability beyond their training sets [122]. As these models mature, they could be incorporated into practice via decision-support systems that, for example, flag a "high-risk" early-stage patient for closer follow-up or prompt consideration of novel adjuvant therapies [123] (Fig. 2). Despite these advances, important limitations and potential sources of bias remain. First, many studies rely on internal validation and lack temporally or geographically external test cohorts, and the reporting of calibration and DCA is inconsistent [124]. Second, patient level splits that separate sites and scanners are often missing, which can allow acquisition related signatures to inflate performance [83]. Third, dynamic models can inadvertently leak post baseline information into the target or comparator, introducing immortal time bias and overly optimistic estimates. Fourth, approaches to censoring, competing risks, and treatment switching vary widely, and concordance alone does not capture calibration over time. Finally, subgroup calibration by sex, age, ancestry, stage, and site is rarely presented, fairness analyses are uncommon, and most reports omit drift monitoring and change control plans [125].

AI in treatment decisionmaking and personalised therapy

AI for treatment decision making in NSCLC is moving beyond proof of concept and toward clinically consequential systems. These models integrate high dimensional, multimodal, and longitudinal data from radiology, pathology, multi omics and ctDNA, and electronic health records, producing calibrated, patient specific inferences that humans cannot reliably integrate at scale. Foundation and multimodal models learn unified representations that support response prediction and toxicity forecasting across therapies. Development and reporting should follow contemporary guidance on external validation, calibration, DCA, and deployment readiness. Critically, models should update risk over time by using serial imaging, circulating biomarkers such as ctDNA kinetics, and real world data streams. This dynamic updating better reflects disease trajectories and treatment effects in practice. When estimating individualized benefit, causal ML approaches can complement prediction to guide treatment escalation or deescalation.

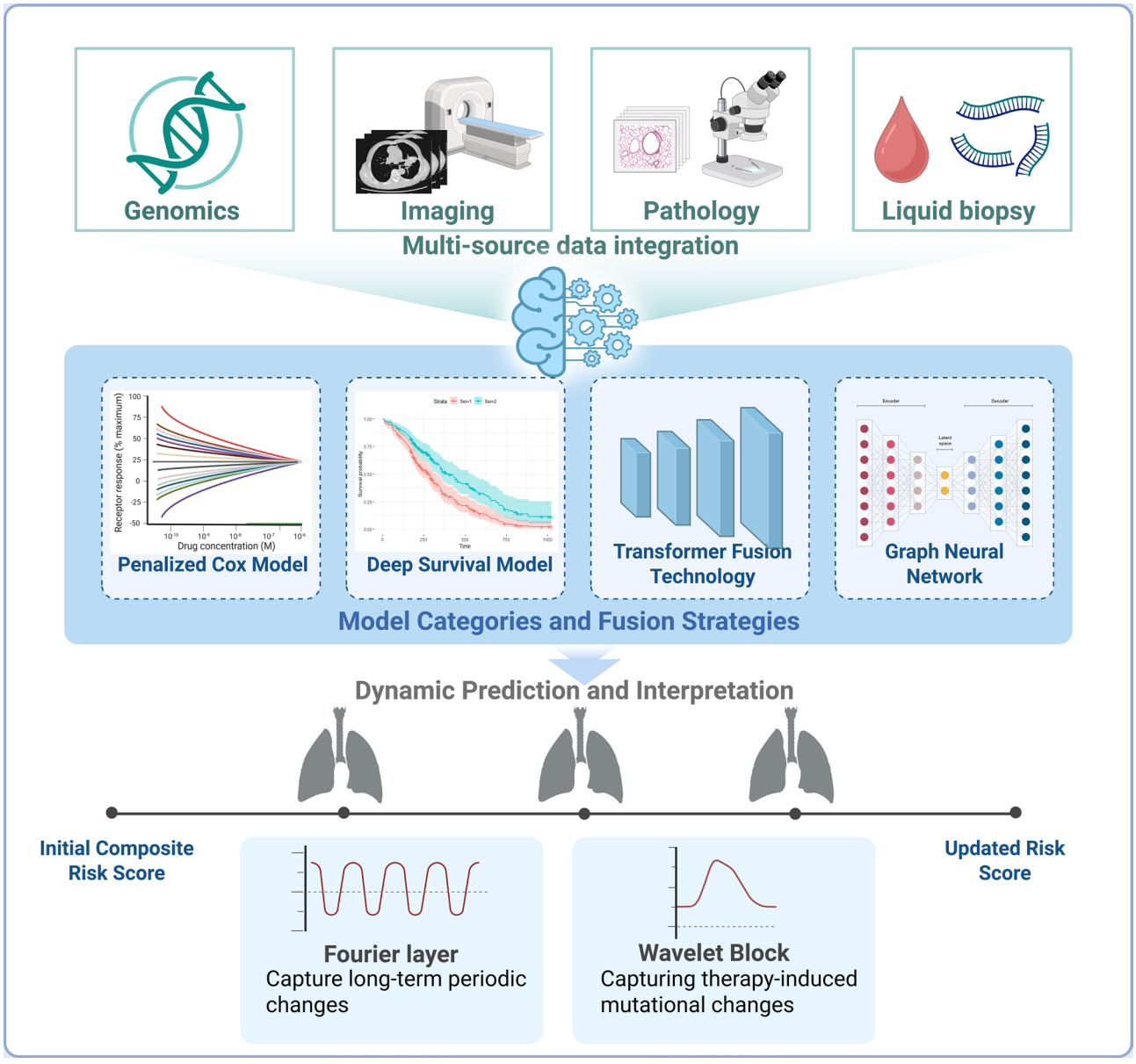

Fig. 2 Multimodal data integration and fusion models for dynamic risk prediction in NSCLC

Reinforcement learning based adaptation is promising but requires explicit safety guardrails and prospective oversight. Against this backdrop, section "Multimodal survival prediction models" summarizes evidence for response and toxicity prediction across radiotherapy, chemotherapy, targeted therapy, and immunotherapy, and highlights actionable operating points and clinical utility;see Fig. 3. section "Response and toxicity prediction across modalities" reviews AI in drug discovery and virtual screening, where generative and structure aware approaches, for example AlphaFold supported pipelines, are accelerating target identification and lead optimization for NSCLC; see Fig. 4.

Response and toxicity prediction across modalities

Radiotherapy

AI improves radiotherapy planning and predicts which patients are likely to benefit from radiation or experience treatment related harm [126]. For planning, recent methodological reviews advocate standardized and externally validated segmentation pipelines with transparent reporting of architectures, key hyperparameters and computational constraints, which enables reliable auto contouring and more consistent plans across sites, including resource limited centers [127–129]. For response prediction, radiomics based models identify imaging features associated with tumor radiosensitivity. For example, radiomic patterns on pre treatment CT have

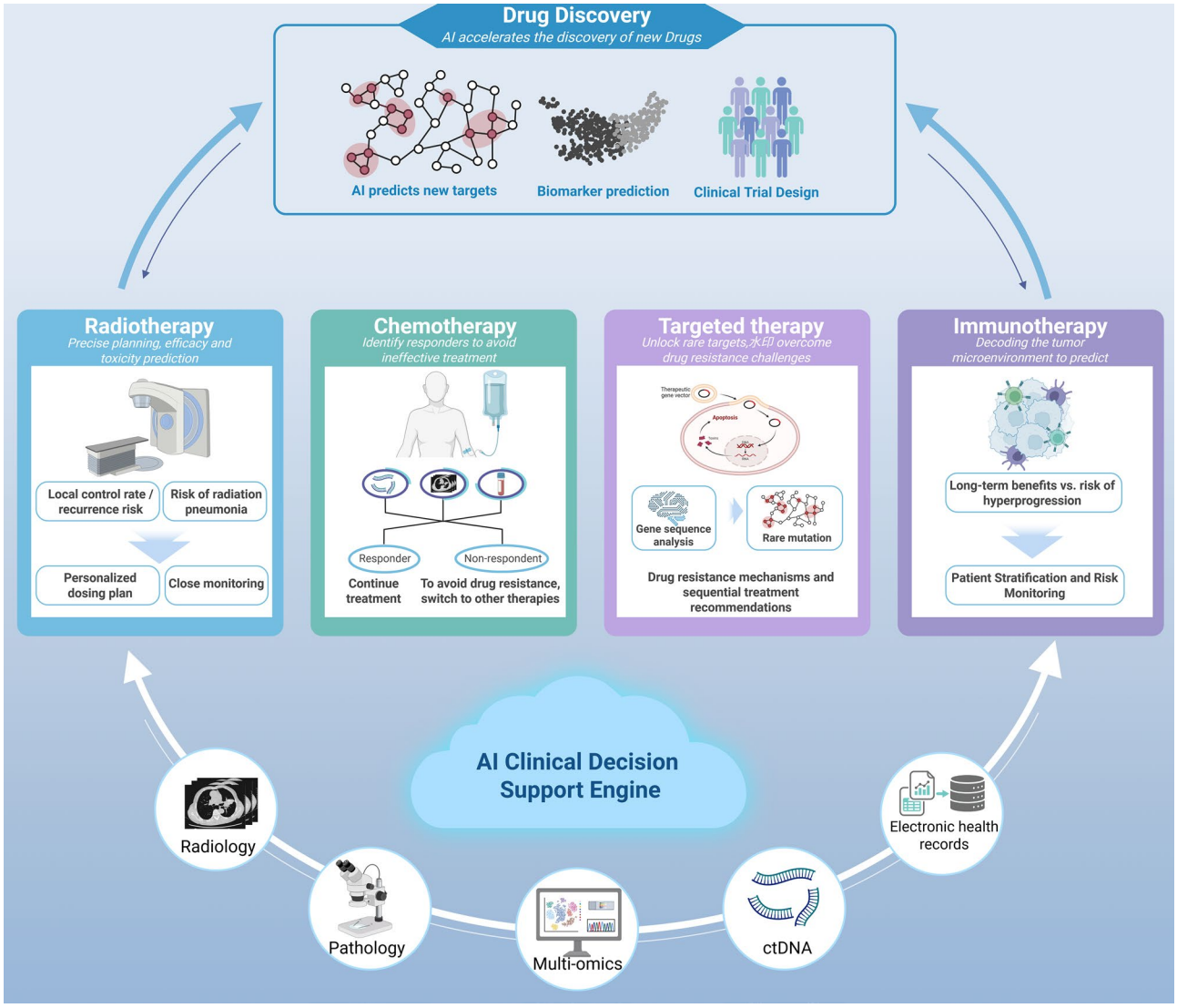

Fig. 3 From multimodal evidence to action: AI guidance for radiotherapy, chemotherapy, targeted therapy, immunotherapy, and drug discovery in NSCLC

been linked to post radiotherapy local control and recurrence risk in NSCLC [130]. In one study, a radiomics signature predicted two year local recurrence after definitive chemoradiation with an AUROC of approximately 0.75, outperforming traditional stage based estimates [131]. AI models have also been developed to forecast radiation induced toxicities [38]. A recent multi institutional deep learning model combined dosiomics, radiomics, and clinical data to predict grade two or higher radiation pneumonitis in locally advanced NSCLC [132]; The integrated model achieved an external validation AUROC of approximately 0.80 and yielded well calibrated risk estimates. Similarly, ML models predict radiation related cardiac toxicity and pulmonary fibrosis by analyzing pre treatment scans together with dose volume parameters [133]. These tools could enable personalized radiotherapy by identifying patients with a high predicted pneumonitis risk who may benefit from alternative approaches or closer monitoring [134].

