Human Immunology 87 (2026) 111774

Contents lists available at ScienceDirect

Human Immunology

journal homepage: www.elsevier.com/locate/humimm

Review

Translational Gaps in Immuno-AI: From algorithmic accuracy to clinical trust

Emmanuel O. Oisakede [a,b], Raphael Igbarumah Ayo Daniel [c], Olabanke Florence Olawuyi [d], John Oluwatosin Alabi [e], Claret Chinenyenwa Analikwu [f], David B. Olawade [g,h,i,*]

[a] Department of Clinical Oncology, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom
[b] Department of Health Research, University of Leeds, Leeds, United Kingdom
[c] Department of Social and Health Sciences, Faculty of Social and Life Sciences, Wrexham Glyndŵr University, Wrexham LL11 2AW, United Kingdom
[d] Department of Biomedical Sciences, College of Medical Sciences, University of Calabar, Calabar, Cross River State, Nigeria
[e] Department of Business and Management, University of Sussex Business School, University of Sussex, Falmer, Brighton BN1 9RH, United Kingdom
[f] Department of Microbiology, Frimley Health NHS Foundation Trust, Surrey GU16 7UJ, United Kingdom
[g] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London E16 2RD, United Kingdom
[h] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham, Kent ME7 5NY, United Kingdom
[i] Department of Public Health, York St John University, London E14 2BA, E14 2BA, United Kingdom

ARTICLE INFO

Keywords:
Immunotherapy
Artificial intelligence
Immune checkpoint inhibitors
Clinical translation
Validation
Interpretability

ABSTRACT

Artificial intelligence has shown remarkable promise in predicting patient responses to immune checkpoint inhibitors across cancers. However, despite high statistical performance, clinical translation remains minimal. This disconnect between algorithmic accuracy and clinical adoption, termed the translational gap, reflects unresolved challenges in validation, interpretability, and regulatory integration. This review critically examines key barriers preventing translation of Immuno-AI systems from research prototypes to clinically trusted decision-support tools. It analyzes methodological, regulatory, ethical, and infrastructural factors limiting implementation and proposes strategies for developing clinically trustworthy AI in immuno-oncology. A structured literature search was conducted across PubMed, Embase, Scopus, and Web of Science for studies published 2018–2025 reporting AI or machine learning models predicting ICI response or toxicity in human cohorts. Narrative synthesis was applied, focusing on translational bottlenecks. Three dominant factors underpin the translational gap: (1) insufficient external and prospective validation, leading to overestimation of model performance; (2) limited interpretability and absence of explainable frameworks suitable for clinical use; and (3) regulatory and infrastructural immaturity, including lack of harmonised standards for adaptive AI systems. These limitations contribute to absence of clinician confidence and hinder regulatory approval. Bridging the translational gap in Immuno-AI requires a shift from model-centric optimisation to system-level accountability. Clinically trustworthy AI must be validated across institutions, designed for interpretability, and governed by transparent, ethical frameworks. Collaborative efforts among researchers, clinicians, and regulators are essential to ensure future Immuno-AI systems achieve algorithmic excellence, clinical credibility, and social legitimacy.

1. Introduction

The application of artificial intelligence (AI) in immuno-oncology has developed rapidly over the past decade, particularly in the field of immune checkpoint inhibitor (ICI) therapy. Machine-learning and deep-learning models have shown promising predictive performance in oncology, including the prediction of survival/recurrence outcomes, treatment response, and treatment-related toxicities or adverse drug reactions across multiple cancer settings [1,2]. Yet despite this computational progress, only a small fraction of these predictive models has achieved successful translation into clinical practice [3]. The gap between algorithmic accuracy and clinical utility represents a major challenge for the future of precision immunotherapy.

Early enthusiasm for AI-based prediction models stemmed from their ability to process large, multidimensional datasets that traditional statistical approaches could not adequately capture. Studies integrating

* Corresponding author.
E-mail address: d.olawade@uel.ac.uk (D.B. Olawade).

https://doi.org/10.1016/j.humimm.2026.111774
Received 13 December 2025; Received in revised form 15 May 2026; Accepted 1 June 2026
Available online 10 June 2026

genomic, histopathological and clinical features have reported area under the curve (AUC) values exceeding 0.8 for ICI response prediction [4–6]. However, when tested in independent or real-world cohorts, immuno-AI models often show reduced performance on external validation, with documented declines ranging from modest to clinically meaningful depending on the model and dataset [7]. This recurring loss of accuracy underscores a fundamental problem: computational excellence within a development dataset does not guarantee reproducibility or clinical trust.

The reasons for this translational gap are multifactorial. Previous reviews of oncology AI identify data heterogeneity, limited multicenter external validation, non-standardized workflows/reporting, and limited explainability as major barriers to clinical adoption [8,9]. End-user studies in oncological pathology likewise show that implementation is slowed by concerns about workflow disruption, accountability, and trust, reflecting the inherently risk-sensitive nature of clinical environments [10,11]. Without clear interpretability, validated reproducibility, and transparent governance, clinicians are reluctant to base treatment decisions on algorithmic recommendations, regardless of technical sophistication [3].

The field now faces a paradox: the technological capability to predict outcomes has advanced faster than the systems required to implement these predictions responsibly. The term Immuno-AI, the application of AI to immunotherapy has emerged to describe these models. Yet the clinical reality reveals an ecosystem of fragmented data sources, uncoordinated validation strategies, and minimal regulatory oversight. Translating computational findings into actionable clinical tools requires more than improved algorithms, it demands a systematic framework for validation, transparency, and clinician engagement.

