

Imaging-Centered Artificial Intelligence for Precision Cancer Immunotherapy: Multimodal Biomarkers, Treatment-Effect Estimation, and Clinical Translation

Shengyang Zhong

Shaanxi Normal University <https://orcid.org/0009-0000-0902-8139>

Siyuan Li

University of Georgia

Wanrong Yang

University of Liverpool

Zheng Zheng

The First Hospital of Hebei Medical University

Zhengliang Liu

University of Georgia

Jin Xie

University of Georgia

Jin Li

Shaanxi Normal University

Bin Yi

University of Georgia

Lin Zhao

New Jersey Institute of Technology

Xiang Li

Massachusetts General Hospital and Harvard Medical School

Wei Liu

Mayo Clinic

Lian Zhang

The First Hospital of Hebei Medical University

Tianming Liu

University of Georgia

Ning Qiang

qn315@snnu.edu.cn

Systematic Review

Keywords: artificial intelligence, medical imaging, cancer immunotherapy, radiomics, multimodal biomarkers, radiotherapy

Posted Date: July 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10468208/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Imaging-Centered Artificial Intelligence for Precision Cancer Immunotherapy: Multimodal Biomarkers, Treatment-Effect Estimation, and Clinical Translation

Shengyang Zhong^{a,1}, Siyuan Li^{b,1}, Wanrong Yang^d, Zheng Zheng^e, Zhengliang Liu^b, Jin Xie^c, Jin Li^a, Bin Yi^f, Lin Zhao^g, Xiang “Shaun” Li^h, Wei Liuⁱ, Lian Zhang^e, Tianming Liu^b, Ning Qiang^{a,*}

^a*College of Physics and Information Technology, Shaanxi Normal University, 620 West Chang'an Avenue, Xi'an, 710119, Shaanxi, China*

^b*School of Computing, University of Georgia, 415 Boyd Research and Education Center, Athens, 30602-7404, GA, USA*

^c*Department of Chemistry, Franklin College of Arts and Sciences, University of Georgia, 312 East Campus Road, Suite 1299B, Athens, 30602, GA, USA*

^d*School of Computer Science and Informatics, University of Liverpool, Foundation Building 765 Brownlow Hill, Liverpool, L69 7ZX, UK*

^e*The First Hospital of Hebei Medical University, No. 89 Donggang Road, Yuhua District, Shijiazhuang, 050031, Hebei, China*

^f*Department of Pharmaceutical and Biomedical Sciences, College of Pharmacy, University of Georgia, 250 West Green Street, Athens, 30602, GA, USA*

^g*Department of Biomedical Engineering, New Jersey Institute of Technology, University Heights, Newark, 07102, NJ, USA*

^h*Center for Advanced Medical Computing and Analysis, Mass General Research Institute, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, 399 Revolution Drive, 11th Floor, Somerville, 02145, MA, USA*

ⁱ*Medical Physics, Department of Radiation Oncology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, 5777 E. Mayo Blvd., Phoenix, 85054, AZ, USA*

Abstract

Immune checkpoint inhibitors (ICIs) can produce durable tumor control, yet many patients experience primary or acquired resistance, mixed lesion-level responses, or immune-related adverse events. Tissue and blood biomarkers provide biologically informative but spatially incomplete views of a

*Corresponding author.

Email address: qn315@snnu.edu.cn (Ning Qiang)

¹These authors contributed equally to this work.

heterogeneous and time-varying tumor–host system. Radiologic imaging complements these assays by repeatedly measuring whole-body tumor burden, organ-specific disease patterns, lesion-level heterogeneity, and treatment-related change. This critical narrative review evaluates imaging-centered artificial intelligence (AI) for precision cancer immunotherapy, including baseline and serial CT, MRI, and PET; radiomics and learned image representations; lesion-level and whole-body heterogeneity; toxicity assessment; and integration with pathology, genomics, circulating tumor DNA, laboratory measurements, and electronic health records. Representative cohorts range from early radiogenomic studies of approximately 100–200 patients to externally validated and longitudinal models involving hundreds or thousands of patients. Rather than treating all reported performance estimates as equally actionable, we appraise the clinical question, cohort design, endpoint, validation strategy, calibration, and intended clinical action. Current evidence supports noninvasive immune phenotyping, risk stratification, longitudinal monitoring, and multimodal biomarker integration. However, most models predict outcomes among patients already receiving ICIs and therefore remain prognostic under treatment rather than predictive of differential treatment benefit. Radiotherapy (RT) combined with ICI serves as a focused imaging use case because lesion identity, dose distribution, fractionation, sequence, and thoracic and normal-tissue exposure must be linked to defined actions. Existing patient-level AI studies estimate efficacy or toxicity after treatment selection; none has prospectively shown that an AI-selected RT strategy improves patient outcomes. Translation will require patient-level leakage control, calibration, geographic and temporal validation, robustness to missing modalities, and prospective evaluation of prespecified model-guided decisions.

Keywords: artificial intelligence, medical imaging, cancer immunotherapy, radiomics, multimodal biomarkers, radiotherapy

Contents

1	Introduction	5
1.1	Clinical need for more complete immunotherapy biomarkers	5
1.2	Why imaging should be central	5
1.3	Scope and objectives	6
2	Review Strategy and Evidence Appraisal	8
3	Clinical and Biological Context for Imaging AI	11
3.1	A spatial and temporal cancer-immunity system	11
3.2	Established and emerging biomarkers	11
4	Imaging-Centered AI for Cancer Immunotherapy	14
4.1	Dataset and endpoint design	14
4.2	Cohort provenance, curation, and applicability	14
4.3	Acquisition, segmentation, and image representation	16
4.4	Baseline CT radiomics and radiogenomic phenotyping...	17
4.5	Lesion-level, organ-specific, and whole-body heterogeneity	18
4.6	MRI, PET, and quantitative functional imaging	19
4.7	Serial imaging and dynamic response monitoring	21
4.8	Response criteria and label uncertainty	23
4.9	Multimodal integration beyond imaging	26
4.10	Cancer-specific evidence and pan-cancer generalization ..	30
4.11	Performance metrics and clinical interpretation	31
4.12	Imaging and EHR data for toxicity	32
4.13	What current performance establishes	33
4.14	From prognosis to treatment-effect estimation	34
5	Radiotherapy plus ICI as a Focused Imaging Use Case	35
5.1	Context-dependent biological effects	36
5.2	Clinical evidence and the resulting opportunity for AI...	39
5.3	AI for efficacy and response prediction	41
5.4	AI for toxicity and safety prediction	41
5.5	Decision-specific research targets	44

6	Methodological and Translational Requirements	45
6.1	Standardization and leakage control	45
6.2	Validation, calibration, and missing modalities	45
6.3	Reproducibility, multicenter collaboration, and data governance	46
6.4	Interpretability, fairness, and clinical workflow	47
6.5	Prospective clinical utility	49
6.6	Human–AI team evaluation and early clinical studies . . .	49
6.7	Lifecycle monitoring, model updating, and regulatory evidence	50
6.8	Implementation and health-economic value	51
7	Research Priorities	52
8	Conclusion	54

1. Introduction

1.1. Clinical need for more complete immunotherapy biomarkers

Immune checkpoint inhibitors (ICIs) targeting PD-1, PD-L1, and CTLA-4 have produced durable responses in non-small-cell lung cancer (NSCLC), melanoma, renal-cell carcinoma, and mismatch-repair-deficient malignancies.^[1,2] In NSCLC, randomized trials established ICI alone or combined with chemotherapy in biomarker-defined and broader populations.^[3,4] Nevertheless, many patients experience primary resistance, acquired resistance, mixed response across lesions, or immune-related adverse events. The central clinical problem has shifted from whether checkpoint blockade can work to determining which patient, disease state, and treatment context are most likely to produce durable benefit. Answering that question requires more than identifying patients with a good prognosis. It requires estimating how outcome would differ under a clinically relevant alternative—a distinction between prognostic and predictive information that we adopt as a principal axis of appraisal throughout this review (Sections 2 and 4). These biomarkers, their individual limitations, and the ways in which those limitations are complementary rather than additive are examined in Section 3. The relevant point here is structural: no currently deployed assay observes the whole of a disease that is spatially heterogeneous and temporally unstable, and the assays that are most biologically specific are the least repeatable.

PD-L1 expression, tumor mutational burden (TMB), microsatellite instability-high/mismatch repair-deficient (MSI-H/dMMR) status, tumor-infiltrating lymphocytes (TILs), immune gene-expression profiles, and circulating tumor DNA (ctDNA) each characterize a component of the tumor-immune interaction.^[5,6,7] Their limitations are complementary. Tissue assays provide biological specificity but sample a limited region; blood biomarkers are repeatable but depend on shedding, timing, and host factors; conventional imaging samples whole-body disease burden but does not directly measure immune cells or molecular programs. This complementarity motivates imaging-centered multimodal AI.

1.2. Why imaging should be central

CT, MRI, and PET are already embedded in staging, response assessment, treatment planning, and surveillance. Their distinguishing value lies in the combination of in vivo acquisition, repeated assessment during routine care, and separate characterization of visible lesions. A biopsy resolves biology at

one site and one time point; ctDNA is repeatable and systemic but spatially unresolved. Imaging therefore provides a complementary view of between-lesion heterogeneity over time and can additionally represent total burden, lesion shape, necrosis, texture, metabolic activity, body composition, and organ distribution.

AI can transform imaging from a predominantly anatomical assessment into a quantitative phenotype. Handcrafted radiomics summarizes intensity, shape, and texture, whereas deep learning derives image representations directly. Imaging can also serve as the spatial and longitudinal backbone for integrating pathology, genomics, ctDNA, laboratory tests, and electronic health record (EHR) data. The purpose is not unrestricted data fusion; it is estimation of a patient state that supports a defined action, such as treatment selection, response confirmation, local therapy for oligoprogression, or intensified toxicity monitoring.

1.3. Scope and objectives

This review focuses on imaging-centered AI for precision cancer immunotherapy. Although reviews of radiomics in immunotherapy are numerous, this review is organized differently in three respects. First, evidence is organized around the clinical decision a model would inform rather than around modality or architecture. Second, we apply an explicit appraisal scheme that separates evidence maturity from the estimand a model actually targets and use it consistently to interpret the primary studies summarized here. Third, we use RT+ICI as a focused case study because it is the setting in which imaging, lesion identity, dose, sequence, and organ toxicity must all be linked to a defined action, and in which the gap between reported performance and demonstrated utility is especially visible.

We examine baseline radiomics, lesion-level heterogeneity, serial imaging, toxicity, and multimodal integration, emphasizing dataset composition, cancer type, sample size, model contribution, reported performance, validation design, and clinical limitations. Digital pathology, omics, ctDNA, and EHR data are discussed as complementary modalities rather than independent review domains. Radiotherapy (RT) combined with ICI is retained as a focused use case because it requires imaging to connect lesion phenotype, dose distribution, sequence, response, and toxicity. The review does not treat model-guided immunoradiotherapy as an established standard. Section 2 states the appraisal scheme and its limitations. Section 3 establishes which

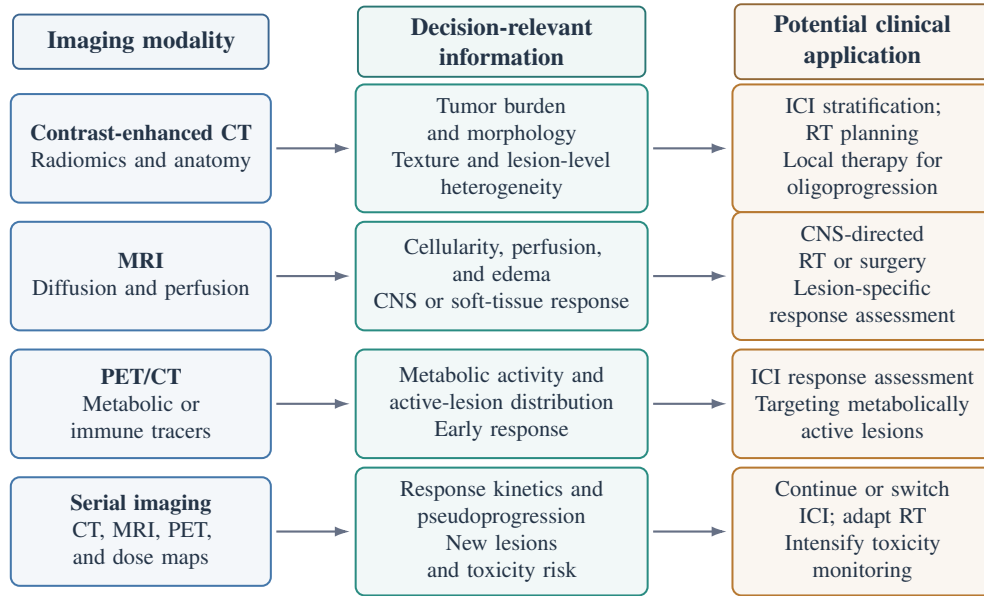


Figure 1: Mapping of imaging modalities to decision-relevant information and potential therapeutic or monitoring applications in precision cancer immunotherapy. Arrows indicate the progression from imaging acquisition to clinical use.

immune and host mechanisms are, and are not, plausibly observable in an image. Section 4 evaluates the primary imaging and multimodal evidence against the scheme. Section 5 applies it to RT+ICI. Sections 6 and 7 set out translational requirements and prioritized study designs.

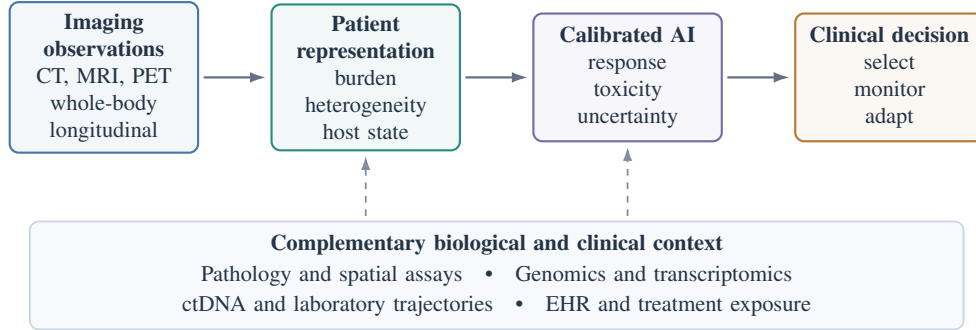


Figure 2: Imaging-centered framework for multimodal decision support in cancer immunotherapy. Imaging provides longitudinal, whole-body observations, while complementary data refine patient representation and model estimation. Clinical use requires calibrated predictions linked to prespecified actions.

2. Review Strategy and Evidence Appraisal

Targeted searches of PubMed, Web of Science, Google Scholar, ClinicalTrials.gov, and relevant journal websites were conducted without a lower publication-date restriction; publications and trial records available through June 30, 2026 were considered. Search terms combined ICI with radiomics, CT, MRI, PET, digital pathology, multimodal AI, ctDNA, EHR, toxicity, response monitoring, RT, chemoradiotherapy, and immunoradiotherapy. Reference lists of pivotal clinical, mechanistic, and computational studies were also screened. Recent imaging-AI and clinical studies were prioritized, while earlier landmark studies were retained when necessary to establish the biological, methodological, and therapeutic context.