Chemotherapy

Predicting chemotherapy response in NSCLC remains challenging, yet AI is increasingly used to identify biomarkers of chemosensitivity [135]. ML models that use genomic signatures and circulating biomarkers have been explored to predict which tumors will respond to platinum based regimens [136]. Gene expression based chemotherapy response scores show potential for predicting response to neoadjuvant chemotherapy in resectable NSCLC and may help avoid futile treatment in non responders [121]. Radiomics has also been evaluated. A deep learning model using radiomic features

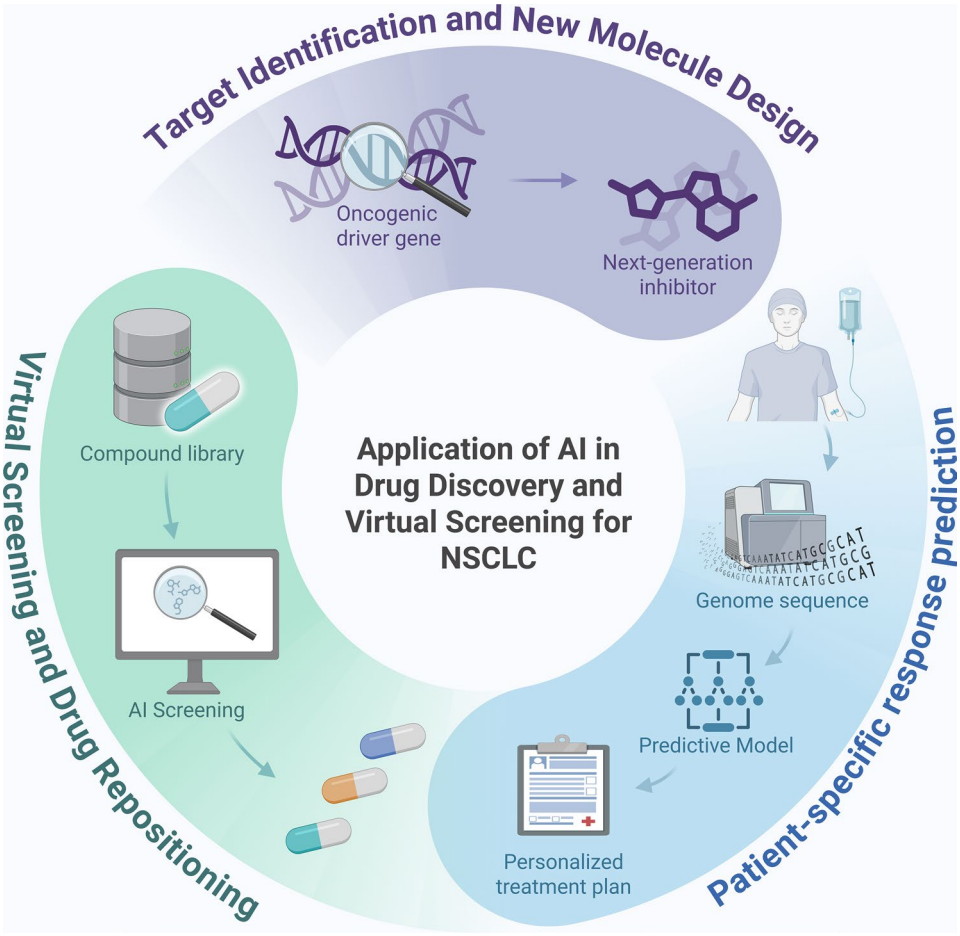

Fig. 4 AI in NSCLC drug discovery: from target identification to patient-specific response prediction

distinguished responders from non responders to first line chemotherapy on baseline CT with modest accuracy of approximately 70% [137]. In addition, AI analysis of blood based biomarkers, for example serum N glycome changes, has been investigated to forecast efficacy [138]. Although no AI test for chemotherapy response is in routine clinical use, these early studies suggest tools that could guide selection between intensive chemotherapy and alternative treatments for individual patients [139].

Targeted therapy (EGFR/ALK inhibitors)

In oncogene driven NSCLC, the key question is not only whether a tumor harbors a targetable alteration but also how durable the treatment response will be [140]. AI models are being developed to predict outcomes with targeted therapies and to identify early in treatment which patients may need additional interventions. Among patients with EGFR mutant NSCLC who start EGFR tyrosine kinase inhibitors, only a subset achieve prolonged progression free survival, whereas others progress rapidly because of de novo resistance [141]. Two recent studies applied ML to baseline clinical and imaging data to predict short progression free survival with EGFR tyrosine kinase inhibitors, thereby flagging high risk patients who might benefit from upfront combination strategies [142]. In these studies, the models identified patients at high risk of early progression within six to nine months with reasonable accuracy. These predictions are clinically actionable. For a patient predicted to respond poorly to EGFR tyrosine kinase inhibitor monotherapy, an oncologist might add chemotherapy or a vascular endothelial growth factor inhibitor at treatment initiation, an approach that can improve outcomes but may increase toxicity [143]. Similarly, in ALK positive NSCLC, where multiple ALK inhibitors are available, AI models are being explored to predict which specific inhibitor a tumor is most likely to respond to based on biological differences derived from omics data [144, 145].

Immunotherapy

Because only about 20 to 30% of unselected NSCLC patients respond to immune checkpoint inhibitors, identifying reliable predictive biomarkers is essential l [146]. AI has been used to discover new biomarkers and to

integrate multiple signals into more accurate composite predictors [38]. In computational pathology, AI models analyze the spatial organization of cells and immune infiltrates on H&E slides to infer tumor immune phenotypes [147]. Rakaee et al. developed Deep-IO, a deep learning model that predicts immune checkpoint inhibitor outcomes from pre treatment H&E slides. In a cohort of 958 patients with advanced NSCLC, the model score independently predicted response and survival and either outperformed or complemented PD-L1 and tumor mutational burden, with an area under the curve of 0.66 for objective response in external validation compared with 0.62 for PD-L1 at at least 50%. Combining the AI score with PD-L1 increased the area under the curve to 0.70 and identified responders with higher precision [148]. Radiomic profiling provides a complementary approach. CT features can distinguish hyperprogressive disease from durable benefit on programmed cell death 1 or PD-L1 inhibitors, and one classifier separated hyperprogressors from ordinary progressors with an area under the curve of about 0.87 [38]. Radiomic biomarkers that reflect heterogeneity or volume dynamics have also correlated with outcomes and with immune related adverse events. For example, a baseline CT signature predicted the risk of immunotherapy induced pneumonitis and enabled closer monitoring [149]. Multimodal models that integrate radiomics, PD-L1 expression, and ctDNA metrics have outperformed individual predictors for one year survival on immunotherapy [26]. By tracking temporal changes during treatment, deep learning models can detect response or progression earlier than the Response Evaluation Criteria in Solid Tumors and thereby prompt earlier treatment changes [102]. Overall, AI driven response prediction across modalities is moving NSCLC care toward personalized therapy by matching each patient to the strategy most likely to maximize benefit and minimize harm. Although many models remain experimental, some pathology based immunotherapy predictors are undergoing prospective evaluation with encouraging early results [148, 150].

AI in drug discovery and virtual screening

Beyond clinical decision support, AI is reshaping the early phases of NSCLC drug development [38]. AI driven discovery applies computational algorithms to identify therapeutic targets, design novel molecules, and repurpose approved agents, often much faster than traditional laboratory screening [151]. NSCLC exhibits numerous genomic alterations and resistance mechanisms, creating many opportunities for new therapies; however, efficient discovery remains challenging. ML models, including deep generative approaches, learn from large chemical libraries and bioassay data to predict compounds that inhibit cancer targets or overcome resistance [152].

One important application is the discovery of next generation inhibitors against established oncogenic drivers [152]. Resistance to third generation EGFR tyrosine kinase inhibitors, such as osimertinib, often emerges through EGFR T790M and C797S mutations [153]. In 2024, Zhou and colleagues used a ML aided approach to identify CDDO-Me as a potential fourth generation EGFR inhibitor active against T790M mutant NSCLC. The model screened hundreds of candidates by learning structure–activity relationships, and it predicted CDDO-Me, which was later confirmed experimentally, to strongly suppress proliferation of T790M mutant NSCLC cells, including in xenograft models. This approach markedly narrowed the pool of candidates for laboratory testing, demonstrating how AI can accelerate lead compound discovery [152]. Similarly, in 2023, Zhang and colleagues used ML with support vector machines and random forests to design new small molecules targeting EGFR active site mutations, achieving external accuracy greater than 95% and guiding feature optimization, with R^2 approximately 0.93 between predicted and experimental activity. These AI designed compounds are now candidates for further preclinical development [154].

AI is used for virtual screening of existing libraries to identify repurposing opportunities in NSCLC [155]. ML algorithms predict drug target interactions and synergistic combinations by mining patterns in historical pharmacologic data [156]. For example, a platform analyzing transcriptomic profiles predicted that a Food and Drug Administration approved kinase inhibitor, not originally indicated for lung cancer, could have activity in KRAS mutant NSCLC; subsequent laboratory testing confirmed this prediction and led to a new clinical trial [157]. Additionally, deep learning models such as GNNs operating on molecular graphs have been used to predict compounds that may inhibit novel targets, for example KRAS G12C and MET exon 14 skipping, thereby guiding medicinal chemistry efforts [158]. In silico approaches also extend to immunotherapy, with neural networks proposing small molecules or peptides that modulate immune checkpoints or the tumor microenvironment [159].

Another emerging paradigm is the prediction of patient specific drug response enabled by AI [157]. Large cell line screens, for example profiling hundreds of compounds across dozens of lung cancer lines, generate data that AI uses to map genomic profiles to likely drug responses [160]. For example, the open source tool D3EGFR provides a deep learning based web server that predicts the sensitivity of EGFR mutant lung cancers to various tyrosine kinase inhibitors and their combinations [161]. As more cell line and organoid screening data accumulate, AI can match NSCLC tumors to optimal therapies, effectively conducting in silico clinical trials to prioritize

treatments. This approach is particularly relevant for rare molecular subsets in which real world trials are difficult to conduct [157].

In summary, AI is accelerating NSCLC drug discovery on several fronts: identifying new candidates such as small molecule and biologic agents for resistance pathways; optimizing leads by predicting how structural changes affect activity and by using reinforcement learning to generate improved analogs; and repurposing or prioritizing agents for defined patient subgroups [162]. The success of AI discovered agents such as CDDO Me against EGFR T790M suggests a future in which AI augmented medicinal chemistry markedly shortens the timeline from target identification to effective therapy [154]. When coupled with rigorous wet laboratory validation, these approaches could expand the therapeutic arsenal for NSCLC, particularly for patients who have exhausted current options [162]. Close collaboration among computational scientists, chemists, and oncologists is essential to ensure that AI predictions are rigorously validated and translated into clinically viable drugs [139]. (Table 4) Despite this progress, important limitations and methodological conflicts remain. First, many studies use random or scaffold split validation, which allows near duplicate chemotypes to enter the test set and inflates performance; temporal splits and external assays from independent sources are needed [163]. Second, assay heterogeneity, batch effects, and differences in experimental conditions can confound activity labels, and few reports include orthogonal confirmation across multiple assay formats [164]. Third, generative models may propose molecules that are difficult to summarize, unstable, or outside the applicability domain; synthesizability metrics, retrosynthesis success rates, and medicinal chemistry review are rarely reported [165]. Fourth, docking and scoring functions used to label data or triage candidates are noisy and can bias learning, and enrichment should be benchmarked against strong baselines with decoys and property matched controls. Fifth, translation from cell lines to patients is limited by context mismatch, off target effects, and absorption, distribution, metabolism, and excretion and toxicity constraints; prospective, blinded screens and pre registered evaluations are uncommon. Sixth, uncertainty and calibration are seldom quantified, and studies frequently omit hit rate, precision at top k, and prospective enrichment metrics that would inform decision making [166].

Table 4 Prospective-facing summary of AI for NSCLC drug discovery and virtual screening

Study (first author, year)	Task/Goal	Key outcomes (quantitative)	Prospective/translational status	Key refs
Zhou, 2024	Identify 4th-gen EGFR inhibitor active against resistance	CDDO-Me suppressed T790M-mutant NSCLC growth (exact IC50/TGI NR)	Preclinical; candidate for further development	[152, 154]
Abramson (AlphaFold3), 2024	Complex structure prediction to inform docking/design	SOTA complex prediction (platform)	Integrated in pipelines; preclinical utility	[151]
Meller (PocketMiner), 2023	Predict cryptic binding pockets	Improved cryptic pocket detection	Design aid; upstream of med-chem	[158]
STTT review + platform exemplar	Repurpose FDA-approved kinase inhibitor for KRAS-mutant NSCLC	Activity confirmed preclinically	Prospective clinical evaluation underway	[155, 157]
Zhao, 2025	Evidential DL for drug-target interactions	Calibrated DTI improvement	Platform tool; supports prioritization	[156]

Challenges, limitations, and ethics

Data bias and quality

Heterogeneity and bias

AI performance in NSCLC depends strongly on data provenance [167]. Variation in CT acquisition parameters such as tube voltage kVp, current time product mAs, slice thickness, reconstruction kernel, and vendor, and differences in PET protocols such as uptake time and calibration, introduce batch effects that distort learned representations.Digital pathology preanalytic factors including fixation, processing, staining, and scanner optics and compression have similar effects [168]. Labels are often noisy because of interreader variability and evolving diagnostic criteria, and class imbalance caused by rare histologies and uncommon fusions predisposes models to majority class bias and to poorer performance in minority subgroups [169].