This review examines the current evidence for predictive Immuno-AI systems and critically analyses the barriers preventing their translation from algorithmic development to clinical adoption. It evaluates the methodological foundations of Immuno-AI models, identifies recurring limitations in validation and implementation, and explores strategies for building clinical trust. By focusing on the interface between algorithmic performance and healthcare integration, this paper seeks to define what is required to achieve clinically reliable, ethically sound and sustainable Immuno-AI systems.

2. Methods

2.1. Search strategy and scope

A structured literature search was conducted to identify relevant studies on AI-based predictive models for immune checkpoint inhibitor response or toxicity. Searches were performed in PubMed/MEDLINE, Embase, Web of Science, and Scopus for articles published between January 2018 and December 2025. The following database-specific Boolean search string was applied, with minor field-tag adaptations for each platform: (“immune checkpoint inhibitor*” OR “checkpoint blockade” OR “anti-PD-1” OR “anti-PD-L1” OR “anti-CTLA-4” OR “immunotherapy”[MeSH]) AND (“artificial intelligence” OR “machine learning” OR “deep learning” OR “neural network*” OR “random forest” OR “predictive model*” OR “classification algorithm*”) AND (“response prediction” OR “survival prediction” OR “toxicity prediction” OR “adverse event*” OR “treatment outcome*” OR “biomarker*”) AND (“validation” OR “external validation” OR “clinical implementation” OR “interpretability” OR “explainability”). In PubMed/MEDLINE, MeSH terms were combined with free-text terms using Boolean OR within each concept block, and blocks were combined using AND. Truncation (*) was applied to capture plural and variant forms. Embase searches additionally used EMTREE controlled vocabulary. In Scopus and Web of Science, title, abstract, and keyword fields (TITLE-ABS-KEY) were searched using equivalent syntax. No language restrictions were applied; however, only English-language full-text articles were included at the full-text review stage.

The review focused on studies that developed or evaluated AI or machine-learning algorithms applied to human subjects receiving immune checkpoint inhibitors. Publications were included if they reported quantitative performance metrics (e.g., AUC, sensitivity, specificity, accuracy) and provided information on validation methodology or clinical application. Preclinical studies using only cell lines or animal models, conference abstracts without peer-reviewed articles, and studies unrelated to predictive modelling were excluded.

2.2. Data extraction and synthesis

Eligible studies were reviewed for study design, cancer type, sample size, model type, data modality (genomic, radiomic, histopathologic, or clinical), validation approach, and reported performance. Special emphasis was placed on differentiating between internal validation, external validation using independent cohorts, and real-world deployment. Methodological transparency, reproducibility, and interpretability were assessed qualitatively. The screening and selection process followed a stepwise approach consistent with PRISMA reporting principles. Title and abstract screening were performed independently by two reviewers, with full-text review applied to all articles meeting provisional inclusion criteria. Disagreements were resolved by consensus. Studies were classified according to the primary translational barrier they addressed: validation deficit, interpretability gap, regulatory immaturity, or implementation challenge. The study selection process is illustrated in Fig. 1; of 4,812 records identified through database searching (PubMed/MEDLINE, Embase, Scopus, and Web of Science) and 83 additional records from other sources, 3,641 records remained after deduplication, of which 592 full-text articles were assessed for eligibility and 57 studies were ultimately included in the narrative synthesis. Final inclusion encompassed studies reporting quantitative performance metrics alongside sufficient methodological detail to evaluate translational applicability.

Given the diversity of algorithms and outcome measures, formal *meta*-analysis was not appropriate. Instead, a narrative synthesis approach was applied, grouping studies according to methodological themes: (1) algorithmic accuracy and validation, (2) interpretability and explainability, (3) regulatory and ethical considerations, and (4) clinical implementation challenges. Supplementary searches were performed using the reference lists of recent systematic reviews and key primary studies to ensure comprehensive coverage of relevant literature.

2.3. Critical appraisal

Each included study was appraised for methodological quality using the Prediction model Risk Of Bias ASsessment Tool (PROBAST), which is specifically designed for studies developing or validating multivariable prediction models and is more appropriate for this literature. PROBAST evaluates risk of bias across four domains: participants, predictors, outcome, and analysis. Studies were rated as low, high, or unclear risk of bias within each domain, with overall risk-of-bias judgement determined by the highest domain-level rating. Particular attention was paid to analytical domain concerns prevalent in Immuno-AI literature, including data leakage, absence of calibration assessment, and single-institution training without external validation. Overall risk of bias was rated high in 38 studies (67%), unclear in 12 (21%), and low in 7 (12%). High risk of bias in the analysis domain was the most prevalent concern, identified in 44 studies (77%), most commonly attributable to absence of calibration assessment, use of inappropriate performance metrics as the sole validation measure, or failure to account for missing data. The participant domain was rated high risk in 26 studies (46%), principally due to single-institution or convenience sampling. The predictors domain was high risk in 18 studies (32%), primarily owing to post hoc biomarker selection without pre-specification. The outcome domain was high risk in 14 studies (25%), reflecting inconsistent or surrogate endpoint definitions across cancer types and treatment