Eligibility was determined by relevance to the clinical and translational questions addressed in this review. Human imaging-AI studies were retained as primary evidence when they specified the target population, cancer type, treatment context, imaging or multimodal inputs, clinical endpoint, evaluable cohort, and validation design. RT-ICI clinical studies were retained when they informed efficacy, toxicity, dose or fractionation, treatment sequence, target selection, treatment volume, or exposure of immune-related normal tissues. Mechanistic studies and methodological, reporting, or regulatory documents were included when required to interpret the clinical evidence and its translational limitations. Duplicate reports were consolidated unless they

contributed distinct follow-up, endpoints, or analyses. Non-oncologic reports, studies without substantive relevance to ICI, RT, or imaging-centered AI, and reports lacking sufficient methodological or outcome information for critical appraisal were not retained as primary evidence. Reviews were used primarily for contextual interpretation and citation chaining.

The final synthesis cited 74 unique sources. For transparent accounting, each source was assigned to one primary category, although some informed more than one section:

- (i) clinical and biological foundations of ICI and biomarkers ($n = 17$);
- (ii) imaging-centered AI, response assessment, and multimodal evidence ($n = 23$);
- (iii) RT-ICI mechanisms, clinical evidence, and direct AI studies ($n = 23$);
- (iv) methodological, reporting, and regulatory frameworks ($n = 11$).

Among eligible sources, representative studies were selected to illustrate distinct combinations of population, imaging modality, endpoint, comparator, validation strategy, and intended clinical use, rather than because they reported the highest performance. Geographic, temporal, multicenter, and trial-linked evaluations were weighted more strongly than random internal splits. Because populations, endpoints, and validation schemes differ, performance estimates are presented descriptively and are not pooled.

Evidence was appraised along two independent axes rather than a single ladder. The first axis is *maturity*: (L2a) retrospective association in development or internally split data; (L2b) robust internal evaluation with patient-level partitioning, leakage control, and internal resampling or leave-one-site-out evaluation; (L3) independent external validity in a geographically, temporally, or platform-distinct cohort not used in development; (L4) prospective evaluation of a locked model against a prespecified action, whether in silent or assisted mode; and (L5) demonstrated clinical utility, in which model-guided care improves a patient-relevant outcome relative to the actual multidisciplinary comparator. These stages map onto the translational sequence in Section 6. The second axis is the *estimand*. A model may be *prognostic under treatment*, estimating outcome among patients who receive ICI, or it may target a *treatment effect*, estimating how outcome would differ under a clinically relevant alternative. Models developed only in ICI-treated cohorts were classified as prognostic-under-treatment irrespective of their maturity, unless a relevant control treatment and a design capable of estimating treatment-effect heterogeneity were present (Section 4). The two axes are independent:

Table 1: Evidence-maturity levels used to appraise imaging AI for immunotherapy.

Maturity level	Minimum supporting design	Interpretation
L2a: Retrospective association	Development or internal split in a defined patient population, time point, and outcome	Demonstrates association and discrimination within the sampled setting.
L2b: Robust internal evaluation	Patient-level partitioning with leakage control and internal resampling or leave-one-site-out evaluation	Tests internal robustness; it does not establish transportability to an untouched population.
L3: Independent external validity	Locked model evaluated in geographically, temporally, or platform-distinct data not used in development	Tests transportability; calibration and subgroup performance remain necessary.
L4: Prospective prespecified action	Locked model evaluated prospectively with a prespecified output, threshold, action, and override pathway	Establishes feasibility and actionability in the intended workflow, with or without influencing care.
L5: Demonstrated clinical utility	Prospective comparison of model-guided care with the actual multidisciplinary workflow	Establishes patient-relevant benefit, harm, or resource consequences.

Interpretive note: Biological plausibility is assessed alongside this maturity axis as a precondition for mechanistic interpretation, not as an evidence level. The estimand—prognosis under treatment or differential treatment effect—is assessed independently.

an externally validated, well-calibrated L3 model may still be uninformative for treatment selection, and this combination describes most of the evidence reviewed here. Biological plausibility is assessed separately, as a precondition rather than a level. It constrains interpretation, whether a signal can correspond to a stated mechanism, but it neither substitutes for empirical evidence nor is required for it, and it is therefore reported alongside, not within, the maturity axis.

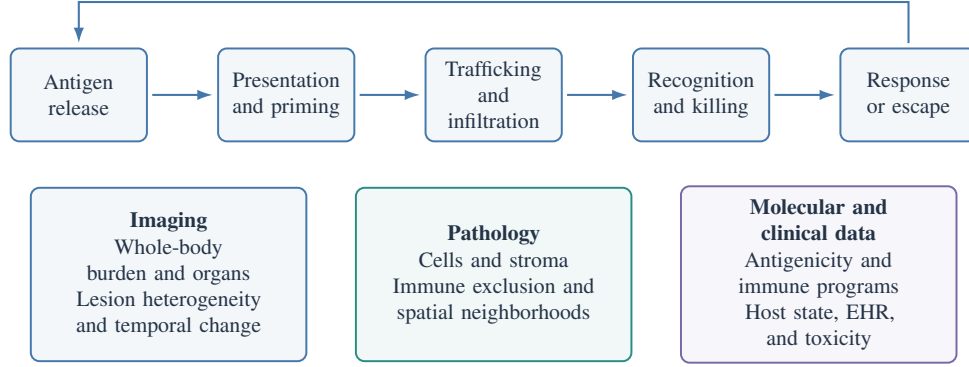


Figure 3: Complementary observation of the cancer–immunity cycle. Imaging supplies whole-body and longitudinal context; pathology resolves local spatial architecture; molecular and clinical data characterize antigenicity, immune programs, host state, and toxicity.

3. Clinical and Biological Context for Imaging AI

3.1. A spatial and temporal cancer–immunity system

The cancer–immunity cycle spans antigen release, presentation, T-cell priming, trafficking, infiltration, recognition, and killing,^[8,9] and resistance can interrupt it at any step—through low antigenicity, defective presentation, impaired interferon signaling, T-cell exclusion, suppressive myeloid or regulatory populations, hypoxia, stromal barriers, exhaustion,^[10,11] or, under treatment pressure, loss of JAK signaling or beta-2-microglobulin.^[12] The question relevant to this review is narrower: which of these steps leaves a signature that an image can plausibly detect, and at what spatial and temporal resolution.

These mechanisms are distributed across lesions and organs and can change during treatment; responses may therefore diverge within the same patient. Pathology describes local cellular and spatial organization. Genomic and transcriptomic assays characterize sampled molecular programs. Blood reflects systemic host state and molecular burden. Imaging repeatedly samples visible disease throughout the body. No modality observes the entire system, and apparent discordance can be informative rather than erroneous.

3.2. Established and emerging biomarkers

PD-L1 is widely used in clinical practice, but its interpretation varies by assay, scoring system, threshold, tissue site, and sampling time. TMB approximates neoantigen opportunity but does not measure antigen presentation

or immune infiltration.^[13,14] MSI-H/dMMR is more biologically specific and supports tissue-agnostic treatment, although response is not universal.^[2,15] TILs and T-cell-inflamed gene-expression profiles characterize immune engagement but remain biopsy- and platform-dependent.^[16,6] Early changes in ctDNA can provide a longitudinal molecular response signal, but interpretation depends on tumor shedding, disease site, assay characteristics, and sampling time.^[7,17]

Imaging should therefore be viewed as complementary to, rather than a replacement for, PD-L1, MSI-H/dMMR, and other molecular biomarkers. Its value depends on whether it adds reproducible, well-calibrated, and clinically actionable information beyond established biomarkers and routine clinical variables.

Table 2: Established biomarkers and the complementary contribution of imaging.

Measure	Principal information	Main limitation	Imaging contribution
PD-L1 IHC	Local checkpoint-ligand expression; established companion or complementary use in several cancers	Assay, threshold, site, temporal, and intratumoral variability	Quantifies whole-body burden and inter-lesion phenotype; cannot replace validated PD-L1 testing.
TMB	Somatic mutation burden and potential neoantigen opportunity	Platform and threshold variation; does not measure presentation or immune engagement	May identify phenotypes associated with burden, necrosis, and heterogeneity for multimodal models.
MSI-H/dMMR	Defective mismatch repair with strong tissue-agnostic clinical evidence	Low prevalence in many cancers; response is not universal	Adds burden, site, and longitudinal response information after molecular eligibility is established.
TILs and immune gene expression	Local immune infiltration and inflammatory programs	Requires tissue; spatial and platform dependence	Samples multiple lesions and can guide where tissue characterization may be most informative.
ctDNA	Repeatable molecular burden and clonal dynamics	Variable shedding, assay sensitivity, and sampling schedule	Localizes lesion behavior and helps distinguish global molecular change from mixed anatomical response.
Routine clinical and laboratory data	Host state, comorbidity, inflammation, organ function, and treatment exposure	Missingness, documentation variation, and limited spatial specificity	Provides anatomical and temporal context for response and organ-specific toxicity.
CT, MRI, and PET	Whole-body disease distribution, morphology, function, and change over time	Acquisition, reconstruction, segmentation, and response-criterion variability	Serves as the spatial backbone for multimodal prediction and longitudinal decision support.

Abbreviations: ctDNA, circulating tumor DNA; dMMR, deficient mismatch repair; IHC, immunohistochemistry; MSI-H, microsatellite instability-high; PD-L1, programmed death-ligand 1; TILs, tumor-infiltrating lymphocytes; TMB, tumor mutational burden.

4. Imaging-Centered AI for Cancer Immunotherapy

4.1. Dataset and endpoint design

A defensible imaging dataset begins with the decision that the model is intended to support. Each study should specify the target population, index treatment, time zero, input window, prediction horizon, unit of prediction, and estimand before model development. A baseline model used before ICI initiation, a landmark model used at the first restaging examination, and a lesion-level model used to identify a resistant site answer different questions even when they use images from the same patient.

Imaging AI studies also target different constructs: a biological phenotype such as CD8-related expression; objective response or durable clinical benefit; progression-free or overall survival; lesion-level response; early treatment change; or toxicity. These targets are not interchangeable. A response classifier requires a reproducible assessment window and rules for death or treatment discontinuation before imaging; a survival model must address censoring and competing events; and a toxicity model must define organ, grade, and prediction horizon. Treatment-effect prediction additionally requires a clinically relevant alternative treatment and cannot be inferred from an ICI-treated cohort alone.

The effective sample size is constrained primarily by the number of independent patients and outcome events, not the raw number of lesions, slices, or image patches. Acquisition protocol, reconstruction, contrast timing, segmentation, cancer site, treatment line, and assessment schedule can create technical shortcuts unless they are represented and evaluated explicitly. Reports should provide patient and event counts, confidence intervals, and, where relevant, site- or time-separated evaluation. The operational requirements for patient-level partitioning and leakage control are specified in Section 6.

4.2. Cohort provenance, curation, and applicability

Dataset quality begins with an explicit intended-use population. A cohort defined only as “patients who received immunotherapy” may combine different cancers, stages, treatment lines, ICI agents, chemotherapy combinations, previous local therapy, and imaging schedules. These factors alter baseline risk and the interpretation of response. Eligibility criteria should therefore specify the index regimen, permitted combinations, treatment line, measurable-disease requirements, baseline and follow-up windows, and

whether patients with brain-only, bone-only, or otherwise non-measurable disease are represented.

The cohort assembly diagram is part of the scientific result. Investigators should report the number of potentially eligible patients, availability and timing of baseline and follow-up imaging, image-quality exclusions, segmentation failures, missing biomarker data, early deaths, treatment changes, and outcome events. Requiring a complete protocol-timed scan pair can condition the cohort on remaining alive, sufficiently well, and under observation until follow-up. Similarly, a model trained only on patients with contrast-enhanced CT and complete sequencing may exclude patients with renal impairment, rapidly progressive disease, transfer of care, or limited tissue. Comparisons between included and excluded patients are therefore needed to define the population to which performance estimates apply.

Data granularity must match the analysis. One patient can contribute multiple examinations, lesions, image series, segmentations, and patches; these observations enrich representation but do not create independent patients. Development reports should provide both patient counts and lower-level counts, describe clustering, and use patient-level resampling. In multi-center studies, the distribution of patients and events across institutions is more informative than the number of image files or nominal sites, because one institution may dominate model fitting and another may contribute too few events for a stable site-specific estimate.

Time zero must be aligned across images, biomarkers, treatments, and outcomes. A pretreatment model should not include laboratory results obtained after ICI initiation or contours selected with knowledge of subsequent response. Longitudinal models require an explicit landmark at which patients are alive, observed, and eligible for the modeled decision; predictions should use only information available by that landmark. Censoring, subsequent treatment, loss to follow-up, and the handling of scans obtained outside the prespecified window should be reported.

Applicability should be examined across the dimensions that can change the clinical meaning of a score: patient population, cancer type, regimen, treatment line, institution, scanner and protocol, anatomical site, and follow-up practice. An ICI-treated cohort generally supports prediction conditional on receiving ICI. A claim of treatment-selection value requires a relevant comparator and a design that estimates treatment-effect heterogeneity. Randomized data provide the clearest basis; observational comparisons require explicit confounding assumptions and sensitivity analyses. Without such con-

trols, a highly accurate model may primarily identify general disease burden, inflammation, or host reserve rather than ICI-specific sensitivity.

4.3. Acquisition, segmentation, and image representation

CT datasets used in immunotherapy research commonly mix routine staging studies acquired with different tube current, reconstruction kernel, field of view, slice thickness, and intravenous contrast phase. These differences affect both handcrafted radiomics and learned features. Restricting the cohort to one protocol improves internal consistency but narrows generalizability; accepting routine variation improves realism but increases technical variance. A defensible analysis should state this trade-off, preserve the native test distribution, and avoid using test outcomes to decide which protocols to exclude.

Preprocessing should be reported as a reproducible sequence. Relevant operations include series selection, orientation, voxel resampling, intensity handling, denoising, registration, segmentation, cropping, and normalization. Radiomic feature definitions should follow standardized conventions such as the Image Biomarker Standardisation Initiative.^[18] Feature-stability filtering can be based on repeat scans, reconstruction experiments, or contour perturbations, but it must occur entirely within development data. Post hoc removal of unstable features after reviewing external performance compromises the independence of the test.

Segmentation determines what the model can observe. A primary-tumor contour excludes metastatic and host information; a target-lesion approach inherits the selection rules of conventional response assessment; and whole-body tumor segmentation offers more complete burden but introduces detection errors and greater annotation cost. The segmentation source should be described as manual, semiautomatic, or automated, together with reader expertise, quality control, and the handling of missed or non-measurable disease. When an automated contour is corrected clinically, the frequency and magnitude of correction are relevant workflow outcomes.

Model input design also encodes assumptions. Two-dimensional slices are computationally efficient but may miss craniocaudal extent and encourage slice-level leakage. Three-dimensional crops preserve volumetric context but increase memory requirements and can include variable surrounding anatomy. Multi-instance approaches can represent several lesions without forcing a fixed count, whereas whole-volume models may learn organ, body-composition, and scanner context. Comparisons should be performed at the patient level and under realistic variation in lesion count, size, and site.