Governance and documentation

Robust governance should include dataset datasheets and provenance logs; capture of preanalytic metadata for pathology, including ischemic time, fixation, stain, scanner model, and International Color Consortium (ICC) profile, and for imaging, including Digital Imaging and Communications in Medicine (DICOM) tags, dose, and reconstruction kernel; and versioned curation pipelines with auditable trails [170]. Such documentation enables reproducibility and supports forensic analysis after deployment [171].

Standardization and harmonization

For radiomics, adhere to IBSI definitions and publish exact parameters and code; apply physics aware harmonization such as ComBat and kernel aware resampling,

and verify stability on repeat scan datasets [172]. For pathology, implement stain normalization, scanner aware augmentation, and color deconvolution, and confirm cross scanner reproducibility through ring trials [173].

Label quality and learning strategies

Use rater agreement protocols and adjudication panels for key endpoints such as PD-L1 TPS near cut points [169]. Leverage weak supervision using MIL, self supervised pretraining, and active learning to maximize label yield. Address class imbalance with cost sensitive losses, calibrated resampling, and careful evaluation in minority strata while avoiding synthetic oversampling artifacts in texture rich tasks [174].

Generalizability, OOD behavior, and robustness

Report internal to external performance gaps explicitly. Decreases of five to ten points in the AUROC across scanners or across sites are common [49]. Characterize distribution shift using stress tests and challenge sets,covering covariate shift,changes in label prevalence, and concept drift. Implement out of distribution (OOD) detection with Uncertainty Quantification (UQ) to enable selective prediction [175]. Maintain model cards that document failure modes, subgroup caveats, and guardrails. In addition, conduct external validation with a locked pipeline. Use patient level splits that keep sites and scanners separate, and apply preprocessing learned only on development data. Report discrimination, calibration, and decision utility together, and provide subgroup results by site, scanner, sex, age, ancestry, stage, and smoking status; see section "Validation, regulation, and workflow integration" for execution details [28]. To improve transportability, embed physics inspired modules that encode signal priors, including FA for global frequency structure, WA for multiscale edges and transients, and low rank tensor decomposition to control capacity. Publish transform settings, low rank factors, and domain shift ablations [30–32].

Privacy and security

Anticipate risks of model extraction, membership inference, and inversion. Where feasible, use FL with secure aggregation and consider differential privacy for sensitive modalities such as whole slide imaging and electronic health records [176]. Perform threat modeling and penetration testing, and monitor for anomalous access and potential data exfiltration [177] (Table 5).

Systematic data pitfalls are pervasive in existing studies (Table 6). Common weaknesses include (i) single-center, small-sample cohorts that inflate internal discrimination but fail under domain shift [178]; (ii) subjective and inconsistent annotations, particularly at near-threshold endpoints such as PD-L1 at 1% and 50%, with minimal adjudication or measurement-error analysis; (iii) systemic case-mix and acquisition biases—including site/scanner fingerprints, stage distribution, smoking status, ancestry, and socioeconomic proxies—that models learn and amplify [179]; (iv) leakage risks (patient overlap across splits and refitting normalization on combined data) that overstate performance [180]; and (v) imbalanced outcomes (rare fusions and never-smoker subsets) that produce unstable thresholds and poor calibration in under-represented groups [181]. We recommend explicit internal–external reporting with locked pipelines; patient-level splits that keep sites and scanners separate;

Table 5 Reporting and governance checklist

Item	Minimum requirement for publication	Recommended best practice (deployment-ready)	Metrics/evidence to report	Key references
Data provenance and metadata	Dataset datasheet; key imaging/pathology metadata (DICOM tags, stain/scanner)	Full provenance logs; versioned curation scripts; audit trails	Data dictionary; inclusion/exclusion flow; preprocessing parameters	[170, 171]
Standardization and harmonization	IBSI-conformant radiomics definitions; code/parameters shared	Physics-aware harmonization (ComBat, kernel-aware resampling); pathology stain normalization and scanner-aware augmentation; ring trials	Test–retest stability; cross-scanner/site reproducibility	[50, 172, 173]
Label quality	Labeler count and roles; consensus rules	Adjudication panels for key endpoints; near-threshold protocols (e.g., PD-L1)	Inter-rater agreement (κ/ICC); sensitivity analyses to relabeling	[169]
Class imbalance and fairness	Class distributions; basic subgroup metrics	Cost-sensitive learning; calibrated resampling; fairness audit (by sex/ancestry/site/scanner)	Calibration-within-groups; equalized-odds/TPR gaps with CIs	[49, 169, 174]
External validity and shift	At least one external test	Internal–external validation across sites/time; stress tests; challenge sets	AUROC/C-index, Brier/ECE; reported internal–external gap (expect 5–10 points); coverage vs. accuracy under selective prediction	[49, 175]
Uncertainty and OOD	—	UQ (ensembles/MC-dropout/evidential); OOD detection; selective deferral policy	Calibration plots; risk-coverage curves; deferral utility/DECISION curves (DCA)	[175]
Privacy and security	Ethics approval; de-identification	FL + secure aggregation; differential privacy (where feasible); pen-testing	DP ε/δ (if used); federation topology; security test report; access monitoring	[176, 177]

Table 6 Data quality pitfalls, subgroup harms, and mitigation/reporting playbook

Pitfall	Typical manifestation in NSCLC AI	Potential clinical harm	Primary mitigation
Single-center, small-sample cohorts	High internal AUROC; drop on external sites/scanners	Mis-triage under domain shift; delayed or missed care	Multi-site curation; locked pipelines; internal–external validation
Inconsistent/subjective labels	Label noise; unstable thresholds; poor calibration	Overtreatment/undertreatment near cutoffs	Adjudication panels; near-threshold SOPs; κ/ICC tracking
Systemic case-mix bias	Subgroup TPR/FPR gaps; risk miscalibration	Disparate false negatives/positives; unequal benefit	Targeted sampling; reweighting; DRO; subgroup thresholds
Acquisition bias	Model keys on device/site signatures	Fragile transportability; scanner-specific failures	Physics-aware harmonization; stain normalization; ring trials
Class imbalance	Minority underperformance; unstable PPV/NPV	Missed rare actionable findings	Cost-sensitive loss; calibrated resampling; synthetic data with caution
Leakage	Inflated metrics; failure at deployment	Unsafe optimism	Patient-level splits; freeze preprocessing; audit trails
OOD/unseen shifts	Confidence mismatch; brittle predictions	Silent failures; unsafe automation	UQ + OOD detection; selective deferral; drift alarms
Privacy/security gaps	Data misuse; membership inference	Legal/ethical risk; loss of trust	FL+secure aggregation; DP where feasible; pen-testing

and joint reporting of discrimination, calibration, and decision-curve analysis, each stratified by site, scanner, sex, age, ancestry, stage, and smoking status [124].

Explainability and clinical trust

From saliency to semantics

Post-hoc saliency methods such as Gradient-weighted Class Activation Mapping, Integrated Gradients, and Layer-wise Relevance Propagation can highlight image regions that influence predictions; however, their faithfulness and stability remain limited without dedicated validation [182]. Where feasible, prioritize intrinsically interpretable designs, including concept-bottleneck models that detect clinical primitives such as spiculation, necrosis, and tumor-infiltrating lymphocyte density; generalized additive models with pairwise interactions (GA^2 M); prototype or nearest-neighbor reasoning; and case-retrieval systems that surface similar prior patients and outcomes [183].

Evaluating explanations

Quantify faithfulness, defined as sensitivity to counterfactual perturbations; quantify stability, defined as repeatability across runs; and quantify utility, assessed by whether clinicians make better decisions faster. Avoid explanation theater by pairing explanations with quantitative UQ and systematic error analyses [184]. In pathology, complement heatmaps with tile-level concept scores that align with pathologist vocabulary.

Uncertainty, calibration, and selective workflows

Distinguish aleatoric and epistemic uncertainty, and use deep ensembles, MC-dropout, evidential networks, or conformal prediction to provide well-calibrated confidence for each output [185]. Enable abstention in low-confidence or OOD cases, and route such cases to expert adjudication or confirmatory testing, for example reflex next-generation sequencing for equivocal EGFR prescreening and manual PD-L1 scoring near 1% or 50% [186].

Communication and documentation

Standardize report templates to list the model name and version, training domains, intended use, contraindications, uncertainty bins, and recommended actions. Maintain accessible factsheets and change logs for tumor boards and quality assurance committees [187].

Validation, regulation, and workflow integration

Validation ladder

Progress deliberately from technical validation that uses cross validation and internal and external splits, to clinical validation through multicenter retrospective studies with prespecified analysis plans, and ultimately to demonstrations of clinical utility through prospective evaluations such as Developmental and Exploratory Clinical Investigations of DEcision support systems driven (DECIDE)-AI style pilots, stepped wedge or cluster trials, and, where feasible, Randomized Controlled Trials [188]. At each stage, report discrimination, calibration, DCA, and net reclassification compared with standard care [28].

High retrospective accuracy is necessary but not sufficient for clinical adoption [189]. Translation requires evidence across technical, clinical, and operational domains. First, demonstrate generalizability and calibration under domain shift on external datasets that differ by time, geography, vendor, and protocol [190]. Beyond discrimination, report calibration slope near 1.0, expected calibration error with a prespecified tolerance, and decision-curve analysis at prespecified thresholds with net benefit and linked actions [189]. Provide subgroup

calibration by site, scanner, age, sex, race or ethnicity, stage, and smoking status, with confidence intervals [28]. Second, characterize uncertainty and OOD behavior and define selective output and deferral policies with predefined coverage and deferral triggers [191]. Route low-confidence or OOD cases to expert review or reflex molecular testing, and disclose the impact on net benefit and resource use [192]. Third, address workflow and interoperability through seamless integration with Radiology Information Systems and Picture Archiving and Communication Systems, DICOM Structured Reports and Segmentation, Fast Healthcare Interoperability Resources (FHIR), and Clinical Decision Support Hooks [193]. In shadow mode, quantify turnaround time, alert burden expressed as alerts per 100 cases, coverage–accuracy trade-offs, and re-review rates relative to standard procedures [194]. Fourth, set explicit service-level agreements and safety guardrails, including targets for end-to-end inference latency, failure rates, audit-log completeness, automatic escalation near decision thresholds such as PD-L1 1 and 50%, and human–AI collaboration with mandatory second confirmation for high-risk tasks such as emergency pulmonary embolism detection [195]. Fifth, manage lifecycle governance and change control under medical-device principles by defining drift triggers such as decreases in external AUROC beyond a preset margin or expected calibration error exceeding a threshold [196]. Establish recalibration and rollback plans with version auditing, monitor multicenter performance for adverse AI events, and track the balance between coverage, accuracy, and workload. Sixth, include fairness and economics in the evidence base by auditing in-group calibration and equalized-odds with confidence intervals, and by reporting time-to-treatment reductions, avoided unnecessary tests, budget impact, cost-effectiveness, and the sensitivity of net benefit to threshold choices [197]. Together, these standards enable a measurable transition from high-performing models to safe, effective, and efficient clinical tools [194].