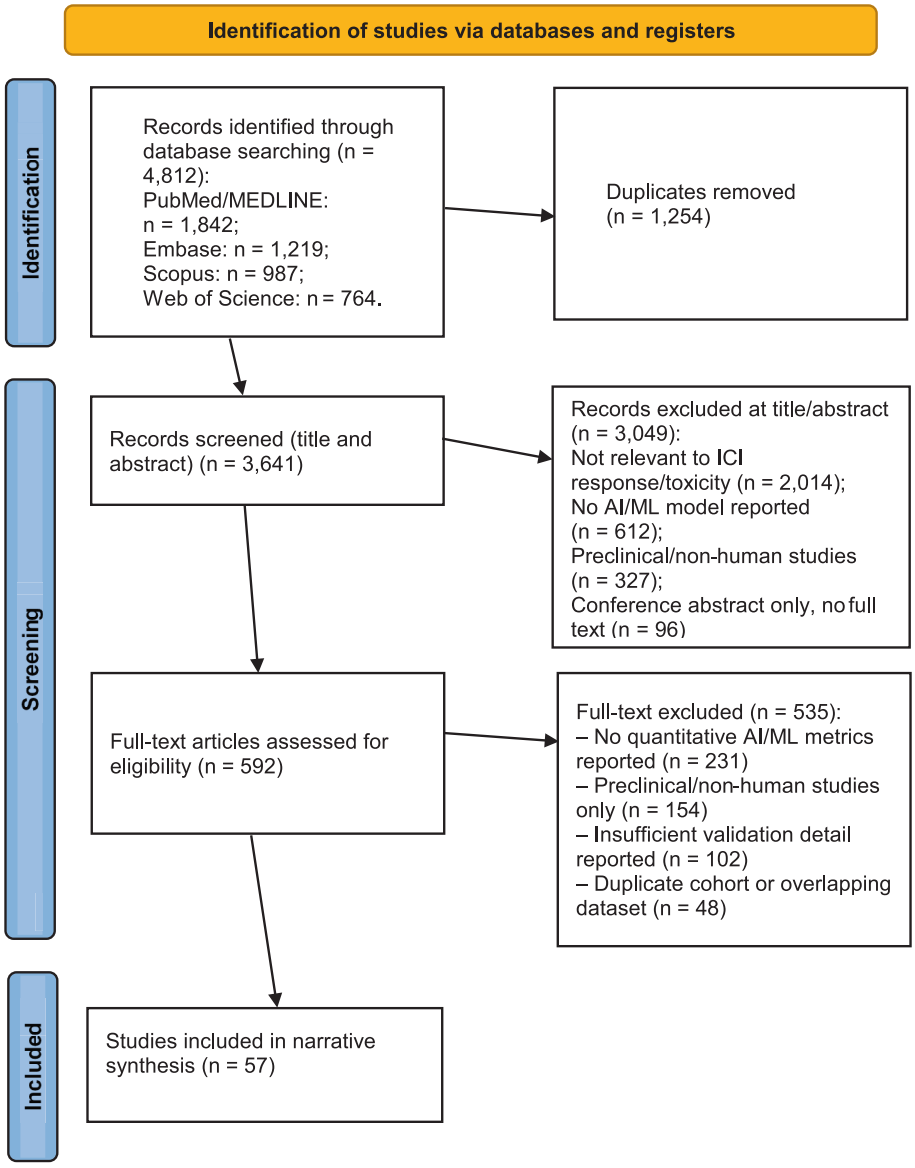

Fig. 1. PRISMA flow diagram for the articles' selection process.

settings.

Reporting quality was assessed against TRIPOD + AI, the updated extension of the original TRIPOD statement incorporating AI-specific requirements for model transparency, data handling, and reproducibility. Where studies reported randomised or prospective evaluation, concordance with CONSORT-AI was additionally assessed. Studies were not excluded on the basis of reporting quality alone; rather, reporting deficiencies were noted as a dimension of translational readiness and factored into the narrative synthesis. Reporting quality findings indicated that concordance with TRIPOD + AI was partial or absent in 46 of 57 included studies (81%). The most frequently missing items were explicit reporting of model calibration (absent in 41 studies, 72%), preregistration of analysis plans (absent in 49 studies, 86%), and disclosure of hyperparameter tuning procedures (absent in 38 studies, 67%). Of the nine studies meeting criteria for CONSORT-AI assessment, full concordance was achieved in three. These findings confirm that reporting deficiencies are pervasive across the Immuno-AI literature and are themselves a barrier to translational progress, as inadequate reporting prevents reproducibility assessment and limits the evidentiary weight of positive findings.

3. The evolution of Immuno-AI

Artificial intelligence in immuno-oncology has developed from exploratory computational models into a core area of translational cancer research (see Fig. 2). Early predictive approaches focused on the identification of single biomarkers such as PD-L1 expression and tumour mutational burden (TMB). These parameters were incorporated into conventional statistical models or logistic regression frameworks to stratify patients for immune checkpoint inhibitors [12]. Although these markers demonstrated some predictive value, their limitations soon became evident. Variable assay methodologies, tumour heterogeneity and lack of reproducibility restricted their clinical reliability across diverse cancer populations [4].