Self-supervised or transfer learning can reduce dependence on outcome labels, but pretraining provenance matters. Pretraining on large clinical archives may improve feature stability while importing site, population, and disease biases. Reports should identify the pretraining source, overlap checks, frozen or fine-tuned components, computational requirements, and performance relative to simpler radiomic and clinical baselines. Dataset scale alone does not establish transferability.

4.4. Baseline CT radiomics and radiogenomic phenotyping

Baseline CT currently provides the most mature evidence base for imaging-centered AI in cancer immunotherapy. Sun et al. developed a CT radiomic signature linked to RNA-sequencing estimates of CD8-related immune infiltration in 135 patients and evaluated it in an independent ICI cohort of 137 patients across cancers.^[19] The reported AUC was 0.74 in training and 0.67 in validation for the CD8 phenotype; a higher score was associated with ICI response and overall survival. The innovation was to anchor a noninvasive imaging phenotype to molecular immune information. The modest validation discrimination and heterogeneous cancer types caution against interpreting the signature as a universal measure of CD8 infiltration.

Trebeschi et al. analyzed 203 patients, including 123 with NSCLC and 80 with melanoma, and 1,055 lesions.^[20] Lesion-level AUC reached 0.83 in NSCLC, whereas performance in melanoma lymph-node lesions was approximately 0.64; patient-level performance reached approximately 0.76. The principal contribution was demonstration that image phenotype and predictability vary by lesion and anatomical site. This result supports organ-aware modeling and explicit aggregation rather than assigning one undifferentiated score to a heterogeneous patient.

Deep-CT extended routine CT assessment to 976 patients with metastatic EGFR/ALK-wild-type NSCLC, including external testing at Stanford.^[21] Combining imaging and clinical variables increased the overall-survival C-index from 0.70 to 0.75. The cohort scale, external evaluation, and use of routine imaging improve relevance. Overall survival, however, reflects disease burden, host fitness, subsequent therapy, and treatment response; performance in an ICI-treated cohort should therefore be interpreted as prognostic under treatment unless differential benefit against a comparator is demonstrated.

These studies do not constitute a common benchmark: they use different biological targets, cancers, input regions, endpoints, and validation designs. Radiomics remains attractive because it can be derived retrospectively from routine images, but feature values can change with reconstruction kernel, slice thickness, contrast phase, motion, voxel resampling, discretization, and contour definition. Robust development should prespecify acquisition windows, lock preprocessing, quantify feature stability, and evaluate performance before and after any harmonization. Harmonization fitted with outcome or test-cohort information creates leakage.

Deep-learning representations reduce dependence on predefined features but do not remove technical shift. Networks can learn scanner, annotation, institutional, or background-anatomy signatures. External testing should preserve native acquisition and report performance by site, scanner, protocol, cancer type, anatomical site, lesion size, and relevant demographic subgroups when sample size permits. Comparison with tumor diameter, total burden, stage, performance status, and PD-L1 is necessary to establish incremental value.

4.5. Lesion-level, organ-specific, and whole-body heterogeneity

Mixed or dissociated response denotes simultaneous regression or stability of some lesions and growth or emergence of others.^[22] Oligoprogression is a narrower clinical state in which a limited number or distribution of sites progresses while most disease remains controlled during active systemic therapy.^[23] Terminology and thresholds vary across studies, and the patterns can overlap, but they are not synonyms: mixed response can be widespread and unsuitable for local treatment, whereas an oligoprogression label implies a prespecified limit on progressing disease and a clinically meaningful background of control. Studies should define lesion count, organ distribution, timing, prior disease state, and whether local treatability is part of the operational definition rather than importing an unstated threshold.

This distinction changes both annotation and modeling. A lesion-level model should retain lesion identity, anatomical site, target or non-target status, prior local treatment, and serial state. A patient-level model must then specify how lesion information is aggregated. The maximum-risk lesion, burden-weighted mean, organ-specific score, fraction of progressing burden, and learned multi-instance model answer different questions. Aggregation rules should be prespecified or validated against lesion-level and patient-level outcomes, rather than selected solely for the best patient-level AUC.

Incomplete lesion representation is itself a source of bias. Manual or automated pipelines may preferentially omit small, infiltrative, poorly contrasted, confluent, or non-measurable lesions. Whole-body studies should report the number, size, site, and reason for exclusion of lesions and should quantify unmatched disease at follow-up. Evaluation should distinguish lesion detection, segmentation, lesion-level state estimation, patient-level response classification, and the downstream decision; high segmentation overlap does not establish correct response interpretation.

Anatomical site is not merely a nuisance covariate. Bone lesions may be non-measurable by RECIST and treatment-related sclerosis can complicate change assessment; nodal enlargement can reflect tumor or inflammation; liver metastases can carry distinct prognostic and immune context; and pleural or meningeal disease may be poorly represented by a target-lesion sum. A whole-body system should permit site-specific uncertainty or abstention and should report site-stratified discrimination and calibration, especially when a score may be used to select a lesion for biopsy or local therapy.

Intracranial disease requires a trajectory separate from extracranial disease because contrast-enhanced MRI findings, neurological status, corticosteroid exposure, surgery, and local radiotherapy can change the interpretation of response. A systemic CT response should not silently overwrite a discordant CNS trajectory; the relevant organ-specific criteria and adjudication fields are detailed in the response-labeling subsection below.

Whole-body imaging also captures host phenotype. Skeletal muscle, adipose distribution, emphysema, interstitial lung abnormality, vascular calcification, and organ reserve may improve prediction of survival or toxicity while reducing treatment specificity because they are relevant across therapies. Studies should separate tumor phenotype, host phenotype, and general disease burden where possible and test whether each component adds information beyond routine clinical variables or interacts with treatment.

4.6. MRI, PET, and quantitative functional imaging

MRI and PET should be positioned by clinical question rather than grouped as uniformly superior alternatives to CT. MRI is the standard anatomical modality for intracranial response assessment and has important organ-specific roles; it can also characterize diffusion, perfusion, permeability, edema, hemorrhage, and treatment-related change. PET provides whole-body metabolic or target-specific measurements. Their additional biological sensitivity is useful only when acquisition, interpretation, and the

downstream action are defined.

Recent MRI studies illustrate both promise and limits. Lin et al. used intratumoral and peritumoral pretreatment MRI radiomics in 172 patients with head-and-neck squamous-cell carcinoma who received neoadjuvant chemoimmunotherapy followed by surgery; the combined model achieved an AUC of 0.849 in an external cohort for pathological complete response.^[24] The pathological endpoint and external cohort strengthen the study, but the model predicts response to a combined regimen and does not isolate an ICI-specific treatment effect. In a retrospective lesion-level analysis, Sanvito et al. analyzed 125 brain metastases in 93 patients and found diffusion–contrast-enhancement phenotypes associated with ICI exposure and, in a 23-lesion histology subset, with CD45-positive immune-cell density; corticosteroid exposure modified these associations.^[25] This supports biological plausibility for MRI phenotyping but is not yet evidence for a locked response-prediction tool or a treatment-changing threshold.

For FDG PET, inflammatory immune-cell recruitment, infection, granulomatous reactions, and immune-related adverse events can increase uptake, while necrosis and small-lesion partial-volume effects can reduce it. Joint practice guidance therefore emphasizes standardized acquisition and immune-treatment context rather than interpreting uptake as viable tumor in isolation.^[26] In a prospective study of 109 patients with NSCLC, Masse et al. obtained FDG PET/CT at baseline, approximately 7 weeks, and 3 months and evaluated pseudoprogression and dissociated response under a protocol that allowed planned reassessment after early metabolic progression.^[27] The study supports serial, protocolized adjudication of atypical PET patterns; it does not make a single early PET category a definitive biological label. Conventional PET Response Criteria in Solid Tumors (PERCIST) and immune-adapted proposals should therefore be recorded as explicit metabolic assessment rules, not silently merged with anatomical RECIST categories or treated as universal biological truth.

Target-specific immunoPET is moving from feasibility toward early clinical validation but remains investigational. In the prospective phase II iCorrelate study, 49 patients contributed 70 matched CD8 PET–biopsy pairs; tracer uptake showed a modest correlation with CD8 immunohistochemistry overall and a stronger correlation in the renal-cell carcinoma subgroup.^[28] The result illustrates the appeal of whole-body immune-cell mapping and the difficulty of validating a spatially distributed signal against core biopsies; the cohort size, sampling mismatch, tracer logistics, and absence of a validated decision

rule preclude autonomous treatment selection.

MRI and PET also introduce modality-specific technical variation. MRI values depend on vendor, field strength, coil, sequence parameters, b-values, contrast timing, distortion correction, and motion; PET depends on injected activity, uptake time, blood glucose, reconstruction, attenuation correction, calibration, and lesion size. Retrospective complete-case analyses may select patients with unusually complete sequences or protocol-conforming scans. Studies should report sequence or tracer availability, acquisition quality, harmonization, minimum evaluable lesion size, missing-modality handling, and performance before and after technical adjustment.

The near-term role of AI in these modalities is most defensible for reproducible quantification, lesion matching, whole-body burden, serial change, uncertainty estimation, and integration with routine anatomical imaging. Specialized acquisitions should demonstrate incremental value over CT, standard MRI interpretation, laboratory data, and simpler quantitative measures. If the additional sequence or tracer does not improve a prespecified decision, greater biological sophistication alone does not justify broad deployment.

4.7. Serial imaging and dynamic response monitoring

Baseline prediction cannot represent treatment-induced change. Serial imaging can quantify trajectories of volume, density, texture, metabolism, organ distribution, and host phenotype. Sako et al. developed a fully automated serial CT response score from pretreatment and approximately 12-week scans in 1,830 patients with advanced NSCLC, using routine-practice development and multicenter test cohorts plus an independent clinical-trial cohort.^[29] The score improved overall-survival stratification relative to RECIST and tumor-volume change. This is strong external prognostic evidence, but the analysis is conditional on availability of an early follow-up scan and does not show that changing treatment at a score threshold improves outcome.

Longitudinal tasks should be separated. Early risk updating, confirmation of an equivocal first progression, recognition of a mixed pattern, identification of acquired resistance or oligoprogression, and toxicity detection have different labels and actions. Each task requires an explicit landmark and prediction horizon. Patients who progress, die, discontinue treatment, or leave the health system before the landmark cannot simply be omitted without changing the target population; they should be represented in cohort flow and handled according to the prespecified estimand.

Longitudinal representation can use explicit difference images, trajectories of engineered features, recurrent or attention-based sequence models, or survival models with time-varying covariates. Explicit change scores are transparent but sensitive to registration and measurement noise. Learned sequence models can represent nonlinear evolution but require adequate patient numbers, actual calendar time, and careful masking. Scans obtained at 6 and 16 weeks should not be treated as equally spaced adjacent observations unless that assumption is explicitly modeled and tested.

Registration and lesion matching deserve separate reporting. Rigid registration may be adequate for compact anatomy but is unreliable for deforming lung, bowel, postoperative cavities, and changing effusions. Deformable registration can improve local alignment while creating non-biological intensity or shape changes. Lesions can merge, split, disappear, recur, or arise de novo; treated lesions may no longer follow the natural systemic trajectory. Studies should specify matching rules, retain new and unmatched lesions, quantify matching uncertainty, and avoid analyzing only lesions that are easy to follow.

Prediction updates must be evaluated dynamically. A model that produces baseline and 12-week risks should show the incremental calibration and discrimination of the update, lead time before conventional progression, and whether a prespecified action would change. Net reclassification alone is insufficient when reclassified patients would receive the same care. External evaluation should reproduce the intended scan windows and decision point rather than retrospectively choose the most informative examination.

A major longitudinal failure mode is informative observation. Imaging frequency increases when symptoms emerge or response is uncertain and decreases when patients deteriorate, transfer care, or enter hospice. Treating missing visits as random can encode the care process rather than disease biology. Models should describe scheduled and unscheduled examinations, consider observation patterns explicitly, compare protocol and real-world follow-up, and test performance in patients with irregular imaging.

Representative studies spanning baseline CT, organ-specific MRI, PET, and serial CT are summarized in Table 3.

Table 3: Representative imaging-centered studies for model development and longitudinal response assessment during immune-checkpoint inhibitor therapy.

Study; modality and setting	Dataset and analytic target	Principal result	Interpretation for this review
Sun et al. ^[19] ; CT, pan-cancer	RNA-anchored CD8 radiomics; $n = 135$ development and $n = 137$ ICI validation	CD8-phenotype AUC 0.74 and 0.67; score associated with response and OS	Innovative molecular anchoring, but modest external discrimination and heterogeneous cancers limit use as a universal immune measure.
Trebeschi et al. ^[20] ; CT, NSCLC and melanoma	203 patients and 1,055 lesions; lesion- and site-aware response modeling	Lesion AUC up to 0.83 in NSCLC; patient AUC about 0.76	Establishes inter-lesion and organ-specific heterogeneity; small site subsets yield unstable estimates.
Deep-CT ^[21] ; CT, metastatic NSCLC	Routine CT plus clinical deep learning; $n = 976$ with external testing	OS C-index increased from 0.70 to 0.75	Larger external evidence, but OS remains a prognostic endpoint and treatment specificity was not established.
Lin et al. ^[24] ; MRI, HNSCC	Pretreatment intratumoral and peritumoral radiomics; $n = 172$ receiving neoadjuvant chemoimmunotherapy and surgery	External AUC 0.849 for pathological complete response	Pathology-anchored multicenter evidence for a combined regimen; it does not isolate the ICI contribution.
Sanvito et al. ^[25] ; MRI, brain metastases	125 lesions in 93 patients; diffusion-enhancement phenotypes with histology in 23 lesions	Phenotypes associated with ICI exposure and CD45-positive-cell density	Supports biological plausibility; steroids and prior local therapy complicate interpretation, and no locked clinical prediction rule was tested.
Masse et al. ^[27] ; FDG PET/CT, NSCLC	Prospective $n = 109$ with baseline, approximately 7-week, and 3-month imaging	Characterized pseudoprogression and dissociated metabolic response under planned reassessment	Supports protocolized serial PET adjudication rather than a definitive single-scan label.
Sako et al. ^[29] ; serial CT, advanced NSCLC	Pretreatment and approximately 12-week CT; $n = 1,830$ across routine-practice and trial cohorts	OS C-index 0.73 in routine-practice testing and 0.71 in trial validation	Strong longitudinal prognostic validation; conditional on reaching follow-up, with no trial of score-guided treatment change.

Abbreviations: AUC, area under the receiver operating characteristic curve; CT, computed tomography; FDG, fluorodeoxyglucose; HNSCC, head-and-neck squamous-cell carcinoma; ICI, immune-checkpoint inhibitor; MRI, magnetic resonance imaging; NSCLC, non-small-cell lung cancer; OS, overall survival; PET, positron-emission tomography.

4.8. Response criteria and label uncertainty

RECIST 1.1 standardized anatomical response assessment and remains an essential comparator for imaging AI.^[30] It is not a biological ground truth. Diameter-based categories compress continuous and lesion-specific change, exclude some bone, pleural, meningeal, and small lesions from target measurement, and depend on target selection and reader measurement. A model trained against a RECIST category may therefore reproduce the criterion’s

measurement behavior rather than identify an independent immune phenotype.

iRECIST introduced standardized trial handling of unconfirmed and confirmed progression during immunotherapy.^[31] For dataset construction, its main contribution is temporal discipline: the first apparent progression, new lesions, clinical stability, confirmatory assessment, and subsequent state should be stored explicitly. Continued treatment after unconfirmed progression is considered only in trials that permit it and for clinically stable patients; iRECIST itself is a trial data standard, not evidence that treatment beyond progression benefits a particular patient. An AI model should not convert an iRECIST state directly into a treatment recommendation.