Regulatory frameworks

For AI and ML based software as a medical device, align with the International Medical Device Regulators Forum risk frameworks; the United States Food and Drug Administration total product lifecycle principles and predetermined change control plans for learning systems; the European Union Medical Device Regulation and In Vitro Diagnostic Regulation and the EU AI Act; and the United Kingdom Medicines and Healthcare products Regulatory Agency change programme. Operate under a quality management system compliant with International Organization for Standardization 13485, risk management according to International Organization for Standardization 14971, software lifecycle standards International Electrotechnical Commission 62304 and International Electrotechnical Commission 82304–1, current cybersecurity guidance, and Good ML Practice [198]. Plan postmarket surveillance with real world performance monitoring and define explicit triggers for recalibration or rollback [199].

Clinical integration and interoperability

Embed AI into routine clinical systems and data pipelines, including DICOM Segmentation and Structured Report objects, Fast Healthcare Interoperability Resources Observation and DiagnosticReport resources, Clinical Decision Support (CDS) Hooks, and Health Level Seven International order and result messages. Define inference service level agreements, maintain auditable logs and user controls, and run shadow-mode pilots before go-live to quantify alert burden and false positive externalities [193].

MLOps in healthcare

Manage models as living systems. Maintain dataset and feature versioning; use gated continuous integration and continuous delivery deployments; run synthetic and real world regression tests; monitor for distribution shift; schedule recalibration and periodic reapproval; and maintain documented rollback plans. Define clear roles, responsibilities, and escalation paths for adverse AI events [58, 200].

General limitations and comparability across studies

Most multimodal fusion studies summarized here are retrospective and frequently single center, and they report only marginal gains over unimodal baselines, which calls into question clinical significance without net benefit analysis and action linked thresholds [28]. For example, a systematic review by Yu and colleagues of externally validated imaging models found that most algorithms performed worse on external datasets, highlighting the gap between internal discrimination and transportability [201]. In digital pathology, a 2024 meta analysis reported variable accuracy and frequent risk of bias, and a public audit of commercial products showed that only about forty percent had peer reviewed external validation, underscoring limited generalizability [82]. Even in promising cases such as the EAGLE pathology system for EGFR prescreening, internal and external AUROC values were about 0.85 and 0.87, and a prospective silent evaluation reached 0.89; however, these gains require translation into net benefit, explicit decision thresholds, and measurable reductions in time to treatment to establish clinical value [202]. Across prognostic models, reporting of calibration and DCA remains inconsistent, despite long standing guidance that clinical usefulness should be expressed as net benefit across

clinically relevant thresholds [203]. Methodological pitfalls also persist, including potential leakage of post baseline information in dynamic models that leads to immortal time bias, heterogeneous handling of censoring, competing risks, and treatment switching, and insufficient subgroup calibration by sex, age, ancestry, stage, and site. Finally, cross paper comparisons are often not directly comparable because datasets, endpoints, preprocessing, and evaluation metrics differ, which limits interpretation of reported incremental gains unless studies share harmonized benchmarks, analysis plans, and reporting checklists such as TRIPOD+AI and DECIDE AI. Together, these observations support prospective multicenter evaluations with preregistered analysis plans, external and temporal test cohorts, decision relevant metrics such as net benefit and reclassification, and routine subgroup and site specific calibration to demonstrate durable generalizability and practical clinical impact.

External validation and calibration

Execution. Predefine the full pipeline and keep it fixed during testing [204]. Use external cohorts that differ in geography and time and that span multiple vendors and protocols. Fit preprocessing only on development data and apply it unchanged to external data. Enforce patient level splits that separate sites and scanners. Report AUROC and area under the precision recall curve with 95% confidence intervals [205]. Report calibration in the large, the calibration slope and intercept, smooth calibration curves, and expected calibration error. Provide DCA across prespecified thresholds and report the net reduction in interventions at the chosen operating point. For survival outcomes, report the concordance index (C index), time dependent AUROC, and the integrated Brier score, and account for competing risks. Aim for at least 100 events and 100 non events for binary outcomes and at least 200 events for survival, or use internal external cross validation when event counts are limited. Follow TRIPOD plus AI and DECIDE AI for study planning and reporting [28].

Common pitfalls include tuning on the external set, refitting normalization on combined data, mixing patients across sites, altering class priors without recalibration, reporting discrimination without calibration or decision analysis, and omitting subgroup calibration by site, scanner, sex, age, ancestry, and stage. Preregistration, harmonized reporting templates, and routine stratification by site and subgroup help mitigate these risks [28].

Future outlook: integration, interpretability, and equity

Realizing the clinical promise of AI in NSCLC requires a shift from isolated single task models to an integrated clinician centered ecosystem that is scalable, interpretable, and equitable (Fig. 5). Foundation and cross modal Transformer backbones trained on diverse medical corpora can unify signals from imaging, pathology, multi omics and ctDNA together with electronic health records. The resulting shared representation supports NSCLC specific fine tuning for diagnosis, biomarker inference, risk trajectories, and treatment ranking. Beyond static baselines, temporal and patient centric modeling that links longitudinal imaging and ctDNA kinetics with digital twin simulations can update risk in real time and anticipate counterfactual treatment responses. Mechanism aware and causal methods align predictions with biology and estimate individualized benefit under confounding. Continual and FL keep models current as scanners, protocols, and therapies evolve while preserving privacy. Successful translation depends on clinician co design, usable explanations, and layered safety that includes uncertainty awareness and OOD aware abstention. It also depends on training in AI literacy and on routine evaluation of cognitive load and time to decision. Standards and equity are essential. Priorities include interoperability through the IBSI, DICOM Structured Report and Segmentation, FHIR, and Clinical Decision Support Hooks; benchmarking on open multicenter data sets; proactive fairness auditing with remediation; and pathways for deployment in low resource settings with clear consent and governance under emerging regulations.

Deeper multimodal fusion at scale

Foundation and transformer paradigms

The field is converging on foundation models trained on diverse medical corpora, including CT and positron emission tomography, WSI, clinical notes, and structured laboratory data, paired with cross modal Transformers that learn joint latent spaces [206]. Fine tuning for NSCLC can produce unified outputs such as diagnosis, biomarker inference, risk trajectories, and ranked treatment options from a shared backbone [15]. Self supervised and weakly supervised objectives, including masked modeling and contrastive pairing of image, omics, and text, reduce labeling burden and improve transfer across institutions [20]. Concretely, foundation models tuned for NSCLC typically comprise three components. The first is a modality specific tokenizer, for example three dimensional patch embedding for CT, a tile encoder for WSI, a gene set projector for omics, and a text encoder for clinical notes. The second is a shared Transformer with between twelve and forty eight layers and between twelve and twenty four attention heads, connected by cross attention bridges for intermediate fusion [29]. The third is a personalization layer using adapters and low rank adaptation (LoRA) for site specific adaptation. For longitudinal use, FA or WA blocks can be inserted into temporal layers to couple slow trends and abrupt shifts

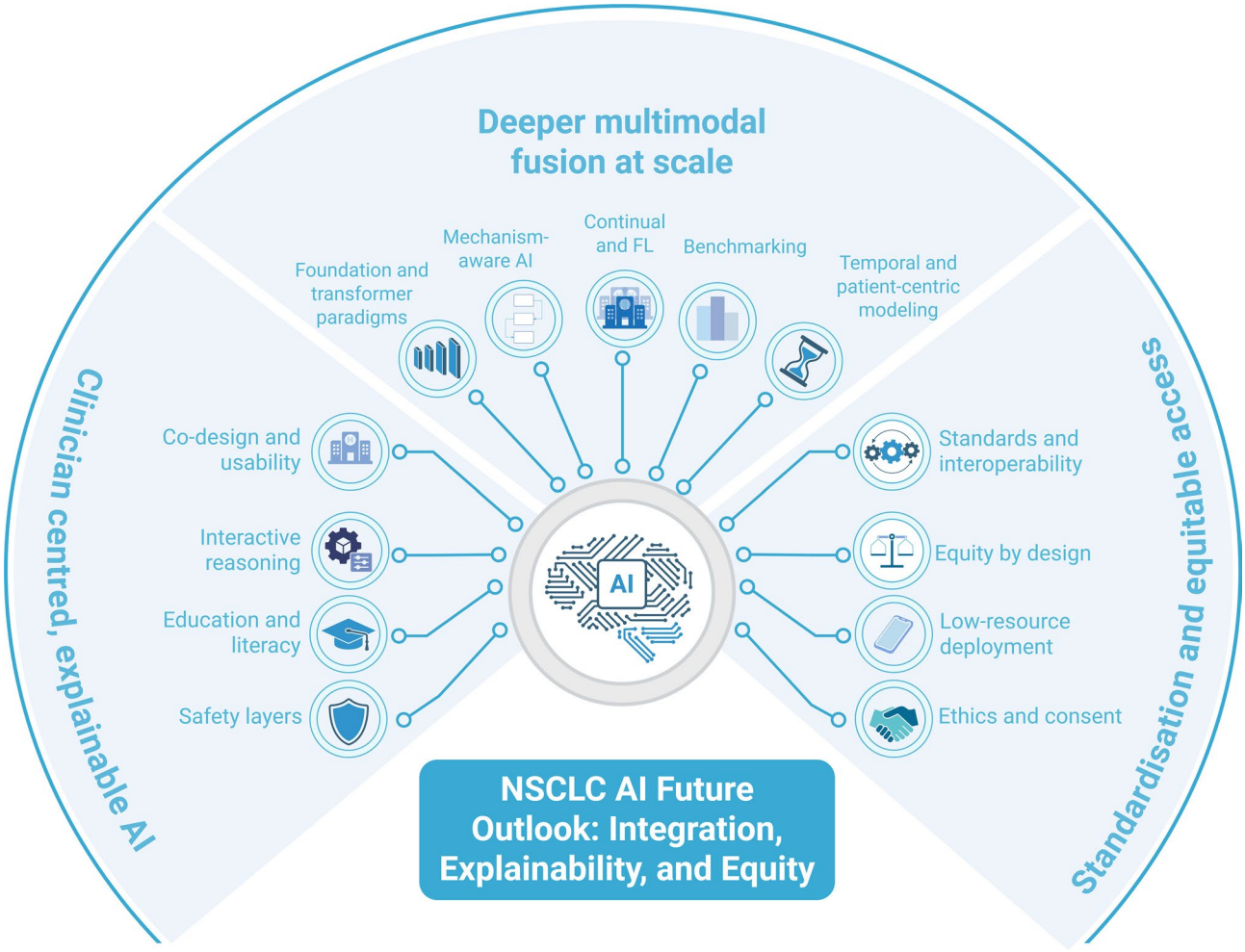

Fig. 5 NSCLC AI future outlook: integration, explainability, and equity

[30–32]. We recommend that future studies report the number of layers, the number of attention heads, the hidden size, the token size and stride, the attention variant, the parameter count, the pretraining corpus and modalities, and the adapter rank.

Temporal and patient-centric modeling

Move from snapshots to trajectories. Use sequence models to represent longitudinal imaging, ctDNA kinetics, laboratory trends, and therapy timelines, and update risk in real time [103]. Combine these models with digital twin constructs that simulate counterfactual responses under alternative regimens and schedules, enabling "what if" exploration during tumor boards [207].

Mechanism-aware AI

Integrate pathway knowledge and causal constraints to reduce spurious associations by using pathway regularized networks, graph causal models that link radiomic heterogeneity to hypoxia and immune evasion programs, and joint models of tumor and host interactions [208]. For drug response modeling, estimate individualized treatment benefit with treatment effect methods such as uplift modeling, causal forests, and targeted maximum likelihood estimation, and explicitly account for confounding through appropriate adjustment or identification strategies [209]. This section's discussion of causal forests, uplift modeling, and mechanism-aware models is reiterated in the Conclusion as a key set of emerging directions for individualized therapy response prediction.

Continual and FL

Anticipate continual updates as scanners, protocols, and therapies evolve [210]. Combine federated training with privacy preserving analytics to keep models current across networks while respecting local data governance requirements [112]. Distribute personalization layers using lightweight adapters that align models to site specific distributions without modifying the shared backbone [206].

Benchmarking

Establish open, multicenter NSCLC benchmarks that span multiple modalities and endpoints, including pathologic complete response, major pathologic response, event free survival, progression free survival, overall survival, and toxicity. Use standardized preprocessing and transparent protocol documentation, and maintain clearly separated training, validation, and test cohorts with temporally separated external tests to enable fair head to head comparisons and reproducibility [210].