Between 2018 and 2023, a major shift occurred with the introduction of machine learning algorithms capable of processing complex, multidimensional datasets. Models began to integrate histopathological images, genomic data and clinical variables to improve response prediction. The most successful examples include the SCORPIO and LORIS frameworks, which demonstrated higher accuracy compared with single-biomarker models [4,13,14]. Although these models have successfully outperformed traditional models with AUC > 0.85, external

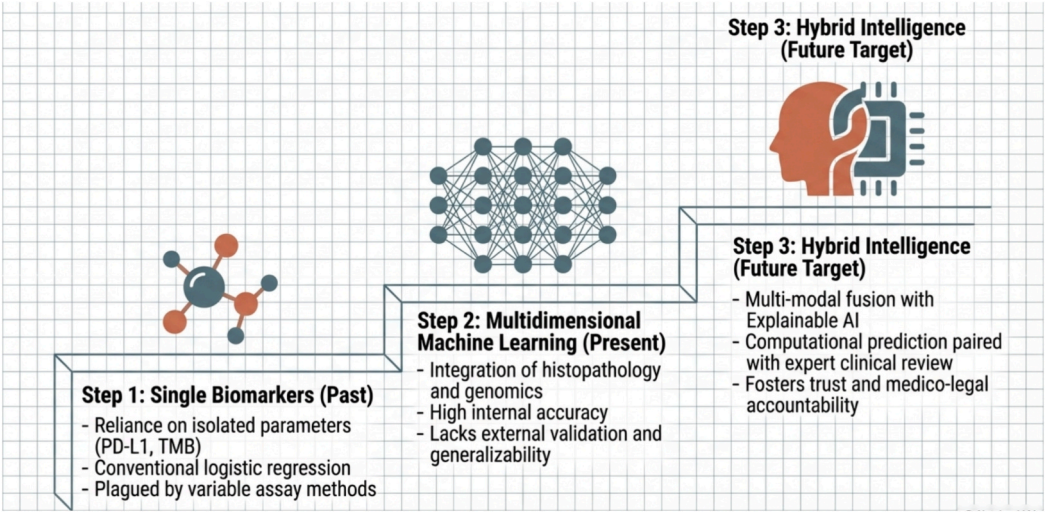

Fig. 2. Evolution of Immuno-Oncology Models.

validation outside the development institutions was often limited, revealing persistent generalisability issues [4].

Recent years have seen a shift toward multi-modal and explainable AI systems. These integrate imaging, transcriptomic and clinical data while providing interpretable outputs through tools such as Shapley additive explanations (SHAP) or saliency mapping [15,16]. This has encouraged more human-in-the-loop workflows in which computational predictions are interpreted alongside expert pathological assessment. Illustrative examples from adjacent computational biology fields underscore both the promise and the interpretability challenges of this approach. xSiGra, an interpretable graph-based AI model leveraging multimodal spatial transcriptomics data, employs hybrid graph transformer architectures and a novel gradient-weighted class activation mapping (graph Grad-CAM) component to identify pivotal genes and cells contributing to spatial cell type identification in tumour microenvironments [17]. Similarly, SpaIM, a style-transfer learning framework, integrates single-cell RNA sequencing data with spatial transcriptomics profiles to impute unmeasured gene expressions across 53 diverse datasets, demonstrating superior performance over 12 existing methods and enhancing downstream analyses including ligand–receptor interaction inference [18]. These models exemplify the trajectory towards architectures that are not only accurate but structurally interpretable, an essential design criterion for future Immuno-AI deployment. However, even with these developments, evidence suggests that most Immuno-AI systems remain at the proof-of-concept stage. Few have progressed through prospective clinical trials or regulatory submission, indicating that technological maturity has outpaced translational readiness.

The evolution of Immuno-AI therefore reflects an ongoing tension between innovation and implementation. Each generation of models has produced measurable gains in analytical performance, yet the same fundamental weaknesses, such as insufficient validation, interpretability and standardisation have persisted. This historical context provides an essential foundation for examining why algorithmic accuracy has not yet translated into clinical trust.

4. Algorithmic Accuracy: The computational success Story

Algorithmic performance has been one of the most visible strengths of Immuno-AI research. In development cohorts, multimodal machine-learning and deep-learning models often report higher discriminatory performance than conventional biomarkers such as PD-L1 or TMB. For example, Wang et al. [19] reported an AUC of 0.77 for DeepAFM, a multimodal model integrating histopathology, genomic alterations, and clinical variables in advanced NSCLC. In a broader oncology example, Goyal et al. [6] reported AUCs of 0.91 internally and 0.84 externally for a combined whole-slide-imaging and clinicopathologic model predicting breast-cancer recurrence risk. Such results have helped fuel optimism that AI could improve biomarker discovery and patient stratification for immunotherapy.

A clearer taxonomy of Immuno-AI models is needed to evaluate where translation fails most consistently. Across the reviewed literature, models can be broadly categorised by (i) data modality (e.g., genomic, radiomic, histopathological, or multimodal); (ii) clinical endpoint, ICI response prediction, immune-related adverse event (irAE) prediction, or overall survival; (iii) cancer type, such as non-small cell lung cancer (NSCLC), melanoma, urothelial carcinoma, or pan-cancer cohorts; and (iv) validation stage, i.e., internal only, external retrospective, or prospective. Genomic and multimodal models have shown strong discriminatory performance in retrospective cohorts, while histopathology-based and radiomic models provide additional clinically scalable examples. However, external validation frequently reduces performance, and most published models remain retrospective rather than prospectively tested. Radiomic models exhibit greater generalisability across imaging platforms but are highly sensitive to scanner-protocol heterogeneity. Multimodal fusion architectures, while computationally superior, are the least consistently validated externally and carry the greatest risk of overfitting. Importantly, models predicting irAEs remain substantially underrepresented in the literature relative to response prediction models, despite irAEs representing a major clinical determinant of treatment continuation. This imbalance limits the translational scope of current Immuno-AI evidence. Critically, the translational implications differ substantially across the three primary endpoint categories identified in this review. For response prediction models, the dominant barrier is external validation failure: internal AUC values typically exceed 0.80 but decline meaningfully on external cohorts, and the clinical decision context, whether to initiate, continue, or switch ICI therapy, requires not just discrimination but calibrated probability estimates. For survival prediction models, the principal challenge is endpoint heterogeneity: overall survival, progression-free survival, and disease-specific survival are frequently conflated across studies, making cross-study synthesis and regulatory evaluation difficult. For irAE prediction models, the translational barrier is primarily one of evidence scarcity and dataset size: irAEs are heterogeneous, organ-specific, and relatively low-frequency events that require large, prospective datasets to model accurately, and most published irAE prediction tools have been developed on small single-institution cohorts with limited generalisability. The narrative in subsequent sections reflects these endpoint-specific distinctions rather than treating Immuno-AI prediction as a

unified task. Table 1 critically exemplifies representative Immuno-AI studies by modality, endpoint, cancer type, and validation stage exposing where translation tends to weaken.