Pseudoprogression is clinically important but context-dependent and should not be treated as the default explanation for early enlargement. At the first enlarging examination, pseudoprogression is not an observable ground truth. The contemporaneous label should remain unconfirmed or indeterminate progression, with a separate flag for suspected immune-related change. A retrospective pseudoprogression label requires prespecified subsequent imaging and clinical evidence, supplemented by pathology when available, that is sufficient to distinguish it from true progression or another cause. Cases without adequate verification should retain their criterion state and verification status rather than be forced into a binary pseudoprogression-versus-progression label.

Response annotation should separate four layers that are often collapsed: raw lesion measurements and lesion identities; the criterion-based patient state, such as RECIST or iRECIST; the cross-lesion pattern, such as concordant response, mixed or dissociated response, oligoprogression candidate, or CNS-systemic discordance; and the certainty of the interpretation. Mixed response and oligoprogression should be encoded as patterns, not as interchangeable alternatives to RECIST categories. An oligoprogression candidate additionally requires a prespecified definition of limited progression during otherwise controlled disease and should not imply that local therapy is beneficial unless that action has been separately evaluated.

Organ-specific assessment standards should be retained rather than collapsed into one label. Intracranial disease may require RANO-BM assessment on MRI, with iRANO confirmation principles when applicable, whereas extracranial disease is assessed with RECIST-based criteria.^[32,33] Brain lesions exposed to surgery, stereotactic radiation, corticosteroids, or anti-angiogenic therapy need treatment-context fields; bone-only or

otherwise non-measurable disease requires an explicit non-target framework; and inflammatory nodal or pulmonary changes may require alternative diagnoses. If organ-specific and systemic states disagree, the aggregation rule should be prespecified and the discordance retained.

Label uncertainty has at least three sources: measurement and reader variability; biological ambiguity among viable tumor, inflammation, edema, necrosis, infection, and treatment effect; and incomplete verification caused by missing, delayed, or informative follow-up. Independent review and adjudication can reduce but not eliminate these sources. Datasets should report reader disagreement, adjudication rules, and the proportion of indeterminate or technically non-evaluable examinations. Probabilistic, ordinal, interval-censored, or multistate labels and sensitivity analyses across defensible definitions are preferable to silent exclusion or majority-vote dichotomization when the underlying state is not identifiable.

Outcome choice ultimately determines what the model learns. Objective response emphasizes shrinkage; durable clinical benefit adds a time threshold; progression-free survival incorporates new lesions and clinical decisions; and overall survival is influenced by subsequent therapy and host reserve. A rigorous study should prespecify one primary target and use alternative endpoints as sensitivity analyses. Concordance across endpoints strengthens interpretation, whereas discordance can reveal whether the model primarily captures tumor shrinkage, general prognosis, the observation process, or clinician behavior. Figure 4 translates these principles into a longitudinal annotation and adjudication workflow.

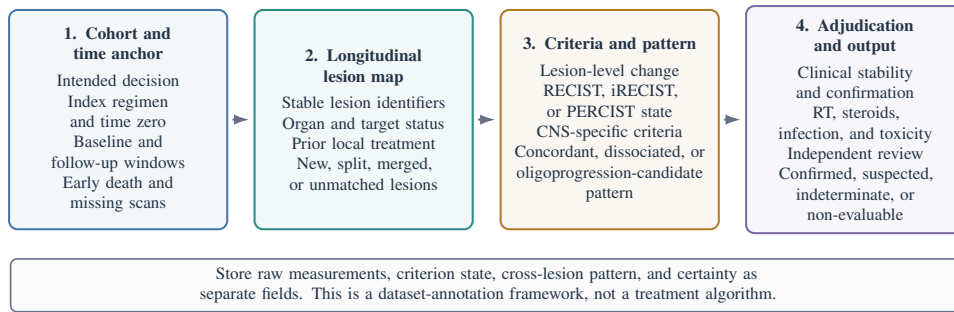


Figure 4: Longitudinal annotation and adjudication framework for ICI imaging datasets. It preserves lesion identity, organ-specific criteria, cross-lesion patterns, and verification uncertainty rather than collapsing these into one binary label.^[30,31,26,32,33,22,23]

4.9. Multimodal integration beyond imaging

Imaging provides spatial and longitudinal characterization of disease but has limited biological specificity. Digital pathology characterizes local tissue architecture and immune context; genomic and transcriptomic assays capture molecular programs; ctDNA can reflect molecular burden and evolving resistance; and laboratory and EHR data represent systemic inflammation, comorbidity, medication exposure, and toxicity history. Multimodal integration should be justified by demonstrable incremental value, biological plausibility, and robustness when one or more inputs are unavailable.

Johannet et al. combined pretreatment H&E slides and clinical variables in metastatic melanoma.^[34] The development cohort included 121 patients at New York University and the external Vanderbilt cohort 32 patients; AUCs were 0.803 and 0.793. The external validation was important, although the external sample was small. Deep-IO later extended histology-based ICI modeling to multicenter NSCLC cohorts.^[35]

Vanguri et al. integrated CT, PD-L1 pathology, genomics, and clinical data from 247 patients with advanced NSCLC.^[36] The multimodal model achieved an AUC of 0.80, compared with 0.73 for PD-L1 and 0.61 for TMB. In a preprint, Yeghaian et al. combined pre- and on-treatment blood measurements, medications, and CT-derived organ volumes in 694 immunotherapy-treated patients across cancers.^[37] Their temporal multimodal model reported AUCs of 0.84, 0.83, 0.82, and 0.81 for mortality at 3, 6, 9, and 12 months, respectively. These results show complementary longitudinal signal but remain prognostic and require peer-reviewed external validation.

LORIS used six routinely available clinical and genomic features in 2,881 patients treated with ICIs and 841 patients not treated with ICIs, spanning 18 solid tumor types.^[38] Inclusion of the non-ICI cohorts enabled a partial assessment of treatment specificity. SCORPIO studied 9,745 patients across 21 cancers, with model development in 1,628 patients and validation in external real-world cohorts and 10 phase III trials.^[39] Time-dependent survival AUCs ranged from 0.725 to 0.763, and clinical-benefit AUCs ranged from 0.641 to 0.714. Performance in non-ICI arms indicated that a substantial component of the model signal was prognostic rather than treatment-specific.

Fusion strategy affects robustness. Early fusion concatenates all features but can be unstable when modalities differ in scale, quality, and missingness. Late fusion combines modality-specific predictions and permits independent quality control and fallback pathways. Joint representation learning can model cross-modal interactions but requires larger cohorts and careful

regularization. Attention coefficients should not be treated as explanations without stability analysis and orthogonal biological support.

Complete-case analysis can select an atypically well-characterized population. Patients without evaluable tissue, broad sequencing, or protocol-timed imaging may differ systematically from those with complete data, so development reports should describe why each modality is missing and compare included with excluded patients. Section 6 addresses the technical evaluation of modality dropout and prespecified fallback models.

Incremental value should be tested sequentially: routine clinical baseline, imaging alone, each additional modality, a parsimonious combination, and the full model. Improvements in AUC should be accompanied by calibration, reclassification where appropriate, decision-curve analysis, and workflow cost. A complex model is justified only if the additional information changes a safe and beneficial decision.

Digital pathology has a special role because it provides cellular and spatial resolution that conventional radiology cannot. Whole-slide models can quantify tumor architecture, lymphocyte-rich regions, necrosis, stroma, and spatial neighborhoods, while radiology describes the distribution of all visible lesions. Their combination is conceptually attractive: the slide supplies local biological specificity and the scan supplies whole-body context. However, the modalities are rarely sampled from identical tissue at identical times. A resection specimen, small biopsy, and metastatic CT lesion may represent different disease states.

Spatial mismatch should therefore be modeled rather than ignored. When the biopsied lesion is known, lesion-level image features can be linked directly to the sampled site and distinguished from patient-level burden. When the site is unknown, pathology should be treated as a patient-level partial observation. Attention-based fusion does not solve this identification problem; a high attention weight can reflect predictive convenience rather than correct biological correspondence.

Genomic and transcriptomic variables introduce additional dimensionality. TMB, specific pathways, copy-number changes, expression signatures, and HLA-related features can be biologically informative, but feature selection in small cohorts is unstable. Restricting molecular inputs to prespecified, reproducible variables may outperform an unconstrained omics pipeline in external data. If dimensionality reduction is learned, it must be fitted inside each development resample and locked before validation.

Routine clinical data should remain a strong comparator. Stage, treat-

ment line, performance status, burden, laboratory inflammation, corticosteroid exposure, antibiotics, and comorbidity can explain substantial outcome variation. An imaging model that is not compared with these inputs may appear more useful than it is. Conversely, a modest improvement over a strong clinical baseline can still matter when imaging is already acquired and the model is automated, calibrated, and inexpensive to run.

Multimodal systems need a quality-control layer as well as a prediction layer. The system should verify image series, segmentation, pathology tissue adequacy, assay completeness, units, temporal alignment, and out-of-distribution inputs. It should then provide the prediction, uncertainty, contributing modalities, and any fallback used. Silent substitution of a missing or corrupted modality can create confident but uninterpretable outputs.

Table 4: Representative multimodal AI studies and the information added beyond imaging alone.

Study; population	Modalities and innovation	Reported performance	Principal limitation
Vanguri et al. ^[36] ; 247 NSCLC	CT radiomics combined with PD-L1 and TMB	Response AUC 0.80	Retrospective treated cohort; no direct estimate of differential ICI benefit.
Yeghaian et al. ^[37] ; 694 pan-cancer patients (preprint)	Pre- and on-treatment blood, medications, and CT-derived organ volumes	Mortality AUC 0.84 at 3 months; 0.81 at 12 months	Prognostic preprint; external transportability and missing-modality behavior require testing.
LORIS ^[38] ; 2,881 ICI and 841 non-ICI patients, 18 cancers	Six routinely available clinical and genomic variables; non-ICI controls examined treatment specificity	Performance varied by cohort and cancer	Pan-cancer averages may obscure disease-specific calibration; residual prognostic signal remains.
SCORPIO ^[39] ; 9,745 patients, 21 cancers	Routine blood and clinical variables; external real-world and 10 phase III trial validations	Survival AUC 0.725–0.763; benefit AUC 0.641–0.714	Performance in non-ICI arms showed that substantial signal was prognostic.
Johannet et al. ^[34] ; melanoma	Histology representations combined with clinical variables	External AUC 0.793	Tissue availability and small external sample limit generalization.
Lippenszky et al. ^[40] ; more than 2,200 ICI-treated recipients	Longitudinal routine EHR data for organ-specific adverse events	AUC 0.729–0.755 for selected toxicities	Requires explicit time horizons, calibration, diagnostic adjudication, and an intervention pathway.

Abbreviations: AUC, area under the receiver operating characteristic curve; CT, computed tomography; EHR, electronic health record; ICI, immune-checkpoint inhibitor; NSCLC, non-small-cell lung cancer; PD-L1, programmed death-ligand 1; TMB, tumor mutational burden.

4.10. Cancer-specific evidence and pan-cancer generalization

NSCLC has dominated imaging-based AI studies of ICI response because CT is routinely acquired, pulmonary lesions are frequently measurable, and relatively large biomarker-defined cohorts are available across multiple treatment lines. Although this concentration has facilitated the development of larger datasets, it may also create a misleading impression that imaging AI for ICI has reached pan-cancer maturity. Melanoma, urothelial carcinoma, renal-cell carcinoma, head-and-neck cancer, and gastrointestinal tumors differ substantially in imaging modality, metastatic distribution, response assessment criteria, treatment regimens, and the prevalence and clinical relevance of molecular biomarkers. Therefore, a model developed using thoracic CT in NSCLC should not be transferred to another malignancy solely because both patient populations receive PD-1/PD-L1 blockade.

Pan-cancer modeling can increase sample size and identify shared patterns, but cancer type may dominate the prediction. Hierarchical approaches can share representations while allowing cancer-, organ-, or regimen-specific intercepts and effects, provided that the hierarchy is prespecified and adequately supported by outcome events. Stratified calibration is necessary, and very small groups should trigger wider uncertainty or abstention rather than unstable disease-specific claims. External validation should test both new institutions within the same cancer and genuinely new cancers because performance averaged across diseases can conceal failure in smaller groups.

Organ site also requires explicit treatment. Liver, lung, bone, brain, lymph-node, and soft-tissue metastases differ in image appearance, measurement reliability, immune microenvironment, and clinical risk. Lesion-level models should report site-specific results and should avoid comparing AUCs from very small organ subsets as if they were stable estimates.

For melanoma, external lesions and nodal or visceral metastases produce a broader range of imaging appearances than a thoracic-only NSCLC cohort, and brain involvement is clinically important. In renal-cell carcinoma, enhancement, necrosis, vascularity, and heterogeneous response may require multiphase CT or MRI. Urothelial and head-and-neck cancers present distinct primary-site anatomy, treatment combinations, and inflammatory confounders. Gastrointestinal cancers include molecularly defined subgroups for which MSI-H/dMMR testing remains central. These differences support disease-specific model components rather than one undifferentiated pan-cancer score.

Treatment regimen is equally important. ICI monotherapy, dual-checkpoint blockade, chemoimmunotherapy, antiangiogenic combinations, and consolidation after chemoradiotherapy have different response rates, toxicity patterns, and assessment schedules. A model trained in one regimen may remain prognostic in another while becoming poorly calibrated. External validation should report regimen-specific calibration and should not pool therapies solely to increase sample size.

Prevalence shift changes predictive values even if ranking is stable. A response model developed in a selected clinical-trial cohort may overestimate benefit in a later-line real-world population. Recalibration can correct an intercept or slope when relationships are otherwise transportable, but it cannot repair missing predictors or reversed associations. Studies should distinguish recalibration from redevelopment and should evaluate whether the original threshold retains net benefit.

4.11. Performance metrics and clinical interpretation

Binary response AUC is common but incomplete. It measures ranking and is unaffected by systematic over- or underestimation of risk. Calibration-in-the-large, calibration slope, and plots across clinically relevant ranges are required when a probability will guide care. Sensitivity, specificity, positive predictive value, and negative predictive value depend on threshold and outcome prevalence and should be reported for a defined use.

Survival models require concordance or time-dependent AUC together with calibration at prespecified horizons. Censoring, subsequent therapy, and competing events should be handled transparently. For toxicity, cumulative incidence may be more appropriate than a binary label when death or treatment discontinuation prevents observation. Confidence intervals are essential; small apparent differences between models may not be statistically or clinically meaningful.

Decision-curve analysis estimates net benefit across thresholds under explicit assumptions, but it does not prove that an intervention is safe or effective. A threshold should correspond to a feasible action and should be selected without using the final test cohort. The decisive outcome remains whether model-assisted care improves patient benefit, reduces harm, or uses resources more effectively than standard multidisciplinary interpretation.