Cliniciancentred, explainable AI

Co-design and usability

Co-develop solutions with radiologists, pathologists, oncologists, and nurses so that outputs align with clinical mental models, including risk categories with associated confidence, flags tied to clinically relevant thresholds such as PD-L1 near one percent or fifty percent, and concise rationales linked to recognizable clinical features [65]. Evaluate cognitive load and time to decision in simulated tumor boards and iteratively refine the user experience [34].

Interactive reasoning

Provide what if and counterfactual tools, link to exemplar cases, and allow threshold adjustment based on patient preferences. Enable selective automation by fully automating low risk, high confidence tasks while requiring human approval for high risk or low confidence outputs [211].

Education and literacy

Incorporate AI literacy into oncology training, covering uncertainty interpretation, calibration, fairness trade offs, and the limits of generalization. Define competencies and credentialing, including safe overrides and adverse event reporting [212].

Safety layers

Embed multilayer safeguards, including abstention based on OOD detection and UQ, fail safe defaults, automatic escalation for biomarker predictions near decision thresholds, and drift alarms that notify stakeholders before clinically meaningful degradation [213].

Standardisation and equitable access

Standards and interoperability

Consolidate around IBSI for radiomics, adopt common staining and scanning protocols for digital pathology, and ensure interoperability through DICOM SR and DICOM SEG, FHIR, and CDS Hooks. Publish reference implementations and open test suites to promote reproducibility [15, 50].

Equity by design

Proactively include underrepresented populations and sites in training and in external validation. Report calibration within groups and equalized odds by sex, ancestry, socioeconomic status, and geography [49]; When disparities are detected, apply remediation through reweighting, domain specific adapters, or targeted data acquisition.

Low-resource deployment

Optimize for edge inference, minimize dependencies, and support offline operation. Provide tiered models matched to local infrastructure, and consider pooled procurement and public private partnerships to reduce costs [214].

Ethics and consent

Clarify consents for secondary use, empower governance bodies to evaluate cross-border data flows and legal alignment, and reinvest benefits from AI deployments to improve access and outcomes in the communities that contributed data [215].

Verified explainability should be treated as a first-class requirement. Foundation and cross-modal models for NSCLC should be deployed with verified explanations as first-class outputs, audited according to the minimum evidence package, and accompanied by site-level adapters that maintain both prediction and explanation stability. Prospective evaluations should include human-factors endpoints and automation-bias monitoring. Regulatory dossiers should treat explanation verification with the same importance as discrimination and calibration.

Conclusions

AI is moving from proof of concept tools to clinically consequential systems across the NSCLC continuum [15]. Beyond matching specialist level accuracy on narrow tasks, the distinctive value of AI is to integrate high dimensional, multimodal, longitudinal information from radiology, pathology, multi omics, and electronic health records into calibrated, individualized inferences that humans cannot reliably integrate at scale [206]. When properly validated and deployed, AI can improve timing, including earlier detection and risk stratification; targeting, through more precise therapy selection with less futile toxicity; and throughput, by enabling standardized and efficient workflows that return clinician time to complex decisions [210]. Current evidence supports AI as an adjunct that enhances screening fidelity, histologic and molecular subtyping, prognostication, and prediction of treatment response and toxicity. However, heterogeneity in external performance, domain shift across scanners and laboratories, and variability in preanalytics mean that models should not be judged by AUROC alone;

calibration, transportability, uncertainty awareness, and clinical net benefit are also required [15]. In practice, the most useful systems generalize across settings with bounded performance loss, disclose uncertainty through selective prediction and abstention, and explain recommendations in clinically meaningful terms such as spiculation, necrosis, tumor infiltrating lymphocytes, and dose and volume drivers to support expert oversight [211, 216].

Causal and mechanism-aware modeling is a key emerging area central to individualized therapy-response prediction [217]. Causal forests and uplift modeling estimate individual treatment effects under confounding and heterogeneity [218]. Mechanism-aware representation learning encodes biological and physical priors, improving transportability, interpretability, and safety under distribution shift [219]. Clinical credibility requires multicenter evaluation with preregistered analysis plans, geographically and temporally external test cohorts, decision-relevant metrics such as net benefit and reclassification, and calibration reported within subgroups and by site [28]. Prospective pilots and pragmatic trials that embed these methods in tumor-board workflows can translate predictions into actionable strategies that balance benefit and risk for each patient.

Looking forward, three inflection points stand out: first, multimodal foundation models tuned for NSCLC should replace fragmented single task pipelines and produce unified outputs from a shared backbone with parameter efficient site personalization; second, causal and longitudinal modeling should become standard by combining treatment effect estimators with dynamic risk models that update after each imaging assessment, ctDNA measurement, or laboratory test; third, AI derived computable biomarkers should advance along formal qualification pathways, progressing from analytical validity and multicenter clinical validity to prospective demonstrations of clinical utility that inform trials, guidelines, and reimbursement [103, 118, 206, 209, 220]. Clinical translation requires an explicit pathway. We recommend a validation ladder that begins with internal and external splits, advances to preregistered multicenter retrospective studies, and culminates in DECIDE AI prospective evaluations and pragmatic trials where feasible [28, 214, 221]. At every stage, report calibration and DCA alongside discrimination, define operating points tied to actions, and prespecify recalibration plans [60, 222]. Deployment should follow Software as a Medical Device principles with postmarket surveillance, drift monitoring, rollback triggers, and change control plans for learning systems [223]. In summary, beyond accuracy and calibration, explanations verified for faithfulness, stability, and clinical utility, integrated with uncertainty-aware selective workflows, are essential for adoption in high-stakes NSCLC decisions. Our proposed minimum evidence package facilitates the implementation of this requirement.

Abbreviations

NSCLC	Non-small cell lung cancer
LDCT	Low-dose computed tomography
EGFR	Epidermal growth factor receptor
AI	Artificial intelligence
ML	Machine learning
CT	Computed tomography
MIL	Multiple instance learning
PD-L1	Programmed Death Ligand 1
H&E	Hematoxylin and Eosin
FA	Fourier attention
WA	Wavelet Attention
CtDNA	circulating tumor DNA
GNNs	Graph Neural Networks
AUROC	rea Under The Receiver Operating Characteristic Curve
PET-CT	Positron Emission Tomography-Computed Tomography
MRI	Magnetic Resonance Imaging
IBSI	Image Biomarker Standardisation Initiative
LoG	Laplacian of Gaussian
DCA	Decision Curve Analysis
IHC	Immunohistochemistry
WSI	Whole Slide Images
TPS	Tumor Proportion Score
ALK	Anaplastic Lymphoma kinase
ROS1	ROS proto oncogene 1
FL	Federated Learning
UQ	Uncertainty Quantification
OOD	Out-Of-Distribution
ICC	International Color Consortium
DICOM	Digital Imaging and Communications in Medicine
FHIR	Fast Healthcare Interoperability Resources

Acknowledgements
Figures 1, 2, 3, 4 and 5 were created with BioRender.com.

Author contributions
Conceptualization, Shilong Song; Study design, Luyuan Chang; Manuscript draft, Luyuan Chang; Draft review and editing, Haipeng Li.; Acquisition work, Wenzong Wu; Literature review and analysis Xinyu Liu; Literature review and analysis, Jiaqi Yan; Icon creation, Zuo Chen; Literature review and analysis, Huan Wu.

Funding
This work was supported by National Natural Science Foundation of China (82403993), Scientific Research Project of Anhui Province of Health Commission (AHWJ2023A30167), and Longhu Talents Program (LH250104010).

Data availability
All data for this study have been provided.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors have seen and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]The First Department of Clinical Medicine (First Affiliated Hospital), Bengbu Medical University, Bengbu, Anhui, China

[2]Department Mental Health, Bengbu Medical University, Bengbu, Anhui, China
[3]The Department of Radiotherapy of the First Affiliated Hospital of Bengbu Medical University, Bengbu, Anhui, China

Received: 28 August 2025 / Accepted: 12 December 2025
Published online: 23 December 2025

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
2. Hendriks LEL, Remon J, Faivre-Finn C, Garassino MC, Heymach JV, Kerr KM, et al. Non-small-cell lung cancer. Nat Rev Dis Primers. 2024;10(1):71.
3. Wakelee H, Liberman M, Kato T, Tsuboi M, Lee SH, Gao S, et al. Perioperative pembrolizumab for early-stage non-small-cell lung cancer. N Engl J Med. 2023;389(6):491–503.
4. Heymach JV, Harpole D, Mitsudomi T, Taube JM, Galffy G, Hochmair M, et al. Perioperative durvalumab for resectable non-small-cell lung cancer. N Engl J Med. 2023;389(18):1672–84.
5. Kratzer TB, Bandi P, Freedman ND, Smith RA, Travis WD, Jemal A, et al. Lung cancer statistics, 2023. Cancer. 2024;130(8):1330–48.
6. Frankell AM, Dietzen M, Al Bakir M, Lim EL, Karasaki T, Ward S, et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. Nature. 2023;616(7957):525–33.
7. Al Bakir M, Huebner A, Martínez-Ruiz, Grigoriadis K, Watkins TBK, Pich O, et al. The evolution of non-small cell lung cancer metastases in TRACERx. Nature. 2023;616(7957):534–42.
8. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773–84.
9. Waqas A, Tripathi A, Ramachandran RP, Stewart PA, Rasool G. Multimodal data integration for oncology in the era of deep neural networks: a review. Front Artif Intell. 2024;7:1408843.
10. Mo C-K, Liu J, Chen S, Storrs E, Targino da Costa ALN, Houston A, et al. Tumour evolution and microenvironment interactions in 2D and 3D space. Nature. 2024;634(8036):1178–86.
11. Sun Q, Hong Z, Zhang C, Wang L, Han Z, Ma D. Immune checkpoint therapy for solid tumours: clinical dilemmas and future trends. Signal Transduct Target Ther. 2023;8(1):320.
12. Wu B, Zhang B, Li B, Wu H, Jiang M. Cold and hot tumors: from molecular mechanisms to targeted therapy. Signal Transduct Targeted Ther. 2024;9(1):274.
13. Shafighi S, Geras A, Jurzysta B, Sahaf Naeini A, Filipiuk I, Rączkowska A, et al. Integrative spatial and genomic analysis of tumor heterogeneity with Tumoroscope. Nat Commun. 2024;15(1):9343.
14. Hill W, Lim EL, Weeden CE, Lee C, Augustine M, Chen K, et al. Lung adenocarcinoma promotion by air pollutants. Nature. 2023;616(7955):159–67.
15. Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Severson K, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med. 2024;30(10):2924–35.
16. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023;24(8):936–44.
17. Eisemann N, Bunk S, Mukama T, Baltus H, Elsner SA, Gomille T, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. Nat Med. 2025;31(3):917–24.
18. Mikhael PG, Wohlwend J, Yala A, Karstens L, Xiang J, Takigami AK, et al. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. J Clin Oncol. 2023;41(12):2191–200.
19. Venkadesh KV, Aleef TA, Scholten ET, Saghir Z, Silva M, Sverzellati N, et al. Prior ct improves deep learning for malignancy risk estimation of screening-detected pulmonary nodules. Radiology. 2023;308(2):e223308.
20. Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30(3):850–62.
21. Ito H, Yoshizawa A, Terada K, Nakakura A, Rokutan-Kurata M, Sugimoto T, et al. A deep learning-based assay for programmed death Ligand 1 Immunohistochemistry scoring in non-small cell lung carcinoma: does it help pathologists score? Mod Pathol. 2024;37(6):100485.
22. Campanella G, Kumar N, Nanda S, Singi S, Fluder E, Kwan R, et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. Nat Med. 2025.
23. Paverd H, Zormpas-Petridis K, Clayton H, Burge S, Crispin-Ortuzar M. Radiology and multi-scale data integration for precision oncology. NPJ Precis Oncol. 2024;8(1):158.
24. Truhn D, Eckardt JN, Ferber D, Kather JN. Large language models and multimodal foundation models for precision oncology. NPJ Precis Oncol. 2024;8(1):72.
25. Yang H, Yang M, Chen J, Yao G, Zou Q, Jia L. Multimodal deep learning approaches for precision oncology: a comprehensive review. Briefings Bioinf. 2025;26(1).
26. Captier N, Lerousseau M, Orlhac F, Hovhannisyan-Baghdasarian N, Luporsi M, Woff E, et al. Integration of clinical, pathological, radiological, and transcriptomic data improves prediction for first-line immunotherapy outcome in metastatic non-small cell lung cancer. Nat Commun. 2025;16(1):614.
27. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, et al. Pathomic fusion: an integrated framework for fusing Histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans Med Imag. 2022;41(4):757–70.
28. TRIPOD+AI statement. Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:q902.
29. Wang F, Ke H, Tang Y. Fusion of generative adversarial networks and non-negative tensor decomposition for depression fMRI data analysis. Inf Process Manag. 2025;62(2):103961.
30. Ke H, Wang F, Bi H, Ma H, Wang G, Yin B. Unsupervised deep frequency-channel attention factorization to non-linear feature extraction: a case study of identification and functional connectivity interpretation of Parkinson's disease. Expert Syst Appl. 2024;243:122853.
31. Wang F, Ke H, Cai C. Deep wavelet self-attention non-negative tensor factorization for non-linear analysis and classification of fMRI data. Appl Soft Comput. 2025;182:113522.
32. Wang F, Ke H, Ma H, Tang Y. Deep wavelet temporal-frequency attention for nonlinear fMRI factorization in ASD. Pattern Recognit. 2025;165:111543.
33. Wang C, Shao J, He Y, Wu J, Liu X, Yang L, et al. Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. Nat Med. 2024;30(11):3184–95.
34. Wenderott K, Krups J, Zaruchas F, Weigl M. Effects of artificial intelligence implementation on efficiency in medical imaging-a systematic literature review and meta-analysis. NPJ Digit Med. 2024;7(1):265.
35. De Luca GR, Diciotti S, Mascalchi M. The pivotal role of baseline LDCT for lung cancer screening in the era of artificial intelligence. Arch Bronconeumol. 2025;61(6):359–67.
36. Hendrix W, Hendrix N, Scholten ET, Mourits M, Trap-de Jong J, Schalekamp S, et al. Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans. Commun Med (Lond). 2023;3(1):156.
37. Sarkar S, Teo PT, Abazeed ME. Deep learning for automated, motion-resolved tumor segmentation in radiotherapy. NPJ Precis Oncol. 2025;9(1):173.
38. Zhu E, Muneer A, Zhang J, Xia Y, Li X, Zhou C, et al. Progress and challenges of artificial intelligence in lung cancer clinical translation. NPJ Precis Oncol. 2025;9(1):210.
39. Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol Artif Intell. 2023;5(5):e230024.
40. Link KE, Schnurman Z, Liu C, Kwon YJ, Jiang LY, Nasir-Moin M, et al. Longitudinal deep neural networks for assessing metastatic brain cancer on a large open benchmark. Nat Commun. 2024;15(1):8170.
41. Rogasch JMM, Michaels L, Baumgärtner GL, Frost N, Rückert JC, Neudecker J, et al. A machine learning tool to improve prediction of mediastinal lymph node metastases in non-small cell lung cancer using routinely obtainable [(18)F]FDG-PET/CT parameters. Eur J Nucl Med Mol Imag. 2023;50(7):2140–51.
42. Salehjahromi M, Karpinets TV, Sujit SJ, Qayati M, Chen P, Aminu M, et al. Synthetic pet from ct improves diagnosis and prognosis for lung cancer: proof of concept. Cell Rep Med. 2024;5(3):101463.