Closer analysis of these results reveals important limitations. Many of the highest-performing models are trained on single-institution datasets with homogeneous populations and tightly controlled imaging or sequencing protocols [4]. This setting minimises confounding variables but does not reflect the variability of real-world oncology practice. Oncology-focused methodological studies show that sample selection and centre-specific bias, limited representativeness, and overfitting can make internally validated machine-learning models appear more accurate than they are in new clinical settings, thereby weakening real-world performance [36]. Concrete cases illustrate the scale of this problem. In a multicenter cohort of 958 patients with advanced NSCLC treated with ICI monotherapy, Rakaee et al. [21] reported that a deep-learning pathology model achieved an AUC of 0.75 in the internal test set but only 0.66 in the external validation cohort. Similarly, in a study of neoadjuvant immunotherapy response prediction in NSCLC, She et al. [37] found that a combined deep-learning model achieved AUC 0.77 in internal validation and 0.75 in external validation. Together, these studies suggest that performance typically becomes more modest when models are tested beyond their development setting.

Another concern is the narrow reliance on statistical indicators such as AUC or F1 score as proxies for clinical utility. These metrics quantify discrimination but not calibration, which refers to the alignment between predicted probabilities and observed outcomes. A model can achieve an AUC of 0.85 but still produce unreliable predictions in a clinical context if calibration is poor. Studies evaluating calibration plots and decision-curve analyses are comparatively rare in Immuno-AI research, limiting confidence in the practical reliability of reported models. Across the 57 included studies, calibration was formally assessed in only 16 studies (28%), of which 11 used calibration plots, 4 reported Hosmer-Lemeshow goodness-of-fit tests, and 1 used isotonic regression recalibration. Decision-curve analysis, which evaluates net clinical benefit across decision thresholds and is considered the preferred framework for assessing clinical utility of prediction models, was performed in only 8 studies (14%). Clinically actionable decision thresholds, defined as pre-specified probability cut-offs with associated clinical actions (e.g., "initiate ICI if predicted response probability $>$ 0.65"), were absent in 49 studies (86%). This systematic absence of calibration reporting, decision-curve analysis, and threshold specification means that most published Immuno-AI models cannot be directly translated into clinical decision rules, regardless of their discrimination performance. Future studies should report calibration metrics alongside AUC, conduct decision-curve analysis as standard, and pre-specify clinically meaningful decision thresholds to support translational credibility. However, findings from studies have reported increasedincreased accuracy when calibration protocols are used to validate models with higher AUC. For example, a recent study by Veeraraghavan et al. [38] on survival prediction model for oral squamous cell carcinoma explicitly included a calibration assessment to further validate an AUC $>$ 0.8 with a strong accuracy margin.

Benchmarking against traditional biomarkers reveals a subtler issue. Although some AI models report better discriminatory performance than single biomarkers such as PD-L1 or TMB, improved AUC alone does not guarantee clinical utility. In oncology, implementation depends on whether models are well calibrated, integrated into workflow, and linked to actionable treatment recommendations rather than isolated risk scores [39]. Many published models still lack clearly defined decision thresholds or explicit guidance on how predictions should alter treatment [40]. Without clinical interpretability and decision relevance, high numerical accuracy alone offers limited value at the point of care.

Table 1
Immuno-AI models by data modality, clinical endpoint, cancer type, validation stage, and translational implication.

Representative study	Data modality	Clinical endpoint	Cancer type	Validation stage	Reported performance	What it shows
Li et al. [20]	Genomic (ctDNA mutation profiles; XGBoost)	Progression-free survival / immunotherapy benefit	Advanced NSCLC	Retrospective train/ validation/test split	AUC 0.82 in training, 0.79 in validation, 0.77 in test set	Genomic models can achieve strong discrimination, but this is still not the same as broad external or prospective validation.
Rakaee et al. [21]	Histopathology (H&E whole-slide deep learning)	ICI response stratification	Advanced NSCLC	External retrospective validation across US/EU cohorts	AUC 0.75 internally, falling to 0.66 on independent external validation; combining Deep-IO with PD-L1 improved external AUC to 0.70	A good example of why validation stage matters: performance drops when the model leaves the development setting.
Fa'ak et al. [22]	Histopathology (deep CNN on pretreatment slides)	ICI response / progression-free survival risk	Metastatic melanoma	Multi-site retrospective generalizability testing	Response prediction AUC 0.72	Cross-site generalizability is possible, but performance remains moderate rather than "clinical-grade."
Provenzano et al. [23]	Radiomic (PET/CT radiomics integrated with real-world clinical data)	24-month overall survival / clinical benefit	Advanced NSCLC	Retrospective test-set evaluation in real-world data	Best model predicted 24-month OS with AUC 0.79 and outperformed PD-L1 alone	Radiomic pipelines can be competitive and clinically relevant, especially when combined with clinical variables.
Vanguri et al. [24]	Multimodal (radiology + pathology + genomics)	Response to PD-(L)1 blockade	Advanced NSCLC	Retrospective proof-of-concept cohort	Multimodal model AUC 0.80, versus 0.73 for PD-L1 TPS and 0.61 for TMB	A classic example of multimodal fusion outperforming standard single biomarkers in a curated retrospective cohort.
Wang et al. [19]	Multimodal (histopathology + genomics + clinical data)	Immunotherapy response prediction	Advanced NSCLC	Retrospective held-out testing	AUC 0.77; attention heatmaps highlighted pathological patterns, genomic mutations, and clinical indicators associated with response	Multimodal fusion can improve prediction while also offering some interpretability, but the study still sits at a pre-prospective stage.
Iivanainen et al. [25]	Clinical / ePRO ML	irAE presence and onset	Mixed cancers treated with ICIs	Early-stage validation	For irAE presence: accuracy 0.97, AUC 0.99; for irAE onset: AUC 0.93	Useful as an irAE example, but also shows how much of the irAE-prediction literature is still early and narrow rather than broadly deployment-ready.