Uncertainty should be separated into statistical uncertainty in performance and uncertainty for an individual prediction. Confidence intervals

around AUC or C-index describe sampling variability; they do not tell a clinician whether one patient’s image is out of distribution or poorly segmented. Individual uncertainty can reflect ensemble variation, conformal prediction, input quality, or distance from development data, but each method requires calibration and a defined response. An uncertain output should lead to review, additional testing, or abstention rather than an unqualified recommendation.

Threshold selection should be based on the relative consequences of false-positive and false-negative actions. A low threshold may be appropriate for inexpensive confirmatory imaging, whereas treatment withholding or RT intensification requires stronger evidence and a higher burden of proof. Reporting the maximum AUC without the intended threshold leaves the clinical problem unspecified. Studies should provide confusion matrices and predictive values at prespecified thresholds in each external cohort.

Model comparison should use identical patients, outcomes, prediction times, and validation sets. Comparing a new network with a published AUC from another cohort is not evidence of superiority. Within-cohort comparisons should include a simple clinical model, conventional radiologic assessment, tumor burden or diameter, and, where available, validated biomarkers. The incremental model should be assessed for calibration and decision value, not only for a statistically significant difference in AUC.

Subgroup analysis should be hypothesis-driven and accompanied by uncertainty. Performance by sex, age, ancestry or race and ethnicity, cancer type, organ site, scanner, institution, and treatment regimen can reveal failure, but small subgroups produce unstable estimates. Apparent parity in AUC can coexist with calibration error if event prevalence differs. Fairness assessment should therefore examine access to required inputs, missingness, error distribution, calibration, and downstream action.

4.12. Imaging and EHR data for toxicity

Immune-related adverse events can limit treatment benefit, mimic progression, and prompt treatment interruption or hospitalization; they also differ in onset, organ distribution, diagnostic certainty, and management. Imaging is central to pneumonitis assessment but must be interpreted alongside symptoms, laboratory results, medication exposure, infection testing, radiation dose, and disease progression. In more than 2,200 ICI-treated patients, EHR-derived models reported AUCs of 0.739 for pneumonitis, 0.729 for hepatitis, 0.755 for colitis, and 0.752 for 1-year survival.^[40]

For clinical use, toxicity prediction must specify organ, grade, time horizon, calibration, and intervention. A high-risk output might trigger earlier imaging, laboratory monitoring, pulmonary review, or a treatment-planning change. A model that produces a risk label without a feasible pathway remains descriptive.

Toxicity labels are often less reliable than survival outcomes. Pneumonitis can overlap with infection, tumor progression, edema, aspiration, pre-existing interstitial disease, and radiation injury; hepatitis and colitis may be identified through combinations of laboratory, imaging, pathology, medication, and narrative documentation. Administrative codes alone can miss events or encode rule-out diagnoses. Adjudication should specify causality, grade, onset, competing diagnoses, and treatment, ideally with blinded multidisciplinary review.

Prediction time also changes the use case. A pretreatment model can support baseline counseling and monitoring intensity; an on-treatment model can incorporate symptoms and early imaging but may be too late to prevent the event; and a diagnostic model distinguishes toxicity from competing conditions after abnormalities appear. Mixing these tasks can produce an apparently strong model with no clear action. Reporting should identify what information is available at the exact prediction time and how much actionable lead time remains.

For thoracic treatment, radiation dose maps, baseline lung appearance, ICI exposure, infection risk, and symptoms should be integrated without assuming that all inflammatory imaging is immune-mediated. A calibrated model might identify patients for pulmonary consultation, closer oxygen monitoring, or a more conservative dose plan, but these actions require evaluation against their costs and the risk of undertreating cancer.

4.13. What current performance establishes

Current imaging AI methods can support noninvasive phenotyping, risk stratification, lesion-level assessment, and longitudinal monitoring. However, they do not, by themselves, establish autonomous treatment selection. The AUC measures discrimination and ranking performance, but does not capture individual-level calibration, clinical utility, or net benefit. Similarly, external validation varies widely, ranging from evaluation in a related institutional cohort to geographic, temporal, platform-level, and clinical-trial validation; these forms of validation should not be treated as equivalent. Saliency maps

and feature-importance scores may aid interpretation, but they should not be conflated with biological mechanisms.

4.14. From prognosis to treatment-effect estimation

Distinguishing prognosis from treatment effect is central to precision immunotherapy. A prognostic model estimates outcomes under the treatment patterns represented in the training data. Such a model may accurately identify patients with low tumor burden, favorable performance status, or indolent disease, but it does not establish whether ICI is preferable to chemotherapy, targeted therapy, supportive care, or combination regimens. Treatment-effect prediction, in contrast, asks how outcomes for a clinically comparable patient would differ under alternative therapeutic strategies. This counterfactual contrast cannot be inferred from a single treatment arm alone.

Randomized trials provide the strongest basis for estimating heterogeneous treatment effects because treatment assignment is independent of measured and unmeasured baseline prognosis within the trial. Imaging biomarkers can be evaluated through prespecified treatment-by-biomarker interactions, multivariable models with appropriate shrinkage, and estimates of absolute benefit across baseline risk. Subgroup response rates alone are insufficient: a marker can be associated with good outcome in both arms and still lack predictive value. Conversely, a modest interaction can be clinically meaningful when baseline risk and treatment toxicity differ substantially.

The effective sample size for treatment-effect modeling is constrained by the smaller treatment group and by the number of outcome events, not by the total number of image patches or lesions. Flexible deep models can overfit interactions even when conventional prognostic performance appears stable. Dimensionality reduction, prespecified imaging representations, penalization, and independent trial validation are therefore especially important. Exploration and confirmation should be separated; a treatment interaction discovered and tested in the same small cohort is hypothesis-generating.

Observational data can extend trial evidence but require an explicit causal target. Eligibility, treatment strategies, time zero, follow-up, outcome, and estimand should be defined as if emulating a target trial. Confounding by indication is substantial because therapy choice reflects PD-L1, molecular alterations, symptoms, comorbidity, access, clinician preference, and calendar time. Propensity methods or outcome regression cannot recover treatment effects when important determinants are absent or measured after treatment

initiation. Sensitivity analyses should examine residual confounding and positivity violations.

Absolute effects are more useful than a binary “benefit” label. A model might estimate the 12-month risk under ICI and under an alternative, their absolute difference, and uncertainty. The same relative effect can imply very different value in patients with different baseline risks. Decisions also require competing outcomes, including severe immune toxicity, early progression, quality of life, and treatment burden. A clinically credible system should present the expected benefit–harm profile rather than a single response probability.

Lesion-level treatment introduces an additional causal layer. Selecting a resistant lesion for biopsy or local RT may change local control without changing the systemic effect of ICI. Training on lesions that were actually irradiated creates selection bias because treatment was influenced by symptoms, size, accessibility, and clinician judgment. Future studies should record the decision process for every candidate lesion and compare explicit local-treatment strategies. Patient-level outcomes must account for the fact that untreated lesions and systemic therapy continue to determine prognosis.

Validation of a treatment-effect model should assess calibration of absolute outcomes within each treatment, calibration of predicted benefit, consistency across clinically important risk groups, and decision value under prespecified treatment thresholds. A model can rank relative benefit while systematically exaggerating its magnitude. Ultimately, a prospective study must test a locked rule in which model output changes a defined decision. Until such evidence exists, the appropriate terminology is exploratory treatment-effect modeling rather than model-guided treatment selection.

5. Radiotherapy plus ICI as a Focused Imaging Use Case

Radiotherapy (RT) plus immune checkpoint inhibition (ICI) serves here as a focused imaging use case rather than a second, independent theme of the review. The setting is particularly informative for imaging-centered AI because treatment is defined spatially and temporally: the relevant questions include which lesion to irradiate, what volume and dose distribution to use, how to fractionate treatment, and when to deliver RT relative to ICI. Imaging can represent whole-body disease and lesion heterogeneity, while RT planning data provide three-dimensional exposure maps for tumors, normal

organs, marrow, lymphoid structures, and surrogate compartments for circulating blood. Clinical evidence is nevertheless substantially more mature for predefined RT+ICI regimens than for model-guided selection among those regimens.

5.1. Context-dependent biological effects

RT can increase tumor-antigen availability, alter the peptide repertoire and major histocompatibility complex class I expression, and activate innate immune signaling that supports dendritic-cell priming and T-cell responses.^[41,42,43] These mechanisms provide a biological rationale for combining RT with ICI, but they do not imply that every radiation exposure is immunologically favorable. The observed effect depends on how tumor killing, inflammatory signaling, immune-cell preservation, stromal remodeling, and normal-tissue injury interact in a particular patient.

Dose and fractionation are therefore mechanistic variables rather than interchangeable descriptions of local treatment. In preclinical models, fractionated RT combined with checkpoint blockade produced systemic antitumor effects that were not reproduced by the tested single-dose regimen.^[44] Conversely, sufficiently high doses per fraction can induce the DNA exonuclease TREX1, reduce cytosolic DNA accumulation, and attenuate cGAS–STING-dependent type-I interferon signaling.^[45] These findings do not define a universally optimal “immunogenic dose.” They instead show that local cytotoxicity and immune stimulation need not increase in parallel and that the preferred regimen may vary with tumor type, lesion size, immune context, and therapeutic objective.^[46]

Treatment sequence adds another layer of dependence. Radiation-induced antigen presentation and inflammatory signaling evolve over time, while adaptive inhibitory pathways can also be induced. In murine models, concurrent PD-L1 blockade overcame adaptive resistance to fractionated RT, supporting the importance of temporal coordination.^[47] However, a preclinical scheduling interaction cannot determine the best clinical sequence. In practice, RT before ICI, concurrent RT+ICI, and consolidation ICI after chemoradiotherapy expose different patient populations and normal tissues, and they should be treated as distinct strategies rather than pooled under a single combination label.

Target selection and exposure of immune compartments may be as important as prescription dose. Irradiating a tumor can release antigen, whereas

elective irradiation of tumor-draining lymph nodes may impair antigen presentation and T-cell priming; in preclinical work, elective nodal irradiation attenuated the efficacy of stereotactic RT combined with immunotherapy.^[48] Larger fields can also expose marrow and circulating lymphocytes. In patients receiving PD-1 inhibitors, prior or concurrent RT-related lymphocyte depletion has been associated with poorer outcomes, and retrospective analyses in stage III non-small-cell lung cancer (NSCLC) have linked higher estimated dose to the host immune system with inferior tumor control and survival.^[49,50] These associations are clinically important but not causal proof that an immune-sparing plan will improve the benefit of ICI.

The net RT–ICI interaction is consequently lesion-, site-, dose-, fractionation-, sequence-, and host-dependent. A useful imaging representation should extend beyond the gross tumor volume and nominal prescription: it should preserve lesion identity, spatial dose, temporal sequence, normal-organ exposure, relevant lymphoid or marrow structures, and baseline pulmonary and immune reserve. Figure 5 summarizes this balance and its connection to individualized decision support.

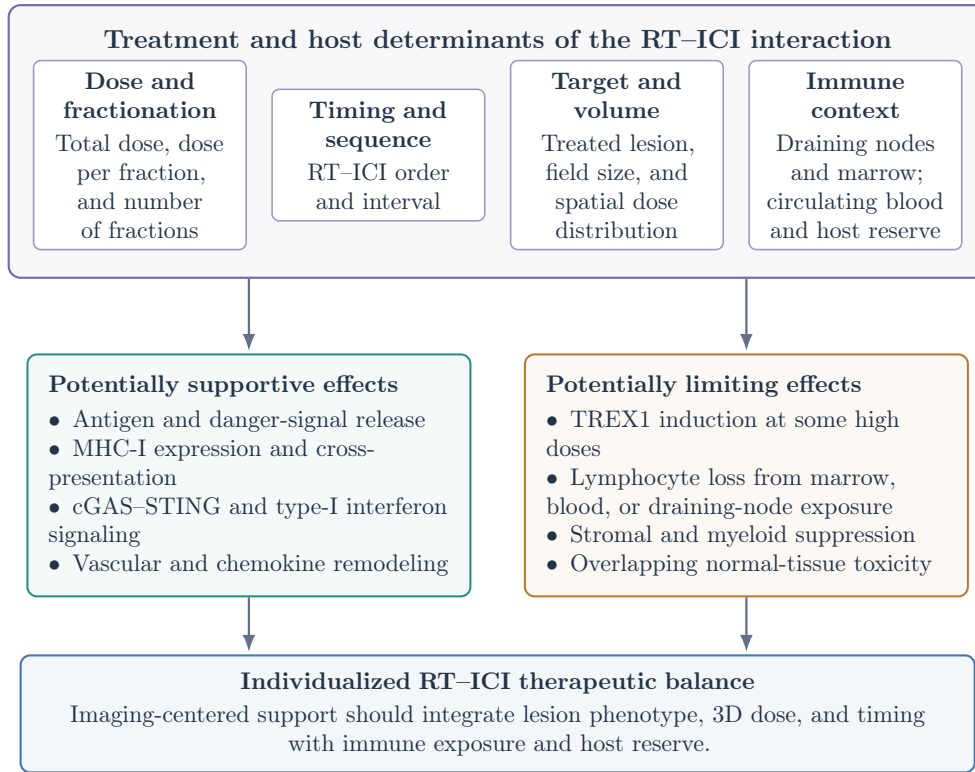


Figure 5: Context-dependent determinants of the interaction between RT and ICI. Dose and fractionation, treatment timing, target selection and volume, and exposure of lymphoid, marrow, and circulating immune compartments can shift radiotherapy toward immune-supportive or immune-limiting effects. Their patient-specific balance motivates imaging-centered tools for individualized treatment design.

5.2. Clinical evidence and the resulting opportunity for AI

The strongest clinical foundation is the PACIFIC paradigm in unresectable stage III NSCLC. Among patients without progression after concurrent chemoradiotherapy, consolidation durvalumab improved progression-free and overall survival relative to placebo.^[51,52] At five years, overall survival was 42.9% versus 33.4%, and progression-free survival was 33.1% versus 19.0%.^[53] This is practice-changing evidence for a defined sequential strategy. It does not establish that RT should be added to ICI in metastatic disease, that concurrent ICI is superior, or that one radiation dose, field, or fractionation is optimal for immune synergy.

In metastatic NSCLC, a pooled analysis of the randomized PEMBRO-RT and MD Anderson phase I/II trials included 148 patients and reported higher out-of-field response and longer survival with pembrolizumab plus RT than with pembrolizumab alone.^[54] The result supports further study but does not resolve which lesion should be treated or which dose and schedule should be used. KEYNOTE-799 subsequently demonstrated the feasibility of pembrolizumab with concurrent chemoradiotherapy in unresectable stage III NSCLC, with objective response rates of approximately 70% and grade 3 or higher pneumonitis in 8.0% and 6.9% of its two nonrandomized cohorts.^[55] More recent phase III evidence is cautionary: PACIFIC-2 and CheckMate 73L did not establish superiority of concurrent intensification over the consolidation durvalumab paradigm.^[56,57]

These studies define clinically relevant settings and endpoints, but they are not evidence of model-guided treatment. Trial-level averages do not identify which patient should receive additional RT, which lesion should be selected, or whether an alternative plan or sequence would improve that patient's outcome. They nevertheless reveal where decision-support research could be useful. Baseline and serial imaging can quantify total burden, organ-specific disease, mixed response, and pulmonary reserve; RT data can provide lesion-resolved dose and immune-compartment exposure; and longitudinal clinical data can align these features with treatment timing and toxicity. To support a treatment choice rather than prognosis alone, an AI study must include clinically relevant alternatives and estimate how outcomes differ between them.