43. Zhong Y, Cai C, Chen T, Gui H, Deng J, Yang M, et al. PET/CT based cross-modal deep learning signature to predict occult nodal metastasis in lung cancer. Nat Commun. 2023;14(1):7513.
44. Koetzier LR, Mastrodicasa D, Szczykutowicz TP, van der Werf NR, Wang AS, Sandfort V, et al. Deep learning image reconstruction for ct: technical principles and clinical prospects. Radiology. 2023;306(3):e221257.
45. Yu P, Zhang H, Wang D, Zhang R, Deng M, Yang H, et al. Spatial resolution enhancement using deep learning improves chest disease diagnosis based on thick slice CT. NPJ Digit Med. 2024;7(1):335.
46. Horvat N, Papanikolaou N, Koh DM. Radiomics beyond the hype: a critical evaluation toward oncologic clinical use. Radiol Artif Intell. 2024;6(4):e230437.
47. Cobo M, Menéndez Fernández-Miranda P, Bastarrika G, Lloret Iglesias L. Enhancing radiomics and deep learning systems through the standardization of medical imaging workflows. Sci Data. 2023;10(1):732.
48. Pai S, Bontempi D, Hadzic I, Prudente V, Sokač M, Chaunzwa TL, et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell. 2024;6(3):354–67.
49. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. Nat Med. 2024;30(10):2838–48.
50. Whybra P, Zwanenburg A, Andrearczyk V, Schaer R, Apte AP, Ayotte A, et al. The image biomarker Standardization Initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. Radiology. 2024;310(2):e231319.
51. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819.
52. Vasey B, Collins GS. Invited commentary: transparent reporting of artificial intelligence models development and evaluation in surgery: the TRIPOD and DECIDE-AI checklists. Surgery. 2023;174(3):727–29.
53. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Pac Symp Biocomput. 2021;26:232–43.
54. Christiansen F, Konuk E, Ganeshan AR, Welch R, Palés Huix J, Czekierdowski A, et al. International multicenter validation of AI-driven ultrasound detection of ovarian cancer. Nat Med. 2025;31(1):189–96.
55. Xiang H, Xiao Y, Li F, Li C, Liu L, Deng T, et al. Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. Nat Commun. 2024;15(1):2681.
56. Ying H, Liu X, Zhang M, Ren Y, Zhen S, Wang X, et al. A multicenter clinical AI system study for detection and diagnosis of focal liver lesions. Nat Commun. 2024;15(1):1131.
57. Ktena I, Wiles O, Albuquerque I, Rebuffi S-A, Tanno R, Roy AG, et al. Generative models improve fairness of medical classifiers under distribution shifts. Nat Med. 2024;30(4):1166–73.
58. Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. NPJ Digit Med. 2024;7(1):120.
59. Qian X, Lu W, Zhang Y. Adaptive wavelet-VNet for single-sample test time adaptation in medical image segmentation. Med Phys. 2024;51(12):8865–81.
60. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
61. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–69.
62. Vickers AJ, Holland F. Decision curve analysis to evaluate the clinical benefit of prediction models. Spine J. 2021;21(10):1643–48.
63. Antonissen N, Tryfonos O, Houben IB, Jacobs C, de Rooij M, van Leeuwen KG. Artificial intelligence in radiology: 173 commercially available products and their scientific evidence. Eur Radiol. 2025.
64. Singh R, Bapna M, Diab AR, Ruiz ES, Lotter W. How AI is used in FDA-authorized medical devices: a taxonomy across 1, 016 authorizations. NPJ Digit Med. 2025;8(1):388.
65. Chae A, Yao MS, Sagreiya H, Goldberg AD, Chatterjee N, MacLean MT, et al. Strategies for implementing machine learning algorithms in the clinical practice of radiology. Radiology. 2024;310(1):e223170.
66. Farič N, Hinder S, Williams R, Ramaesh R, Bernabeu MO, van Beek E, et al. Early experiences of integrating an artificial intelligence-based diagnostic decision support system into radiology settings: a qualitative study. J Am Med Inf Assoc. 2023;31(1):24–34.
67. Geppert J, Asgharzadeh A, Brown A, Stinton C, Helm EJ, Jayakody S, et al. Software using artificial intelligence for nodule and cancer detection in CT lung cancer screening: systematic review of test accuracy studies. Thorax. 2024;79(11):1040–49.
68. Jorg T, Halfmann MC, Stoehr F, Arnhold G, Theobald A, Mildenberger P, et al. A novel reporting workflow for automated integration of artificial intelligence results into structured radiology reports. Insights Imag. 2024;15(1):80.
69. Tanno R, Barrett DGT, Sellergren A, Ghaisas S, Dathathri S, See A, et al. Collaboration between clinicians and vision-language models in radiology report generation. Nat Med. 2025;31(2):599–608.
70. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ. 2024;386:e078276.
71. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, et al. Assessing the Trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol Artif Intell. 2021;3(6):e200267.
72. Huber T, Limmer B, André E. Benchmarking perturbation-based saliency maps for explaining atari agents. Front Artif Intell. 2022;5:903875.
73. Alangari N, El Bachir Menai M, Mathkour H, Almosallam I. Exploring evaluation methods for interpretable machine learning: a survey. Inf [Internet]. 2023;14(8).
74. Gandomkar Z, Khong PL, Punch A, Lewis S. Using occlusion-based saliency maps to explain an artificial intelligence tool in lung cancer screening: agreement between radiologists, labels, and visual prompts. J Digit Imag. 2022;35(5):1164–75.
75. Passaro A, Al Bakir M, Hamilton EG, Diehn M, André F, Roy-Chowdhuri S, et al. Cancer biomarkers: emerging trends and clinical implications for personalized treatment. Cell. 2024;187(7):1617–35.
76. Kludt C, Wang Y, Ahmad W, Bychkov A, Fukuoka J, Gaisa N, et al. Next-generation lung cancer pathology: development and validation of diagnostic and prognostic algorithms. Cell Rep Med. 2024;5(9):101697.
77. El Nahhas OSM, Loeffler CML, Carrero ZI, van Treeck M, Kolbinger FR, Hewitt KJ, et al. Regression-based deep-learning predicts molecular biomarkers from pathology slides. Nat Commun. 2024;15(1):1253.
78. Amidi E, Ramzanpour M, Chen M, Boucher T, Varma M, Samec T, et al. Predicting ROS1 and alk fusions in NSCLC from H&E slides with a two-step vision transformer approach. NPJ Precis Oncol. 2025;9(1):266.
79. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024;630(8015):181–88.
80. van Kolfschooten H, van Oirschot J. The eu artificial intelligence Act (2024). Implications for healthcare. Health Policy. 2024;149:105152.
81. Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, Chow AK, et al. A multimodal generative AI copilot for human pathology. Nature. 2024;634(8033):466–73.
82. McGenity C, Clarke EL, Jennings C, Matthews G, Cartlidge C, Freduah-Agyemang H, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. NPJ Digit Med. 2024;7(1):114.
83. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun. 2021;12(1):4423.
84. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. medRxiv. 2020.
85. Hägele M, Seegerer P, Lapuschkin S, Bockmayr M, Samek W, Klauschen F, et al. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Sci Rep. 2020;10(1):6423.
86. Kothari S, Phan JH, Stokes TH, Osunkoya AO, Young AN, Wang MD. Removing batch effects from histopathological images for enhanced cancer diagnosis. IEEE J Biomed Health Inf. 2014;18(3):765–72.
87. Nakamura Y, Watanabe J, Akazawa N, Hirata K, Kataoka K, Yokota M, et al. ctDNA-based molecular residual disease and survival in resectable colorectal cancer. Nat Med. 2024;30(11):3272–83.
88. Song KJ, Choi S, Kim K, Hwang HS, Chang E, Park JS, et al. Proteogenomic analysis reveals non-small cell lung cancer subtypes predicting chromosome instability, and tumor microenvironment. Nat Commun. 2024;15(1):10164.
89. Wang C, Li J, Chen J, Wang Z, Zhu G, Song L, et al. Multi-omics analyses reveal biological and clinical insights in recurrent stage I non-small cell lung cancer. Nat Commun. 2025;16(1):1477.
90. Jiang L, Xu C, Bai Y, Liu A, Gong Y, Wang Y-P, et al. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. NPJ Precis Oncol. 2024;8(1):4.
91. Ing A, Andrades A, Cosenza MR, Korbel JO. Integrating multimodal cancer data using deep latent variable path modelling. Nat Mach Intel. 2025;7(7):1053–75.