Therefore, while algorithmic accuracy demonstrates the technical potential of Immuno-AI, it is not a sufficient indicator of clinical success. The focus on internal validation and performance metrics has created an illusion of progress that masks the deeper translational barriers limiting clinical adoption.

5. The translational Gap: From In-Silico promise to clinical reality

The translational gap in Immuno-AI refers to the persistent failure of predictive models to maintain performance, reproducibility and trustworthiness when applied beyond their original development environment. This gap arises from interrelated methodological, infrastructural and cultural factors.

One of the most pervasive challenges is the lack of external validation. Many published oncology-AI models show strong performance in retrospective cohorts but falter when tested on independent datasets. A substantial proportion of published models do not advance beyond internal validation, primarily due to reproducibility concerns and the absence of independent cohorts with sufficient sample size and demographic diversity. For instance, a recent *meta*-analysis found that only about one-third of studies (104/315) included external validation [41], and a separate scoping review identified just 56 models meeting both external validation and clinical utility criteria [42]. Although these studies are broader than Immuno-AI specifically, they remain relevant because immunotherapy prediction models are developed, validated, and translated within the same oncology-AI pipeline and therefore face the same core challenges of external validation, reproducibility, multicentre generalisability, and demonstration of clinical utility. The absence of multicentre datasets continues to limit generalisability and accurate estimation of real-world performance.

Another source of translational failure is the absence of standardisation across data acquisition and preprocessing steps. Variations in staining protocols, imaging resolution, sequencing depth and feature extraction pipelines introduce inconsistencies that compromise reproducibility. While international frameworks for data sharing are emerging [33], adoption remains inconsistent, especially in low- and middle-income healthcare systems.

Algorithmic opacity further compounds the problem. Deep learning architectures, despite their predictive strength, often function as "black boxes," providing limited insight into how predictions are generated. This lack of interpretability undermines clinician confidence and complicates regulatory review. Explainable AI methods such as SHAP or Local Interpretable Model-agnostic Explanations (LIME) offer partial transparency but have yet to demonstrate consistent interpretive reliability across datasets [28].

Cultural and systemic factors also play a role. Many clinicians remain cautious about integrating algorithmic predictions into decision-making due to medico-legal liability and ethical uncertainty [43]. In parallel, regulatory agencies face difficulties in evaluating continuously learning systems that adapt over time, as traditional medical device approval pathways were not designed for dynamic algorithms [30]. These structural mismatches between innovation and regulation create additional friction in the translational process.

Ultimately, the translational gap reflects a broader misalignment between the objectives of computational research and the requirements of clinical medicine. While data scientists prioritise performance optimisation, clinicians require interpretability, reproducibility and safety assurance. Bridging this divide demands new frameworks that combine rigorous methodological standards with transparent validation pipelines and meaningful clinical engagement. As illustrated in Fig. 3, the translational bottleneck in Immuno-AI results from misalignment between algorithmic objectives and clinical requirements. While model development prioritises accuracy, successful clinical adoption depends on interpretability, reproducibility, and governance mechanisms that remain underdeveloped.

6. Building clinical trust in Immuno-AI

The transition of AI models from development laboratories to clinical oncology settings depends not only on statistical validation but also on the cultivation of trust among clinicians, regulators and patients. Trust in Immuno-AI is multifactorial, encompassing technical transparency, ethical accountability, interpretability, and consistent performance under real-world conditions. Although several studies have demonstrated impressive algorithmic metrics, the absence of clinical trust remains a major barrier to adoption, especially when models are insufficiently explainable, weakly integrated into workflow, or not convincingly validated outside development settings [4,9,44].

6.1. *Explainability and interpretability*

Interpretability remains a cornerstone of clinical trust. For oncologists, confidence in AI-assisted recommendations depends on the ability to understand why a prediction is made. Deep learning models, despite their power, often lack this transparency. Explainable AI (XAI) methods, such as SHAP, Layer-wise Relevance Propagation (LRP) and LIME, have been proposed to visualise decision pathways or highlight critical features influencing predictions [28].