Table 5: Selected clinical evidence for radiotherapy combined or sequenced with ICI.

Study; setting	Design and regimen	Key result	Interpretation for decision support
PACIFIC Antonia et al. ^[51,52] ; Spigel et al. ^[53] Unresectable stage III NSCLC	Randomized durvalumab versus placebo after concurrent chemoradiotherapy	Five-year OS 42.9% versus 33.4%; PFS 33.1% versus 19.0%	Establishes consolidation durvalumab after chemoradiotherapy, not AI-guided selection or metastatic RT addition.
Pooled phase II NSCLC Theelen et al. ^[54] Metastatic disease	Randomized pembrolizumab with or without RT; pooled $n = 148$	Higher out-of-field response and longer survival reported with RT	Hypothesis- generating; small cohort and unresolved lesion, dose, and schedule selection.
KEYNOTE-799 Jabbour et al. ^[55] Stage III NSCLC	Nonrandomized concurrent pembrolizumab plus chemoradiotherapy	Objective response about 70%; grade ≥ 3 pneumonitis 8.0% and 6.9%	Demonstrated feasibility but lacked a randomized comparator for concurrent intensification.
PACIFIC-2 Bradley et al. ^[56] Stage III NSCLC	Concurrent durvalumab with chemoradiotherapy versus placebo during chemoradiotherapy	Concurrent strategy did not establish superiority over the consolidation paradigm	Underscores that sequence cannot be inferred from biological rationale alone.
CheckMate 73L Peters et al. ^[57] Stage III NSCLC	Nivolumab-based concurrent and consolidation strategies versus durvalumab consolidation	Intensification did not improve the primary outcome	Supports decision-specific randomized testing before adopting more complex schedules.

Abbreviations: AI, artificial intelligence; ICI, immune-checkpoint inhibitor; NSCLC, non-small-cell lung cancer; OS, overall survival; PFS, progression-free survival; RT, radiotherapy.

5.3. AI for efficacy and response prediction

Direct patient-level evidence for AI-based efficacy prediction in combined RT+ICI remains sparse. Li et al. studied 88 patients with unresectable stage III NSCLC treated with concurrent chemoradiotherapy followed by PD-1 inhibition.^[58] Their pretreatment CT-radiomics model reported an AUC of 0.85 in an independent retrospective test subset of 33 patients and 0.90 in a prospectively accrued validation subset of 17 patients. The study provides a relevant example of imaging-based response stratification, but the small validation partitions yield imprecise performance estimates. More importantly, all patients received a predefined combined-modality strategy; the model therefore predicts response conditional on that strategy and does not estimate whether an alternative RT or ICI approach would be preferable.

Mechanistic AI offers a complementary but earlier line of work. Peng et al. developed an AI-based mathematical framework to model interactions between pulsed adaptive RT and immunotherapy.^[59] Such models can formalize hypotheses about dose and timing and may help prioritize experimental schedules. They are not substitutes for patient-level imaging validation, because simulated synergy does not demonstrate transportable prediction or clinical utility. The distinction between mechanistic simulation, post-treatment outcome prediction, and treatment-effect estimation should remain explicit.

Future efficacy studies should prespecify both the endpoint and the decision. Local control of an irradiated lesion, response of nonirradiated lesions, progression-free survival, and overall survival represent different biological and clinical targets. A model intended to decide whether to add RT requires an ICI-alone comparator; a sequencing model requires alternative timing strategies; and a lesion-selection model requires outcomes for all candidate lesions, including those not selected for treatment. Without these contrasts, even strong discrimination may reflect baseline prognosis rather than differential benefit from RT.

5.4. AI for toxicity and safety prediction

Building on the general ICI-toxicity considerations in Section 4.12, this subsection focuses on the additional planning and diagnostic complexity introduced by thoracic RT+ICI. Here, an adverse-event risk estimate is useful only if it is linked to a diagnostic, surveillance, planning, or mitigation pathway, because radiation pneumonitis, immune-mediated pneumonitis, in-

fection, tumor progression, and mixed etiologies can overlap radiologically and clinically.

Yang et al. developed a multimodal deep-learning model in 261 patients with lung cancer treated with RT and immunotherapy, integrating pretreatment CT radiomics, deep image features, and clinical variables; the reported five-fold cross-validated AUC for symptomatic pneumonitis was 0.922.^[60] Kraus et al. combined radiomics, dosimetrics, dose-volume histograms, and clinical variables in 100 patients receiving thoracic RT, 39 of whom received ICI, and reported a best combined-model AUC of 0.77.^[61] Wang et al. studied 102 patients with NSCLC treated with ICI and RT and reported AUCs of 0.918 in training and 0.833 in an internal validation cohort for radiation esophagitis.^[62] Masuda et al. evaluated grade 2 or higher pneumonitis after chemoradiotherapy and durvalumab in 123 patients from 10 institutions, reporting an AUC of 0.88 under five-fold cross-validation.^[63]

These studies show that imaging, dosimetry, and clinical data can be combined to estimate toxicity risk, but their evidence remains predominantly retrospective and internally evaluated. Cross-validation within a multicenter cohort is not equivalent to a locked external test across new institutions, scanners, planning practices, and calendar time. Moreover, a risk score becomes decision support only when it is connected prospectively to a defined action, such as plan modification, immune-compartment sparing, intensified surveillance, diagnostic work-up, or treatment interruption, and when the consequences of that action are evaluated.

The evidence gap is summarized in Figure 6. Current models estimate efficacy or toxicity after treatment selection. None has prospectively shown that an AI-selected choice of whether, where, when, or how to deliver RT during ICI improves a patient-relevant outcome or safety endpoint.

Table 6: Selected imaging-centered AI studies in patients treated with RT and ICI, separated by efficacy and toxicity endpoints.

Study; treated population	Endpoint and model inputs	Reported performance	Clinical-decision interpretation
Li et al. ^[58] ; stage III NSCLC, $n = 88$	Efficacy: response after chemoradiotherapy followed by PD-1 inhibition; pretreatment CT radiomics	AUC 0.90 in a prospectively accrued validation subset ($n = 17$); 0.85 in an independent retrospective test subset ($n = 33$)	Predicts response after a fixed regimen in small validation partitions; no AI-directed RT or ICI choice was evaluated.
Yang et al. ^[60] ; lung cancer, $n = 261$	Toxicity: symptomatic pneumonitis; multimodal deep network using CT radiomics, deep imaging features, and clinical factors	Five-fold cross-validated AUC 0.922	Strong cross-validated discrimination, but no locked independent test or prospective model-guided mitigation pathway.
Kraus et al. ^[61] ; thoracic RT, $n = 100$, including 39 ICI recipients	Toxicity: pneumonitis; radiomics, dosiomics, dose-volume histograms, and clinical factors	Best combined-model AUC 0.77	Small heterogeneous cohort with limited ICI representation; RT+ICI inference and clinical actionability remain exploratory.
Wang et al. ^[62] ; NSCLC, $n = 102$	Toxicity: radiation esophagitis; CT radiomics plus clinical and dosimetric factors	AUC 0.918 in training and 0.833 in internal validation	Single-center internal validation; predicts toxicity rather than whether or how RT should be delivered during ICI.
Masuda et al. ^[63] ; stage III NSCLC, $n = 123$ from 10 institutions	Toxicity: grade ≥ 2 pneumonitis after chemoradiotherapy and durvalumab; pretreatment CT radiomics and clinical variables	AUC 0.88 under five-fold cross-validation	Multicenter sampling, but no locked external test or comparison of alternative RT or toxicity-mitigation strategies.

Interpretive note: Within this selected evidence set, AI estimates efficacy or toxicity after treatment selection; prospective utility for choosing whether, where, when, or how to deliver RT during ICI remains unestablished.

Abbreviations: AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; CT, computed tomography; ICI, immune-checkpoint inhibitor; NSCLC, non-small-cell lung cancer; PD-1, programmed cell death protein 1; RT, radiotherapy.

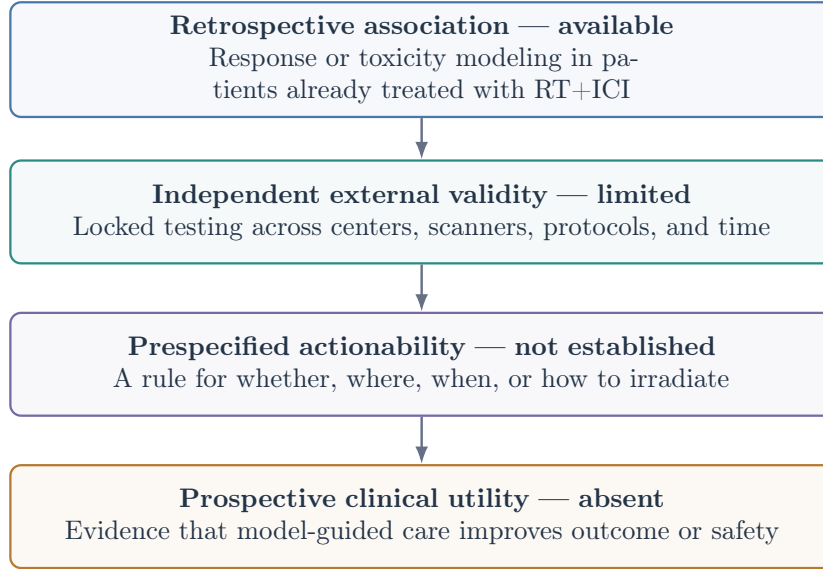


Figure 6: Evidence maturity for AI in immunoradiotherapy. Existing studies estimate efficacy or toxicity after treatment selection, but the prospective clinical utility of AI-selected RT remains unproven. ^[58,60,61,62,63]

5.5. Decision-specific research targets

Future models should be organized around one prespecified clinical decision at a time. A model addressing whether to add RT must compare an RT-containing strategy with a clinically relevant non-RT alternative and estimate treatment-effect heterogeneity. A model addressing which lesion to irradiate requires organ-aware, lesion-level imaging linked to an explicit objective, such as durable local control, symptom prevention, or management of oligoprogression. A model addressing dose, fractionation, or sequence must encode three-dimensional dose distributions and treatment timing together with exposure of lung, marrow, lymphoid structures, and circulating blood. A toxicity model must report calibrated absolute risk and specify the action triggered at a clinically justified threshold.

Among plausible near-term targets, lesion selection for oligoprogression and mitigation of thoracic toxicity may be particularly testable because each can be linked to a discrete imaging-based action and a patient-relevant endpoint. Even in these settings, retrospective association should be treated as hypothesis generation. Translation requires a locked model, an explicit

override pathway, external validation, and prospective comparison of model-guided care with the existing clinical workflow.

6. Methodological and Translational Requirements

6.1. *Standardization and leakage control*

Imaging varies by scanner, protocol, contrast, reconstruction, slice thickness, motion, and segmentation. Pathology, molecular assays, laboratory values, and RT plans introduce additional technical variation. A model can learn these signatures more easily than stable tumor biology. The Image Biomarker Standardisation Initiative provides standardized radiomic definitions^[18]; CLAIM and joint ESTRO/AAPM guidance support transparent imaging and radiation-oncology AI reporting.^[64,65]

All lesions, slides, patches, and time points from one patient must remain in one partition. Preprocessing, harmonization, imputation, feature selection, and hyperparameter tuning must be learned within training data. An untouched test set should be evaluated only after the pipeline is locked. Cross-validation quantifies internal uncertainty but does not create external evidence.

6.2. *Validation, calibration, and missing modalities*

Random internal splits test interpolation within one data-generating process. Temporal validation tests changes in practice; geographic validation tests institutional and population shift; platform validation tests scanners and assays; trial validation tests a protocol-selected population. Performance should include discrimination, calibration intercept and slope, clinically relevant operating points, subgroup error, and failure or abstention rates.

Multimodal models must specify which inputs are mandatory and how predictions change when a biopsy, sequencing assay, or scheduled scan is unavailable. Complete-case analysis can select an atypically well-characterized population. Modality dropout, late fusion, and prespecified fallback models may better reflect routine care.

TRIPOD+AI and PROBAST+AI support reporting and risk-of-bias assessment for clinical prediction models.^[66,67] These frameworks should influence protocol design rather than be applied only after analysis.

6.3. Reproducibility, multicenter collaboration, and data governance

Reproducibility has at least three levels. Computational reproducibility asks whether the same code and data produce the same result. Analytical reproducibility asks whether plausible changes in acquisition, preprocessing, contouring, and feature extraction preserve the estimate. Clinical reproducibility asks whether the locked system performs in a new population and workflow. Reporting only the final model coefficients or network architecture addresses none of these levels adequately.

Radiomics quality frameworks emphasize protocol description, feature robustness, validation, biological or clinical association, comparison with accepted standards, and open science.^[68] For contemporary deep learning, an equivalent record should include series-selection rules, preprocessing versions, segmentation provenance, pretrained weights, augmentation, optimization, stopping criteria, model selection, and random seeds. The purpose is not procedural detail for its own sake; it is to identify which parts of the pipeline are fixed and which may vary across sites.

Data and code sharing can accelerate verification, but unrestricted release is not always feasible for clinical images. Imaging contains facial, anatomical, temporal, and institutional information that may remain identifiable after removal of conventional metadata. RT plans, genomic profiles, rare diagnoses, and longitudinal dates increase re-identification risk. Governance should therefore specify consent or waiver, de-identification, date handling, access control, permitted use, retention, audit, and procedures for withdrawal or correction.

When raw data cannot be shared, reproducibility can be improved through locked containers, executable preprocessing specifications, synthetic test cases, feature dictionaries, model cards, and controlled evaluation servers. A model can be submitted to an external data custodian and evaluated without transferring patient data to the developer. This arrangement preserves a stronger separation between development and testing than a nominal external cohort that is repeatedly queried during model refinement.

Federated learning is attractive for rare cancers and multimodal immunotherapy cohorts because parameters rather than raw data are exchanged. It does not automatically solve privacy, heterogeneity, or bias. Site-specific label definitions, scanner protocols, missing modalities, and outcome prevalence remain, and parameter updates can leak information under some threat models. Federated studies should define the participating-site distribution, aggregation strategy, privacy safeguards, communication

failures, site-level performance, and behavior when a small or atypical center joins or leaves.

Multicenter evaluation should preserve heterogeneity rather than harmonize every site into the development distribution. Site-wise results reveal whether apparent generalization depends on one dominant center. Leave-one-site-out evaluation can be informative during development, but the final external site must remain untouched. If recalibration is permitted, its sample size, variables, and timing should be prespecified, and performance should be reported both before and after recalibration.

Dataset stewardship continues after publication. Corrections to imaging, outcomes, and biomarker values should be versioned; duplicate patients across public and institutional sources should be detected; and train-test overlap should be reassessed when foundation models or external pretrained weights are introduced. Provenance should link every prediction to the model version, input series, preprocessing version, and timestamp. Without this traceability, later error analysis and clinical accountability are compromised.