92. Hartman E, Scott AM, Karlsson C, Mohanty T, Vaara ST, Linder A, et al. Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. Nat Commun. 2023;14(1):5359.
93. Hands I, Kavuluru R. A survey of NLP methods for oncology in the past decade with a focus on cancer registry applications. Artif Intell Rev. 2025;58(10):314.
94. Hong Z, Yue Y, Chen Y, Lin H, Luo Y, Wang MH, et al. Out-of-distribution detection in medical image analysis: a survey. ArXiv. 2024;abs/2404.18279.
95. Jiang S, Hondelink L, Suriawinata AA, Hassanpour S. Masked pre-training of transformers for histology image analysis. J Pathol Inf. 2024;15:100386.
96. Anghel A, Stanisavljevic M, Andani S, Papandreou N, Rüschoff JH, Wild P, et al. A high-performance System for Robust stain normalization of whole-slide images in Histopathology. Front Med (Lausanne). 2019;6:193.
97. Vickers AJ, Van Claster B, Wynants L, Steyerberg EW. Decision curve analysis: confidence intervals and hypothesis testing for net benefit. Diagn Progn Res. 2023;7(1):11.
98. Steyaert S, Qiu YL, Zheng Y, Mukherjee P, Vogel H, Gevaert O. Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. Commun Med. 2023;3(1):44.
99. Tian R, Hou F, Zhang H, Yu G, Yang P, Li J, et al. Multimodal fusion model for prognostic prediction and radiotherapy response assessment in head and neck squamous cell carcinoma. NPJ Digit Med. 2025;8(1):302.
100. Sujit SJ, Aminu M, Karpinets TV, Chen P, Saad MB, Salehjahromi M, et al. Enhancing NSCLC recurrence prediction with PET/CT habitat imaging, ctDNA, and integrative radiogenomics-blood insights. Nat Commun. 2024;15(1):3152.
101. Tsuboi M, Herbst RS, John T, Kato T, Majem M, Grohé C, et al. Overall survival with osimertinib in resected EGFR-Mutated NSCLC. N Engl J Med. 2023;389(2):137–47.
102. Anagnostou V, Ho C, Nicholas G, Juergens RA, Sacher A, Fung AS, et al. ctDNA response after pembrolizumab in non-small cell lung cancer: phase 2 adaptive trial results. Nat Med. 2023;29(10):2559–69.
103. Assaf ZJF, Zou W, Fine AD, Socinski MA, Young A, Lipson D, et al. A longitudinal circulating tumor DNA-based model associated with survival in metastatic non-small-cell lung cancer. Nat Med. 2023;29(4):859–68.
104. Ding H, Yuan M, Yang Y, Xu XS. Identifying key circulating tumor DNA parameters for predicting clinical outcomes in metastatic non-squamous non-small cell lung cancer after first-line chemoimmunotherapy. Nat Commun. 2024;15(1):6862.
105. Jee J, Fong C, Pichotta K, Tran TN, Luthra A, Waters M, et al. Automated real-world data integration improves cancer outcome prediction. Nature. 2024;636(8043):728–36.
106. Orcutt X, Chen K, Mamtani R, Long Q, Parikh RB. Evaluating generalizability of oncology trial results to real-world patients using machine learning-based trial emulations. Nat Med. 2025;31(2):457–65.
107. Derraz B, Breda G, Kaempf C, Baenke F, Cotte F, Reiche K, et al. New regulatory thinking is needed for AI-based personalised drug and cell therapies in precision oncology. NPJ Precis Oncol. 2024;8(1):23.
108. Jayaraman P, Desman J, Sabounchi M, Nadkarni GN, Sakhuja A. A Primer on Reinforcement learning in Medicine for clinicians. NPJ Digit Med. 2024;7(1):337.
109. Tosca EM, De Carlo A, Ronchi D, Magni P. Model-informed Reinforcement learning for enabling precision Dosing via adaptive Dosing. Clin Pharmacol Ther. 2024;116(3):619–36.
110. Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med. 2023;29(11):2929–38.
111. Aboy M, Minssen T, Vayena E. Navigating the eu AI Act: implications for regulated digital medical products. NPJ Digit Med. 2024;7(1):237.
112. Feng B, Shi J, Huang L, Yang Z, Feng S-T, Li J, et al. Robustly federated learning model for identifying high-risk patients with postoperative gastric cancer recurrence. Nat Commun. 2024;15(1):742.
113. Pati S, Kumar S, Varma A, Edwards B, Lu C, Qu L, et al. Privacy preservation for federated learning in health care. Patterns (N Y). 2024;5(7):100974.
114. Yan F, Da Q, Yi H, Deng S, Zhu L, Zhou M, et al. Artificial intelligence-based assessment of PD-L1 expression in diffuse large B cell lymphoma. NPJ Precis Oncol. 2024;8(1):76.
115. Guo J, Miao J, Sun W, Li Y, Nie P, Xu W. Predicting bone metastasis-free survival in non-small cell lung cancer from preoperative ct via deep learning. NPJ Precis Oncol. 2024;8(1):161.
116. Li Y, Chai X, Yang M, Xiong J, Zeng J, Chen Y, et al. Accurate prediction of disease-free and overall survival in non-small cell lung cancer using patient-level multimodal weakly supervised learning. NPJ Precis Oncol. 2025;9(1):197.
117. Zheng S, Guo J, Langendijk JA, Both S, Veldhuis RNJ, Oudkerk M, et al. Survival prediction for stage I-IIIA non-small cell lung cancer using deep learning. Radiother Oncol. 2023;180:109483.
118. Ding H, Feng Y, Huang X, Xu J, Zhang T, Liang Y, et al. Deep learning-based classification and spatial prognosis risk score on whole-slide images of lung adenocarcinoma. Histopathology. 2023;83(2):211–28.
119. Lu CF, Liao CY, Chao HS, Chiu HY, Wang TW, Lee Y, et al. A radiomics-based deep learning approach to predict progression free-survival after tyrosine kinase inhibitor therapy in non-small cell lung cancer. Cancer Imag. 2023;23(1):9.
120. Lin X, Liu Z, Zhou K, Li Y, Huang G, Zhang H, et al. Intratumoral and peritumoral PET/CT-based radiomics for non-invasively and dynamically predicting immunotherapy response in NSCLC. Br J Cancer. 2025;132(6):558–68.
121. Spigel DR, Westeel V, Anderson IC, Greillier L, Guisier F, Bylicki O, et al. Adjuvant chemotherapy for stage IA-IIA non-squamous, non-small-cell lung cancer identified as molecular high-risk by a 14-gene expression profile (AIM-HIGH): an international, randomised, phase 3 trial. Lancet Respir Med. 2025.
122. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. Bmj. 2024;384:e074820.
123. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554.
124. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Publisher correction: reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(10):2218.
125. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng. 2023;7(6):719–42.
126. Hurkmans C, Bibault JE, Brock KK, van Elmpt W, Feng M, David Fuller C, et al. A joint ESTRO and AAPM guideline for development, clinical validation and reporting of artificial intelligence models in radiation therapy. Radiother Oncol. 2024;197:110345.
127. Pang EPP, Tan HQ, Wang F, Niemelä J, Bolard G, Ramadan S, et al. Multicentre evaluation of deep learning ct autosegmentation of the head and neck region for radiotherapy. NPJ Digit Med. 2025;8(1):312.
128. Muneer A, Waqas M, Saad MB, Showkatian E, Bandyopadhyay R, Xu H, et al. From classical machine learning to emerging foundation models: review on multimodal data integration for cancer Research. ArXiv. 2025, abs/2507.09028.
129. Al-Tashi Q, Saad MB, Muneer A, Qureshi R, Mirjalili S, Sheshadri A, et al. Machine learning models for the identification of prognostic and predictive cancer biomarkers: a systematic review. Int J Mol Sci. 2023;24(9).
130. Ogbonna CP, Breen WG, Le Noach P, Rajagopalan S, Hostetter LJ, Maldonado F, et al. Radiomics-based prediction of local recurrence after stereotactic body radiation therapy for early-stage non-small cell lung cancer. Ann Am Thorac Soc. 2025;22(8):1236–43.
131. Ouraou E, Tonneau M, Le WT, Filion E, Campeau MP, Vu T, et al. Predicting early stage lung cancer recurrence and survival from combined tumor motion amplitude and radiomics on free-breathing 4D-CT. Med Phys. 2025;52(3):1926–40.
132. Nie T, Chen Z, Cai J, Ai S, Xue X, Yuan M, et al. Integration of dosimetric parameters, clinical factors, and radiomics to predict symptomatic radiation pneumonitis in lung cancer patients undergoing combined immunotherapy and radiotherapy. Radiother Oncol. 2024;190:110047.
133. Ungvári T, Szabó D, Győrfi A, Dankovics Z, Kiss B, Olajos J, et al. Machine learning-driven imaging data for early prediction of lung toxicity in breast cancer radiotherapy. Sci Rep. 2025;15(1):18473.
134. Horne A, Harada K, Brown KD, Chua KLM, McDonald F, Price G, et al. Treatment response biomarkers: working toward personalized radiotherapy for lung cancer. J Thorac Oncol. 2024;19(8):1164–85.
135. Ligero M, El Nahhas OSM, Aldea M, Kather JN. Artificial intelligence-based biomarkers for treatment decisions in oncology. Trends Cancer. 2025;11(3):232–44.
136. Biswas D, Liu YH, Herrero J, Wu Y, Moore DA, Karasaki T, et al. Prospective validation of oracle, a clonal expression biomarker associated with survival of patients with lung adenocarcinoma. Nat Cancer. 2025;6(1):86–101.