However, interpretability techniques are not without limitations. Many generate post hoc visualisations that may not accurately reflect the model's internal reasoning, leading to the risk of "false

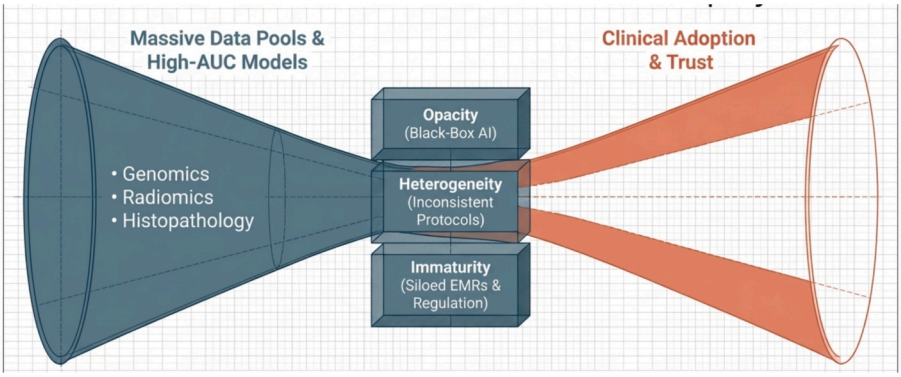

Fig. 3. Conceptual overview of the translational bottlenecks that limit the clinical implementation of Immuno-AI. Predictive models often demonstrate high algorithmic accuracy during internal validation but fail to generalise to independent datasets. Major barriers, such as lack of external validation, data heterogeneity, algorithmic opacity, and regulatory uncertainty undermine clinical confidence and hinder adoption. The figure illustrates how these barriers collectively erode trust and create a feedback loop that reinforces the need for prospective validation, explainable AI design, and ethical oversight.

interpretability" [45]. Furthermore, interpretability outputs often require technical literacy beyond that of the average clinician. For instance, a heatmap overlay on a histopathological image may be scientifically informative but clinically ambiguous (see Fig. 4). The challenge, therefore, is not only to make AI explainable but also to make explanations clinically meaningful. Recent developments in computational biology offer instructive models for how interpretability can be built structurally into AI architectures rather than appended post hoc. The xSiGra model employs a novel graph gradient-weighted class activation mapping (graph Grad-CAM) algorithm to quantify the contribution of individual genes and cells to spatial cell type classification in tumour tissue, achieving a median fidelity score of 0.16 and contrastivity score of 0.77, markedly superior to conventional explainability methods including Saliency, InputXGradient, GuidedBackprop, and Deconvolution [17]. Critically, xSiGra demonstrated that cellular identity is shaped not only by intrinsic gene expression but also by neighbouring cell interactions, a biologically meaningful insight that emerged directly from the interpretable architecture. This approach illustrates what "interpretability by design" can deliver in practice: quantitative, cell-level and gene-level explanations that align with established biological logic and are communicable to domain experts. Immuno-AI models should aspire to equivalent levels of structural transparency, particularly those intended to influence treatment decisions at the point of care.

Effective interpretability frameworks must therefore be co-designed with domain experts. Pathologists and oncologists should be involved from the early stages of model development to ensure that outputs correspond to established clinical logic. Studies where AI and clinicians collaborate in joint interpretive tasks have shown improved diagnostic accuracy and user confidence compared with AI-only or clinician-only assessment [46]. The future of Immuno-AI lies in such hybrid decision systems, where computational insight complements but does not replace human judgement.

6.2. *Human-in-the-Loop validation*

A growing consensus in digital health advocates for "human-in-the-loop" (HITL) validation as a strategy to maintain oversight and accountability during AI deployment. HITL systems integrate continuous clinician interaction at multiple stages of model use; from training data curation to real-time validation of predictions [47]. In oncology, this approach ensures that models learn from clinically verified feedback rather than abstract statistical optimisation (See Fig. 5).

For example, when deep learning models are applied to histopathology, expert pathologists can verify and annotate ambiguous outputs, creating an iterative feedback mechanism that improves both model accuracy and reliability [48]. Similarly, oncologists using predictive models for checkpoint inhibitor response can provide corrective labels based on actual patient outcomes, facilitating real-world learning and adaptation. This iterative validation promotes both algorithmic robustness and clinician trust by establishing a transparent partnership between human expertise and computational inference.

Nevertheless, HITL systems introduce practical challenges. Continuous feedback loops require time, training, and infrastructure for secure data exchange. The design of user interfaces must also allow clinicians to provide input efficiently without workflow disruption. Despite these barriers, evidence suggests that HITL validation remains one of the most effective methods for reconciling algorithmic autonomy with clinical accountability [29,47].

6.3. *Regulatory and ethical oversight*

Regulatory agencies have recognised that AI and machine learning differ fundamentally from traditional medical devices. The U.S. Food and Drug Administration (FDA) has developed a framework for the regulation of software as a medical device (SaMD), distinguishing between "locked" algorithms and those capable of adaptive learning [30]. The European Medicines Agency (EMA) and the UK's Medicines and Healthcare Products Regulatory Agency (MHRA) have adopted similar guidelines, emphasising transparency, risk classification, and post-deployment monitoring [31].

For Immuno-AI systems, regulatory approval hinges on three domains: analytical validity, clinical validity, and clinical utility. Analytical validity ensures the algorithm's technical correctness; clinical validity verifies predictive accuracy across diverse populations; and clinical utility assesses whether model use improves patient outcomes or decision-making. Most published Immuno-AI models fail to meet all three, primarily due to limited multi-institutional validation and absence of prospective clinical trial data.