6.4. Interpretability, fairness, and clinical workflow

Clinicians need the intended population, prediction time, input availability, reference standard, calibrated risk, uncertainty, and known failure conditions. Feature importance and saliency can support review but do not prove causality. Population representation matters because access to imaging, molecular testing, trials, and toxicity assessment differs across groups and institutions.

The FUTURE-AI framework emphasizes fairness, universality, traceability, usability, robustness, and explainability.^[69] WHO guidance emphasizes autonomy, safety, transparency, responsibility, inclusion, and sustainability.^[70] High-risk decision support additionally requires version control, drift detection, recalibration, incident review, and human override.

Table 7: Minimum safeguards for translating imaging and multimodal AI into immunotherapy care.

Domain	Required practice	Failure detected	Evidence to report
Population and time point	Define eligibility, treatment line, prediction time, and available inputs	Target-population drift and use of future information	Cohort flow, exclusions, event count, and prediction timeline.
Patient-level separation	Keep all lesions, images, slides, and time points from one patient together	Patch-, lesion-, or visit-level leakage	Split unit, site allocation, and code or audit of identifiers.
Locked preprocessing	Fit harmonization, imputation, feature selection, and thresholds within training data	Information transfer from validation or test data	Versioned pipeline and untouched-test evaluation.
External validity	Test geographic, temporal, platform, and clinically relevant subgroup shift	Institution-, scanner-, assay-, or prevalence-specific shortcuts	Discrimination, calibration, operating points, uncertainty, and failure rate.
Missing modalities	Define mandatory inputs, dropout training, and fallback models	Complete-case selection and silent failure in routine care	Missingness reasons, comparison of included and excluded cases, and dropout performance.
Action and oversight	Prespecify threshold, recommendation, override, monitoring, and drift response	Non-actionable scores, automation bias, and unsafe deployment	Decision-curve analysis, silent deployment, override reasons, and prospective utility.

6.5. *Prospective clinical utility*

A defensible development sequence is retrospective feasibility, locked external validation, prospective silent evaluation, model-assisted feasibility, and comparative testing of model-guided care. SPIRIT-AI and CONSORT-AI guide protocols and reports when AI is part of the intervention.^[71,72] The comparator should be the actual multidisciplinary workflow rather than an artificially weak baseline.

Clinical utility depends on the action. A response model may trigger confirmatory imaging or biopsy; a toxicity model may trigger monitoring or pulmonary review; a treatment-effect model may alter therapy. Each action has costs and harms. Survival, severe toxicity, quality of life, time on ineffective treatment, resource use, and patient burden are more meaningful than AUC alone.

Prospective silent evaluation is an important intermediate stage. A locked model can run without influencing care while investigators measure data availability, processing time, calibration, subgroup failures, out-of-distribution cases, and disagreement with multidisciplinary assessment. This stage often exposes workflow problems that cannot be identified retrospectively. Only after the output, threshold, and override process are stable should the model influence care.

6.6. *Human–AI team evaluation and early clinical studies*

The unit of clinical evaluation is the human–AI team, not the algorithm in isolation. An imaging model may improve average reader accuracy yet worsen decisions when its confidence is poorly communicated or when users anchor on an incorrect output. Conversely, a moderately discriminative model may be useful if it reliably identifies cases requiring second review. Evaluation should therefore measure unaided clinicians, AI alone, and clinicians assisted by AI on the same representative cases, while controlling for reading order and learning effects. Human-factor outcomes should include alert burden, processing delay, recommendation viewing and adherence, reasons for override, additional testing, and changes in clinician time or resource use.

The interface should show the intended output, prediction time, threshold, uncertainty, input quality, and relevant comparison without implying mechanistic certainty. Lesion overlays can support verification, but saliency maps should not substitute for evidence. For whole-body models, the clinician should be able to identify which examinations and lesions were included,

which were excluded, and whether the system used a fallback because pathology, sequencing, or follow-up imaging was missing.

Early live evaluation should examine feasibility and human factors before efficacy claims. DECIDE-AI provides stage-specific reporting guidance for early clinical evaluation of AI decision-support systems.^[73] Relevant measures include recruitment, data availability, technical failures, processing time, user training, adherence, overrides, workflow change, and unintended consequences. A technically successful prediction delivered after the tumor board or after treatment initiation has no practical utility.

Comparative clinical trials should randomize at the patient, clinician, tumor-board, or site level according to the risk of contamination. Cluster designs may be necessary when clinicians cannot reasonably alternate between assisted and unassisted workflows. Endpoints should match the decision: time to confirmed response, severe toxicity, time to systemic-therapy change, quality of life, or survival may be appropriate, whereas reader agreement alone is not sufficient. Protocols should specify downtime procedures and the authority of human override.

6.7. Lifecycle monitoring, model updating, and regulatory evidence

Deployment changes the data-generating process. Scanner replacements, reconstruction updates, new ICI combinations, changing PD-L1 testing, revised response criteria, and altered referral patterns can shift inputs and outcomes. Monitoring should distinguish input drift, prevalence shift, calibration drift, and changes in clinical action. Drift alerts require prespecified thresholds and ownership; otherwise, dashboards accumulate signals without a safe response.

Performance monitoring requires outcomes that become available at different times. Image quality and missing inputs can be assessed immediately, response after weeks or months, and survival much later. A practical surveillance plan should combine leading indicators with delayed clinical endpoints. It should also stratify by cancer, site, regimen, scanner, and relevant patient groups so that acceptable overall performance does not conceal a localized failure.

Updating is a regulated intervention, not routine software maintenance. The International Medical Device Regulators Forum’s Good Machine Learning Practice principles emphasize representative data, independent testing, human–AI performance, clear user information, monitoring, and control across the total product life cycle.^[74] Each update should state

what changed, why it changed, which data were used, whether the decision threshold changed, and what evidence supports continued safety and effectiveness.

Locked periodic updates are easier to audit than continuous adaptation. A candidate update can be developed on new data, evaluated in a segregated test set, compared with the deployed version, and released only after approval. Continuous learning may be appropriate in limited settings but requires stronger controls against feedback loops. If clinicians preferentially act on high-risk predictions, subsequent outcomes are partly caused by the model, and naive retraining can learn from its own interventions.

Rollback, incident response, and communication should be designed before deployment. The institution needs a mechanism to suspend the model, identify affected predictions, notify users, preserve logs, and revert to a validated version. Material changes in calibration or workflow may require revalidation even when the underlying network is unchanged. Local responsibility should be explicit among the clinical service, imaging department, information technology team, developer, and governance body.

6.8. Implementation and health-economic value

Implementation value depends on more than model accuracy. Direct costs include software, computing, storage, integration, quality control, cybersecurity, maintenance, and staff training. Indirect costs include additional imaging, biopsies, consultations, false alarms, and time spent resolving failed segmentations. These costs should be compared with potential reductions in ineffective treatment, severe toxicity, delayed progression recognition, and manual measurement workload.

Equity is also an implementation outcome. A model that requires specialized PET, broad sequencing, or research-quality follow-up may perform well while widening access disparities. Reporting should identify the proportion of eligible patients who can supply the required inputs and should evaluate a lower-cost fallback pathway. The strongest deployment candidate may be a parsimonious imaging-and-clinical model with slightly lower discrimination but broader availability, faster turnaround, and more stable calibration.

Adoption should proceed only when the system adds value to the real comparator. If an experienced multidisciplinary team already reaches the same decisions, automation may still reduce time or variability, but that benefit must be measured. If the model changes care, the resulting benefits

and harms must be evaluated prospectively. This requirement keeps implementation focused on patient and service outcomes rather than on technical novelty.

7. Research Priorities

The highest-priority studies should use imaging as a longitudinal spatial backbone and add other modalities only when they improve a defined decision. Multicenter cohorts should preserve raw images, lesion identity, contours, dose maps, pathology, molecular assays, ctDNA, medications, laboratory trajectories, outcomes, and reasons for missingness. Shared definitions are more important than unrestricted feature count.

Treatment-specific prediction requires relevant controls. Where randomized trial data exist, models should estimate absolute and relative treatment effects with calibration across baseline risk. Longitudinal models should identify prespecified decision times and account for informative observation. Prospective silent deployment should precede any model-triggered change in therapy.

For RT+ICI, a realistic trial could evaluate local RT for oligoprogression during otherwise effective ICI. A locked model would combine lesion-level serial imaging, organ site, previous response, ctDNA trajectory, RT feasibility, and pulmonary or marrow risk. The intervention would be a prespecified recommendation, not an unstructured score. Endpoints could include time to systemic-therapy change, local and nonirradiated control, severe toxicity, quality of life, and survival.

Table 8: Decision-focused research priorities for imaging AI in cancer immunotherapy.

Clinical question	Minimum data structure	Appropriate evaluation	Patient-relevant endpoint
Who benefits from ICI?	Imaging and routine biomarkers in ICI and relevant non-ICI groups	Treatment-effect modeling with cancer-specific calibration	Absolute benefit, severe toxicity, quality of life, and survival.
Is early change reliable?	Prespecified serial imaging and molecular time points with reasons for missing scans	Landmark analysis, checks for informative observation, and external temporal validation	Time to effective treatment change and avoidance of premature discontinuation.
Which lesion should be treated locally?	Whole-body lesion identity, organ site, serial response, symptoms, and RT feasibility	Lesion- and patient-level evaluation with prespecified aggregation	Local control, time to systemic-therapy change, symptoms, and toxicity.
How should RT and ICI be sequenced?	Three-dimensional dose, fractionation, timing, immune exposure, and a valid comparator	Randomized or target trial-compatible treatment-effect analysis	Survival, non-irradiated control, immune toxicity, and normal-tissue injury.
Can AI improve workflow?	Locked software, real-time data availability, uncertainty, and override logging	Prospective silent deployment followed by comparative model-guided evaluation	Safety, calibration, clinician adherence, resource use, and net patient benefit.

Abbreviations: AI, artificial intelligence; ICI, immune-checkpoint inhibitor; RT, radiotherapy.

8. Conclusion

Imaging-centered AI provides whole-body, longitudinal information that cannot be obtained from a single biopsy or blood draw and is already embedded in oncology care. Current studies provide evidence for radiogenomic phenotyping, lesion-level analysis, risk stratification, serial response assessment, toxicity modeling, and multimodal biomarker integration. Pathology, omics, ctDNA, laboratory measurements, and EHR data add complementary specificity, but increasingly complex fusion should be justified by incremental calibration, robustness, and decision utility.

The principal conceptual boundary is between outcome prediction and treatment-effect prediction. Most current models stratify prognosis among ICI-treated patients. Few include relevant non-ICI controls, and none has established broad clinical utility through a model-guided treatment trial. RT+ICI illustrates this gap: direct AI studies predict response, pneumonitis, or esophagitis, but no prospective evidence shows that an AI-selected RT strategy improves outcome.

Progress should be judged by data provenance, patient-level leakage control, calibration, external validity, missing-modality robustness, treatment specificity, and improvement in a patient-relevant decision. Under these conditions, imaging AI may help move cancer immunotherapy from isolated static biomarkers toward multimodal, longitudinal, and individualized decision support. Its clinical value will ultimately depend not on a higher reported AUC but on whether it improves a defined decision for a defined patient at a defined time.

Statements and Declarations

Funding. This research received no external funding.

Competing interests. The authors declare no competing interests.

CRedit authorship contribution statement. Shengyang Zhong: Conceptualization, Methodology, Project administration, Writing – original draft, Writing – review & editing. Siyuan Li, Wanrong Yang, and Zheng Zheng: Writing – original draft, Writing – review & editing. Zhengliang Liu: Methodology, Writing – review & editing. Jin Xie, Jin Li, Bin Yi, Lin Zhao, Xiang “Shaun” Li, Wei Liu, Lian Zhang, and Tianming Liu: Writing – review & editing. Ning Qiang: Supervision, Project administration, Writing – review & editing. Shengyang Zhong and Siyuan Li contributed equally and share first authorship. All authors reviewed and approved the final manuscript.

Data availability. No new datasets were generated or analyzed in this study.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI’s ChatGPT to assist with language editing, structural organization, literature summarization, and L^AT_EX preparation. The authors independently verified the cited sources, scientific summaries, interpretations, and figure content; reviewed and edited all AI-assisted output as needed; and take full responsibility for the content of the article.

References

1. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med*. 2012;366(26):2443–2454. doi:10.1056/NEJMoa1200690.
2. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*. 2017;357(6349):409–413. doi:10.1126/science.aan6733.
3. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N Engl J Med*. 2016;375(19):1823–1833. doi:10.1056/NEJMoa1606774.
4. Gandhi L, Rodríguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, et al. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *N Engl J Med*. 2018;378(22):2078–2092. doi:10.1056/NEJMoa1801005.
5. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat Rev Cancer*. 2019;19(3):133–150. doi:10.1038/s41568-019-0116-x.
6. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science*. 2018;362(6411):eaar3593. doi:10.1126/science.aar3593.

7. Anagnostou V, Forde PM, White JR, Niknafs N, Hruban C, et al. Dynamics of tumor and immune responses during immune checkpoint blockade in non-small cell lung cancer. *Cancer Res.* 2019;79(6):1214–1225. doi:10.1158/0008-5472.CAN-18-1127.
8. Chen DS, Mellman I. Oncology meets immunology: The cancer-immunity cycle. *Immunity.* 2013;39(1):1–10. doi:10.1016/j.immuni.2013.07.012.
9. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. *Nature.* 2017;541(7637):321–330. doi:10.1038/nature21349.
10. Wherry EJ, Kurachi M. Molecular and cellular insights into t cell exhaustion. *Nat Rev Immunol.* 2015;15(8):486–499. doi:10.1038/nri3862.
11. Kalbasi A, Ribas A. Tumour-intrinsic resistance to immune checkpoint blockade. *Nat Rev Immunol.* 2020;20(1):25–39. doi:10.1038/s41577-019-0218-4.
12. Zaretsky JM, Garcia-Diaz A, Shin DS, Escuin-Ordinas H, Hugo W, et al. Mutations associated with acquired resistance to PD-1 blockade in melanoma. *N Engl J Med.* 2016;375(9):819–829. doi:10.1056/NEJMoa1604958.
13. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, et al. Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science.* 2015;348(6230):124–128. doi:10.1126/science.aaa1348.
14. Samstein RM, Lee CH, Shoushtari AN, Hellmann MD, Shen R, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet.* 2019;51(2):202–206. doi:10.1038/s41588-018-0312-8.
15. Marcus L, Fashoyin-Aje LA, Donoghue M, Yuan M, Rodriguez L, Gallagher PS, Philip R, Ghosh S, Theoret MR, Beaver JA, Pazdur R, Lemery SJ. FDA approval summary: Pembrolizumab for the treatment of tumor mutational burden-high solid tumors. *Clin Cancer Res.* 2021;27(17):4685–4689. doi:10.1158/1078-0432.CCR-21-0327.

16. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJM, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*. 2014;515(7528):568–571. doi:10.1038/nature13954.
17. Bartolomucci A, et al. Circulating tumor DNA to monitor treatment response in solid tumors and advance precision oncology. *NPJ Precis Oncol*. 2025;9:84. doi:10.1038/s41698-025-00876-y.
18. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328–338. doi:10.1148/radiol.2020191145.
19. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: An imaging biomarker, retrospective multicohort study. *Lancet Oncol*. 2018;19(9):1180–1191. doi:10.1016/S1470-2045(18)30413-3.
20. Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. *Ann Oncol*. 2019;30(6):998–1004. doi:10.1093/annonc/mdz108.
21. Saad MB, Hong L, Aminu M, Vokes NI, Chen P, Salehjahromi M, et al. Predicting benefit from immune checkpoint inhibitors in patients with non-small-cell lung cancer by CT-based ensemble deep learning: A retrospective study. *Lancet Digit Health*. 2023;5(7):e404–e420. doi:10.1016/S2589-7500(23)00082-1.
22. Sato Y, Morimoto T, Hara S, Nagata K, Hosoya K, Nakagawa A, Tachikawa R, Tomii K. Dissociated response and clinical benefit in patients treated with nivolumab monotherapy. *Invest New Drugs*. 2021;39(4):1170–1178. doi:10.1007/s10637-021-01077-7.
23. Guckenberger M, Lievens Y, Bouma AB, Collette L, Dekker A, deSouza NM, Dingemans AMC, Fournier B, Hurkmans C, Lecouvet FE, et al. Characterisation and classification of oligometastatic disease: A european society for radiotherapy and oncology and european organisation for research and treatment of cancer consensus recommendation. *Lancet Oncol*. 2020;21(1):e18–e28. doi:10.1016/S1470-2045(19)30718-1.

24. Lin P, Xie W, Li Y, Zhang C, Wu H, Wan H, Gao M, Liang F, Han P, Chen R, et al. Intratumoral and peritumoral radiomics of MRIs predicts pathologic complete response to neoadjuvant chemoimmunotherapy in patients with head and neck squamous cell carcinoma. *J Immunother Cancer*. 2024;12(11):e009616. doi:10.1136/jitc-2024-009616.
25. Sanvito F, Nocera G, Wang Z, Shiuan E, Sun L, Oshima S, Kim A, Kryukov I, Shao G, Cho NS, et al. Diffusion and contrast-enhancement MRI phenotypes associated with immune checkpoint inhibitor exposure and with immune cell infiltration in brain metastases. *Neuro Oncol*. 2025;27(8):2194–2205. doi:10.1093/neuonc/noaf084.
26. Lopci E, Hicks RJ, Dimitrakopoulou-Strauss A, Dercle L, Iravani A, Seban RD, et al. Joint EANM/SNMMI/ANZSNM practice guidelines/procedure standards on recommended use of [^{18}F]FDG PET/CT imaging during immunomodulatory treatments in patients with solid tumors version 1.0. *Eur J Nucl Med Mol Imaging*. 2022;49(7):2323–2341. doi:10.1007/s00259-022-05780-2.
27. Masse M, Chardin D, Tricarico P, Ferrari V, Martin N, Otto J, Darcourt J, Comte V, Humbert O. [^{18}F]FDG-PET/CT atypical response patterns to immunotherapy in non-small cell lung cancer patients: Long-term prognosis assessment and clinical management proposal. *Eur J Nucl Med Mol Imaging*. 2024;51(12):3696–3708. doi:10.1007/s00259-024-06794-8.
28. Postow MA, Zang PD, Pal SK, Mauguén A, Farwell M, Gordon M, Hays D, Wong JYC, Chen D, Ulaner GA, et al. CD8 PET imaging with [^{89}Zr]Zr-berdoxam-crefmirlimab in patients with solid tumors: The phase II icorrelate study. *J Immunother Cancer*. 2026;14:e012208. doi:10.1136/jitc-2025-012208.
29. Sako C, Kurland BF, Schmidt TG, He J, De Groot M, Patel AA, et al. Deep-learning serial CT prediction of survival in immunotherapy-treated non-small cell lung cancer. *JAMA Netw Open*. 2026;9(1):e2555759. doi:10.1001/jamanetworkopen.2025.55759.
30. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228–247. doi:10.1016/j.ejca.2008.10.026.

31. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekas S, et al. iRECIST: Guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol.* 2017;18(3):e143–e152. doi:10.1016/S1470-2045(17)30074-8.
32. Lin NU, Lee EQ, Aoyama H, Barani IJ, Barboriak DP, Baumert BG, Bendszus M, Brown PD, Camidge DR, Chang SM, et al. Response assessment criteria for brain metastases: Proposal from the RANO group. *Lancet Oncol.* 2015;16(6):e270–e278. doi:10.1016/S1470-2045(15)70057-4.
33. Okada H, Weller M, Huang R, Finocchiaro G, Gilbert MR, Wick W, Ellingson BM, Hashimoto N, Pollack IF, Brandes AA, et al. Immunotherapy response assessment in neuro-oncology: A report of the RANO working group. *Lancet Oncol.* 2015;16(15):e534–e542. doi:10.1016/S1470-2045(15)00088-1.
34. Johannet P, Coudray N, Donnelly DM, Jour G, Illa-Bochaca I, Xia Y, et al. Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma. *Clin Cancer Res.* 2021;27(1):131–140. doi:10.1158/1078-0432.CCR-20-2415.
35. Rakae 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):109–118. doi:10.1001/jamaoncol.2024.5356.
36. Vanguri RS, Luo J, Aukerman AT, Egger JV, Fong CJ, Horvat N, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small-cell lung cancer. *Nat Cancer.* 2022;3(10):1151–1164. doi:10.1038/s43018-022-00416-8.
37. Yeghaian M, Bodalal Z, Tareco Bucho TM, Kurilova I, Blank CU, Smit EF, et al. Integrated noninvasive diagnostics for prediction of survival in immunotherapy. *Immuno-oncol Technol.* 2024;24:100723. doi:10.1016/j.iotech.2024.100723.
38. Chang TG, Cao Y, Sfreddo HJ, Dhruba SR, Lee SH, Valero C, et al. LORIS robustly predicts patient outcomes with immune checkpoint

- blockade therapy using common clinical, pathologic and genomic features. *Nat Cancer*. 2024;5(8):1158–1175. doi:10.1038/s43018-024-00772-7.
39. Yoo SK, Fitzgerald CW, Cho BA, Fitzgerald BG, Han C, Koh ES, Pandey A, et al. Prediction of checkpoint inhibitor immunotherapy efficacy for cancer using routine blood tests and clinical data. *Nat Med*. 2025;31:869–880. doi:10.1038/s41591-024-03398-5.
 40. Lippenszky L, Mittendorf KF, Kiss Z, LeNoue-Newton ML, et al. Prediction of effectiveness and toxicities of immune checkpoint inhibitors using real-world patient data. *JCO Clin Cancer Inform*. 2024;8:e2300207. doi:10.1200/CCI.23.00207.
 41. Apetoh L, Ghiringhelli F, Tesniere A, Obeid M, Ortiz C, et al. Toll-like receptor 4-dependent contribution of the immune system to anticancer chemotherapy and radiotherapy. *Nat Med*. 2007;13(9):1050–1059. doi:10.1038/nm1622.
 42. Reits EA, Hodge JW, Herberts CA, Groothuis TA, Chakraborty M, et al. Radiation modulates the peptide repertoire, enhances MHC class I expression, and induces successful antitumor immunotherapy. *J Exp Med*. 2006;203(5):1259–1271. doi:10.1084/jem.20052494.
 43. Burnette BC, Liang H, Lee Y, Chlewicki L, Khodarev NN, et al. The efficacy of radiotherapy relies upon induction of type I interferon-dependent innate and adaptive immunity. *Cancer Res*. 2011;71(7):2488–2496. doi:10.1158/0008-5472.CAN-10-2820.
 44. Dewan MZ, Galloway AE, Kawashima N, Dewyngaert JK, Babb JS, et al. Fractionated but not single-dose radiotherapy induces an immune-mediated abscopal effect when combined with anti-CTLA-4 antibody. *Clin Cancer Res*. 2009;15(17):5379–5388. doi:10.1158/1078-0432.CCR-09-0265.
 45. Vanpouille-Box C, Alard A, Aryankalayil MJ, Sarfraz Y, Diamond JM, et al. DNA exonuclease Trex1 regulates radiotherapy-induced tumour immunogenicity. *Nat Commun*. 2017;8:15618. doi:10.1038/ncomms15618.

46. Demaria S, Guha C, Schoenfeld J, Morris Z, Monjazeb A, et al. Radiation dose and fraction in immunotherapy: One-size regimen does not fit all settings, so how does one choose? *J Immunother Cancer*. 2021;9(4):e002038. doi:10.1136/jitc-2020-002038.
47. Dovedi SJ, Adlard AL, Lipowska-Bhalla G, McKenna C, Jones S, Cheddle EJ, Stratford IJ, Poon E, Morrow M, Stewart R, et al. Acquired resistance to fractionated radiotherapy can be overcome by concurrent PD-L1 blockade. *Cancer Res*. 2014;74(19):5458–5468. doi:10.1158/0008-5472.CAN-14-1258.
48. Marciscano AE, Ghasemzadeh A, Nirschl TR, Theodros D, Kochel CM, Francica BJ, Muroyama Y, Anders RA, Sharabi AB, Velarde E, et al. Elective nodal irradiation attenuates the combinatorial efficacy of stereotactic radiation therapy and immunotherapy. *Clin Cancer Res*. 2018;24(20):5058–5071. doi:10.1158/1078-0432.CCR-17-3427.
49. Pike LRG, Bang A, Mahal BA, Taylor A, Krishnan M, Spektor A, Cagney DN, Aizer AA, Alexander BM, Rahma OE, et al. The impact of radiation therapy on lymphocyte count and survival in metastatic cancer patients receiving PD-1 immune checkpoint inhibitors. *Int J Radiat Oncol Biol Phys*. 2019;103(1):142–151. doi:10.1016/j.ijrobp.2018.09.010.
50. Ladbury CJ, Rusthoven CG, Camidge DR, Kavanagh BD, Nath SK. Impact of radiation dose to the host immune system on tumor control and survival for stage III non-small cell lung cancer treated with definitive radiation therapy. *Int J Radiat Oncol Biol Phys*. 2019;105(2):346–355. doi:10.1016/j.ijrobp.2019.05.064.
51. Antonia SJ, Villegas A, Daniel D, Vicente D, Murakami S, et al. Durvalumab after chemoradiotherapy in stage III non-small-cell lung cancer. *N Engl J Med*. 2017;377(20):1919–1929. doi:10.1056/NEJMoa1709937.
52. Antonia SJ, Villegas A, Daniel D, Vicente D, Murakami S, et al. Overall survival with durvalumab after chemoradiotherapy in stage III NSCLC. *N Engl J Med*. 2018;379(24):2342–2350. doi:10.1056/NEJMoa1809697.
53. Spigel DR, Faivre-Finn C, Gray JE, Vicente D, Planchard D, Paz-Ares L, et al. Five-year survival outcomes from the PACIFIC trial: Durvalumab

- after chemoradiotherapy in stage III non-small-cell lung cancer. *J Clin Oncol.* 2022;40(12):1301–1311. doi:10.1200/JCO.21.01308.
54. Theelen WSME, Chen D, Verma V, Hobbs BP, Peulen HMU, et al. Pembrolizumab with or without radiotherapy for metastatic non-small-cell lung cancer: A pooled analysis of two randomised trials. *Lancet Respir Med.* 2021;9(5):467–475. doi:10.1016/S2213-2600(20)30391-X.
 55. Jabbour SK, Lee KH, Frost N, Breder V, Kowalski DM, et al. Pembrolizumab plus concurrent chemoradiation therapy in patients with unresectable, locally advanced, stage III non-small cell lung cancer: The phase 2 KEYNOTE-799 nonrandomized trial. *JAMA Oncol.* 2021;7(9):1351–1359. doi:10.1001/jamaoncol.2021.2301.
 56. Bradley JD, Sugawara S, Lee KH, et al. Simultaneous durvalumab and platinum-based chemoradiotherapy in unresectable stage III non-small cell lung cancer: The phase III PACIFIC-2 study. *J Clin Oncol.* 2025;43(33):3610–3621. doi:10.1200/JCO-25-00036.
 57. Peters S, Tan DSW, Gerber DE, Urbanic J, Ramalingam S, et al. Nivolumab plus chemoradiotherapy followed by nivolumab with or without ipilimumab for untreated locally advanced stage III NSCLC: A randomized phase 3 trial. *Nat Cancer.* 2026. doi:10.1038/s43018-026-01161-y.
 58. Li Y, Zhang M, Hu Y, Du B, Mo Y, He T, et al. Pretreatment CT-based machine learning radiomics model predicts response in inoperable stage III NSCLC treated with concurrent radiochemotherapy plus PD-1 inhibitors. *Technol Cancer Res Treat.* 2025;24:15330338251351109. doi:10.1177/15330338251351109.
 59. Peng H, Moore C, Zhang Y, Saha D, Jiang S, Timmerman R. An AI-based approach for modeling the synergy between radiotherapy and immunotherapy. *Sci Rep.* 2024;14:8250. doi:10.1038/s41598-024-58684-6.
 60. Yang M, Ma J, Zhang C, Zhang L, Xu J, Liu S, Li J, Han J, Hu S. Multi-modal data deep learning method for predicting symptomatic pneumonitis caused by lung cancer radiotherapy combined with immunotherapy. *Front Immunol.* 2025;15:1492399. doi:10.3389/fimmu.2024.1492399.

61. Kraus KM, Oreshko M, Schnabel JA, Bernhardt D, Combs SE, Peeken JC. Dosimetrics and radiomics-based prediction of pneumonitis after radiotherapy and immune checkpoint inhibition: The relevance of fractionation. *Lung Cancer*. 2024;189:107507. doi:10.1016/j.lungcan.2024.107507.
62. Wang K, Zhao J, Duan J, Feng C, Li Y, Li L, Yuan S. Radiomic and dosimetric parameter-based nomogram predicts radiation esophagitis in patients with non-small cell lung cancer undergoing combined immunotherapy and radiotherapy. *Front Oncol*. 2024;14:1490348. doi:10.3389/fonc.2024.1490348.
63. Masuda T, Kawahara D, Daido W, Imano N, Matsumoto N, Hamai K, et al. A radiomics-based machine learning model for predicting pneumonitis during durvalumab treatment in locally advanced NSCLC. *AI*. 2026;7(1):32. doi:10.3390/ai7010032.
64. Tejani AS, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300. doi:10.1148/ryai.240300.
65. Hurkmans C, Bibault JE, Brock KK, van Elmpt W, Feng M, 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. doi:10.1016/j.radonc.2024.110345.
66. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
67. Moons KGM, Damen JAA, Kaul T, Hooft L, et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
68. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749–762. doi:10.1038/nrclinonc.2017.141.

69. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, et al. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554. doi:10.1136/bmj-2024-081554.
70. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva: World Health Organization; 2021. ISBN 978-92-4-002920-0.
71. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nat Med*. 2020;26:1351–1363. doi:10.1038/s41591-020-1037-7.
72. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat Med*. 2020;26:1364–1374. doi:10.1038/s41591-020-1034-x.
73. 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):924–933. doi:10.1038/s41591-022-01772-9.
74. International Medical Device Regulators Forum. Good Machine Learning Practice for Medical Device Development: Guiding Principles. Technical Report IMDRF/AIML WG/N88 FINAL:2025. International Medical Device Regulators Forum. Published January 29, 2025. Accessed July 20, 2026. Available at: <https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles>.