137. She Y, He B, Wang F, Zhong Y, Wang T, Liu Z, et al. Deep learning for predicting major pathological response to neoadjuvant chemoimmunotherapy in non-small cell lung cancer: a multicentre study. EBioMedicine. 2022;86:104364.
138. He K, Baniasad M, Kwon H, Caval T, Xu G, Lebrilla C, et al. Decoding the glycoproteome: a new frontier for biomarker discovery in cancer. J Hematol Oncol. 2024;17(1):12.
139. Lotter W, Hassett MJ, Schultz N, Kehl KL, Van Allen EM, Cerami E. Artificial intelligence in oncology: current landscape, challenges, and future Directions. Cancer Discov. 2024;14(5):711–26.
140. Schneider JL, Lin JJ, Shaw AT. ALK-positive lung cancer: a moving target. Nat Cancer. 2023;4(3):330–43.
141. Lou N, Gao R, Shi Y, Han X. Plasma metabolomics profiling of EGFR-mutant NSCLC patients treated with third-generation EGFR-TKI. Sci Data. 2024;11(1):1369.
142. Huang D, Li Z, Jiang T, Yang C, Li N. Artificial intelligence in lung cancer: current applications, future perspectives, and challenges. Front Oncol. 2024;14:1486310.
143. Planchard D, Jänne PA, Cheng Y, Yang JC, Yanagitani N, Kim SW, et al. Osimertinib with or without chemotherapy in EGFR-Mutated advanced NSCLC. N Engl J Med. 2023;389(21):1935–48.
144. Berko ER, Witek GM, Matkar S, Petrova ZO, Wu MA, Smith CM, et al. Circulating tumor DNA reveals mechanisms of lorlatinib resistance in patients with relapsed/refractory ALK-driven neuroblastoma. Nat Commun. 2023;14(1):2601.
145. Cai Z, Apolinário S, Baião AR, Pacini C, Sousa MD, Vinga S, et al. Synthetic augmentation of cancer cell line multi-omic datasets using unsupervised deep learning. Nat Commun. 2024;15(1):10390.
146. Hong L, Aminu M, Li S, Lu X, Petranovic M, Saad MB, et al. Efficacy and clinicogenomic correlates of response to immune checkpoint inhibitors alone or with chemotherapy in non-small cell lung cancer. Nat Commun. 2023;14(1):695.
147. Parra ER, Zhang J, Jiang M, Tamegnon A, Pandurengan RK, Behrens C, et al. Immune cellular patterns of distribution affect outcomes of patients with non-small cell lung cancer. Nat Commun. 2023;14(1):2364.
148. Rakaee M, Tafavvoghi M, Ricciuti B, Alessi JV, Cortellini A, Citarella F, et al. Deep learning model for predicting immunotherapy response in advanced non-small cell lung cancer. JAMA Oncol. 2025;11(2):109–18.
149. Delasos L, Khorrami M, Viswanathan VS, Jazieh K, Ding Y, Mutha P, et al. Novel radiogenomics approach to predict and characterize pneumonitis in stage iii NSCLC. NPJ Precis Oncol. 2024;8(1):290.
150. Saad MB, Al-Tashi Q, Hong L, Verma V, Li W, Boiarsky D, et al. Machine-learning driven strategies for adapting immunotherapy in metastatic NSCLC. Nat Commun. 2025;16(1):6828.
151. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500.
152. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023;616(7958):673–85.
153. Chmielecki J, Mok T, Wu YL, Han JY, Ahn MJ, Ramalingam SS, et al. Analysis of acquired resistance mechanisms to osimertinib in patients with EGFR-mutated advanced non-small cell lung cancer from the AURA3 trial. Nat Commun. 2023;14(1):1071.
154. Zhou R, Liu Z, Wu T, Pan X, Li T, Miao K, et al. Machine learning-aided discovery of T790M-mutant EGFR inhibitor CDDO-Me effectively suppresses non-small cell lung cancer growth. Cell Commun Signal. 2024;22(1):585.
155. Xia Y, Sun M, Huang H, Jin WL. Drug repurposing for cancer therapy. Signal Transduct Target Ther. 2024;9(1):92.
156. Zhao Y, Xing Y, Zhang Y, Wang Y, Wan M, Yi D, et al. Evidential deep learning-based drug-target interaction prediction. Nat Commun. 2025;16(1):6915.
157. Sinha S, Vegesna R, Mukherjee S, Kammula AV, Dhruba SR, Wu W, et al. Perception predicts patient response and resistance to treatment using single-cell transcriptomics of their tumors. Nat Cancer. 2024;5(6):938–52.
158. Meller A, Ward M, Borowsky J, Kshirsagar M, Lotthammer JM, Oviedo F, et al. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. Nat Commun. 2023;14(1):1177.
159. Sobhani N, Tardiel-Cyril DR, Chai D, Generali D, Li JR, Vazquez-Perez J, et al. Artificial intelligence-powered discovery of small molecules inhibiting CTLA-4 in cancer. BJC Rep. 2024;2.
160. Lawrence PJ, Burns B, Ning X. Enhancing drug and cell line representations via contrastive learning for improved anti-cancer drug prioritization. NPJ Precis Oncol. 2024;8(1):106.
161. Shi Y, Li C, Zhang X, Peng C, Sun P, Zhang Q, et al. D3EGFR: a webserver for deep learning-guided drug sensitivity prediction and drug response information retrieval for EGFR mutation-driven lung cancer. Brief Bioinform. 2024;25(3).
162. Zhang K, Yang X, Wang Y, Yu Y, Huang N, Li G, et al. Artificial intelligence in drug development. Nat Med. 2025;31(1):45–59.
163. Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model. 2013;53(4):783–90.
164. Dahlin JL, Nissink JW, Strasser JM, Francis S, Higgins L, Zhou H, et al. Pains in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. J Med Chem. 2015;58(5):2091–113.
165. Handa K, Thomas MC, Kageyama M, Iijima T, Bender A. On the difficulty of validating molecular generative models realistically: a case study on public and proprietary data. J Cheminform. 2023;15(1):112.
166. Ash JR, Hughes-Oliver JM. Confidence bands and hypothesis tests for hit enrichment curves. J Cheminform. 2022;14(1):50.
167. Jiang X, Liu H, You Y, Zhong G, Ruan Z, Liao J, et al. Multi-omics reveals the protective effects of curcumin against AFB1-induced oxidative stress and inflammatory damage in duckling intestines. Comp Biochem Physiol C Toxicol Pharmacol. 2024;276:109815.
168. Sadiq NM, Abdulwahid RT, Aziz SB, Woo HJ, Kadir MFZ. Chitosan as a suitable host for sustainable plasticized nanocomposite sodium ion conducting polymer electrolyte in EDLC applications: structural, ion transport and electrochemical studies. Int J Biol Macromol. 2024;265(Pt 1):130751.
169. Goto H, Kitahara H, Matsumoto T, Tateishi K, Saito Y, Kato K, et al. Comparison of very early-phase vascular response to the CD34 antibody-covered sirolimus-eluting stent versus durable polymer-coated everolimus-eluting stent. Cardiovasc Interv Ther. 2025;40(3):527–35.
170. Dal Buono A, Faita F, Armuzzi A, Jairath V, Peyrin-Biroulet L, Danese S, et al. Assessment of activity and severity of inflammatory bowel disease in cross-sectional imaging techniques: a systematic review. J Crohns Colitis. 2025;19(2).
171. Priya S, Dhruba DD, Sorensen E, Aher PY, Narayanasamy S, Nagpal P, et al. ComBat harmonization of myocardial radiomic features sensitive to cardiac MRI acquisition parameters. Radiol Cardiothorac Imag. 2023;5(4):e220312.
172. Ling L, Khan H, Qianqian L, Qiumei L. Minimum wage standard adjustment and employment: heterogeneity effects on the human capital investment. Heliyon. 2024;10(3):e25097.
173. Tuminello S, Turner WM, Untalan M, Ivic-Pavlicic T, Flores R, Taioli E. Racial and socioeconomic disparities in non-small cell lung cancer molecular diagnostics uptake. J Natl Cancer Inst. 2025;117(1):112–19.
174. Li J, Zhang X, Liu X, Liao X, Huang J, Jiang Y. Co-upcycling of plastic waste and biowaste via tandem transesterification reactions. JACS Au. 2024;4(8):3135–45.
175. Machado R, Jacques PD, Nummer AR. Mesozoic/Cenozoic strike-slip tectonics in the catarinense shield and its correlation with structures associated with the continental rift in southeastern Brazil. An Acad Bras Cienc. 2022;94(suppl 4):e20211033.
176. Carstens DD, Maselli DJ, Cook EE, Mu F, Chen J, Yang D, et al. Real-world effectiveness of benralizumab among patients with asthma and concomitant chronic obstructive pulmonary disease. Int J Chron Obstruct Pulmon Dis. 2024;19:1813–18.
177. Whitaker L, Sherman N, Ahmed I, Etkin Y. A review of the current recommendations and practices for hemodialysis access monitoring and maintenance procedures. Semin Vasc Surg. 2024;37(2):133–49.
178. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. 2022;5(1):48.
179. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683.
180. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and ct scans. Nat Mach Intel. 2021;3(3):199–217.
181. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health. 2022;4(6):e406–14.
182. Zhang J, Chao H, Dasegowda G, Wang G, Kalra MK, Yan P. Revisiting the Trustworthiness of saliency methods in radiology AI. Radiol Artif Intell. 2024;6(1):e220221.

183. Abgrall G, Holder AL, Chelly Dagdia Z, Zeitouni K, Monnet X. Should AI models be explainable to clinicians? Crit Care. 2024;28(1):301.
184. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. NPJ Digit Med. 2023;6(1):206.
185. Liu Y, Liu C, Zheng J, Xu C, Wang D. Improving explainability and integrability of medical AI to promote Health care professional acceptance and use: mixed systematic review. J Med Internet Res. 2025;27:e73374.
186. Dvijotham KD, Winkens J, Barsbey M, Ghaisas S, Stanforth R, Pawlowski N, et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. Nat Med. 2023;29(7):1814–20.
187. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924–33.
188. Roth CJ, Petersilge C, Clunie D, Towbin AJ, Cram D, Primo R, et al. HIMSS-SIIM Enterprise imaging community white papers: reflections and future Directions. J Imag Inf Med. 2024;37(2):429–43.
189. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
190. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283–86.
191. Zhu F, Zhang XY, Cheng Z, Liu CL. Revisiting confidence estimation: towards reliable failure prediction. IEEE Trans Pattern Anal Mach Intell. 2024;46(5):3370–87.
192. Popat R, Ive J. Embracing the uncertainty in human-machine collaboration to support clinical decision-making for mental health conditions. Front Digit Health. 2023;5:1188338.
193. Tejani AS, Cook TS, Hussain M, Sippel Schmidt T, O'Donnell KP. Integrating and adopting AI in the radiology workflow: a Primer for standards and integrating the healthcare Enterprise (IHE) profiles. Radiology. 2024;311(3):e232653.
194. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904.
195. Pencina MJ, D'Agostino RB Sr, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Stat Med. 2012;31(2):101–13.
196. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J. 2023;65(8):e2200302.
197. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, et al. Consolidated Health economic evaluation reporting standards 2022 (cheers 2022) statement: updated reporting guidance for health economic evaluations. J Manag Care Spec Pharm. 2022;28(2):146–55.
198. Zhou K, Gattinger G. The evolving regulatory paradigm of AI in MedTech: a review of perspectives and where we are Today. Ther Innov Regul Sci. 2024;58(3):456–64.
199. Singh V, Cheng S, Kwan AC, Ebinger J. United States Food and drug Administration regulation of clinical Software in the era of artificial intelligence and machine learning. Mayo Clin Proc Digit Health. 2025;3(3):100231.
200. Wells BJ, Nguyen HM, McWilliams A, Pallini M, Bovi A, Kuzma A, et al. A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare. NPJ Digit Med. 2025;8(1):514.
201. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell. 2022;4(3):e210064.
202. Campanella G, Kumar N, Nanda S, Singi S, Fluder E, Kwan R, et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. Nat Med. 2025;31(9):3002–10.
203. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.
204. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics. 2014;15(3):526–39.
205. Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vera Garcia DV, Singh Y, et al. Mitigating bias in radiology machine learning: 1. Data handling. Radiol Artif Intell. 2022;4(5):e210290.
206. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–65.
207. Laubenbacher R, Mehrad B, Shmulevich I, Trayanova N. Digital twins in medicine. Nat Comput Sci. 2024;4(3):184–91.
208. Li H, Han Z, Sun Y, Wang F, Hu P, Gao Y, et al. Cgmega: explainable graph neural network framework with attention mechanisms for cancer gene module dissection. Nat Commun. 2024;15(1):5997.
209. Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. Nat Med. 2024;30(4):958–68.
210. Perez-Lopez R, Ghaffari Laleh N, Mahmood F, Kather JN. A guide to artificial intelligence for cancer researchers. Nat Rev Cancer. 2024;24(6):427–41.
211. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, et al. A visual-language foundation model for computational pathology. Nat Med. 2024;30(3):863–74.
212. Car J, Ong QC, Erlikh Fox T, Leightley D, Kemp SJ, Švab I, et al. The digital Health competencies in medical education framework: an International consensus statement based on a Delphi study. JAMA Netw Open. 2025;8(1):e2453131.
213. Sahlsten J, Jaskari J, Wahid KA, Ahmed S, Glerean E, He R, et al. Application of simultaneous uncertainty quantification and segmentation for oropharyngeal cancer use-case with Bayesian deep learning. Commun Med (Lond). 2024;4(1):110.
214. Han G-R, Goncharov A, Eryilmaz M, Ye S, Palanisamy B, Ghosh R, et al. Machine learning in point-of-care testing: innovations, challenges, and opportunities. Nat Commun. 2025;16(1):3165.
215. Gilbert S. The eu passes the AI Act and its implications for digital medicine are unclear. NPJ Digit Med. 2024;7(1):135.
216. Zhou Q, Z-H C, Y-H C, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit Med. 2021;4(1):154.
217. Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019;47:1179–203.
218. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol. 2022;22(1):316.
219. Vasey B, Nagendran M, Campbell B, Clifton D, Collins G, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904.
220. Campanella G, Kumar N, Nanda S, Singi S, Fluder E, Kwan R, et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. Nat Med. 2025.
221. Martindale APL, Llewellyn CD, de Visser RO, Ng B, Ngai V, Kale AU, et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. Nat Commun. 2024;15(1):1619.
222. Liou L, Scott E, Parchure P, Ouyang Y, Egorova N, Freeman R, et al. Assessing calibration and bias of a deployed machine learning malnutrition prediction model within a large healthcare system. NPJ Digit Med. 2024;7(1):149.
223. Hillis JM, Visser JJ, Cliff ERS, van der Geest-Aspers K, Bizzo BC, Dreyer KJ, et al. The lucent yet opaque challenge of regulating artificial intelligence in radiology. NPJ Digit Med. 2024;7(1):69.

Publisher's Note