Ethical oversight complements regulatory control by ensuring

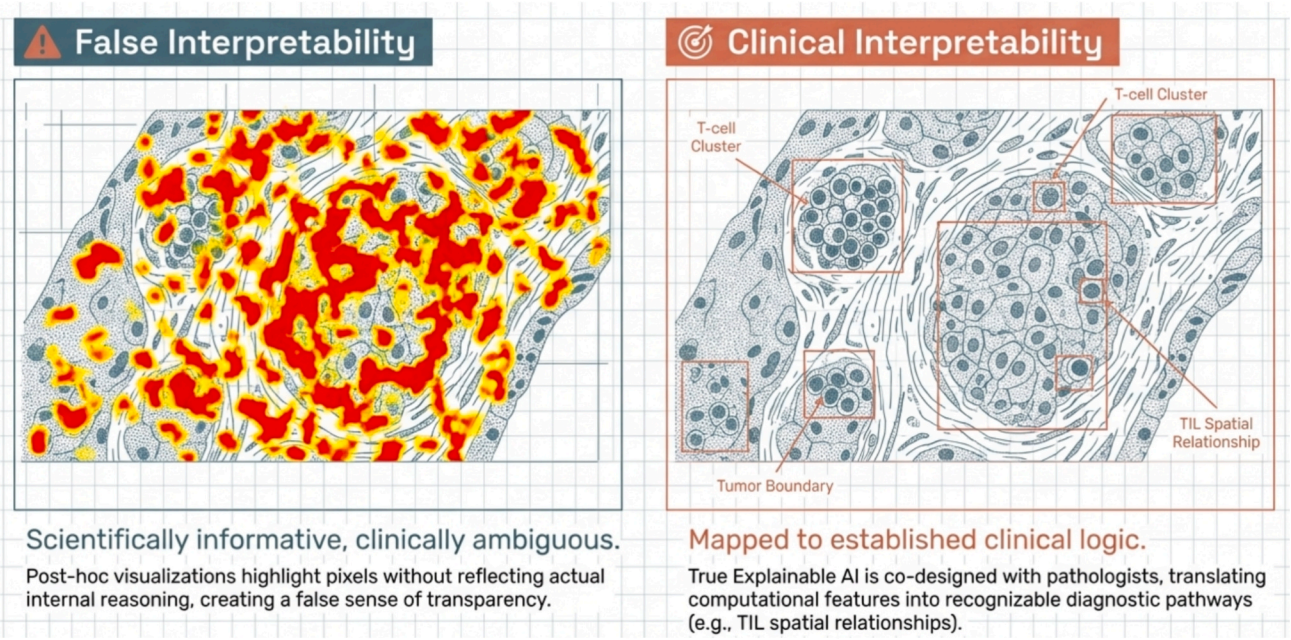

Fig. 4. False Interpretability Vs Clinical Logic.

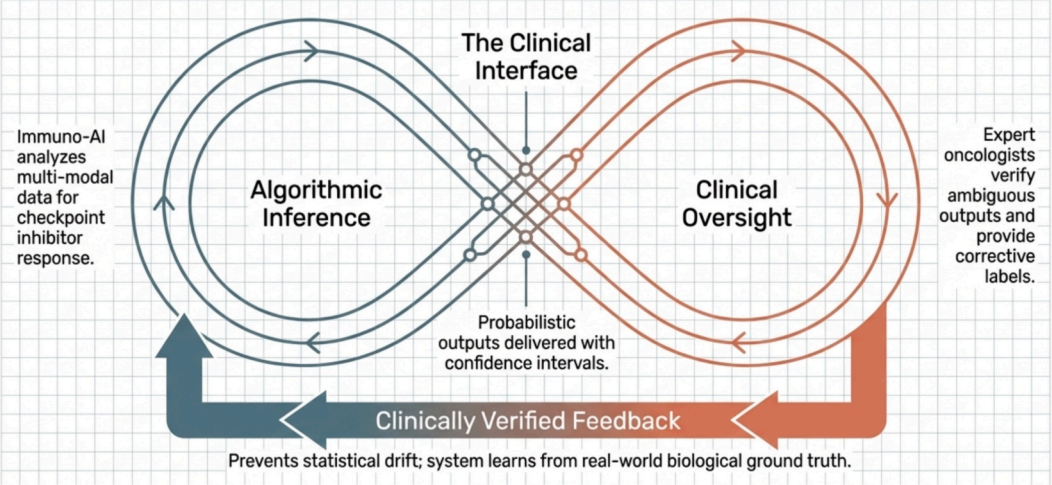

Fig. 5. Engineering trust via Human-in-the-Loop Validation.

fairness, accountability, and privacy protection. Algorithmic bias arising from imbalanced datasets or unrepresentative sampling can exacerbate existing health disparities if not adequately addressed. The STANDING Together Consensus Recommendations emphasize transparent documentation of who is represented in health datasets, how people are represented, and how data are used, alongside proactive evaluation of dataset limitations and their effects across different groups. These recommendations are intended to help identify and mitigate algorithmic bias, improve dataset representativeness, and support more equitable model performance across diverse patient populations [32]. Furthermore, the consensus advocates for ongoing bias auditing and transparency reporting to enhance accountability and improve model reliability in real-world clinical settings. Ethical frameworks must also safeguard patient autonomy, ensuring that AI recommendations do not override shared decision-making processes between clinicians and patients. The recommendations highlight that AI should support, not replace, the clinician-patient relationship, thereby ensuring that patient preferences and values are respected in the decision-making process.

6.4. *Transparency and reproducibility*

Transparency is central to both scientific credibility and clinical trust. Many high-impact AI studies in oncology have been criticised for insufficient reporting of training data, preprocessing steps, and code availability [49]. Without transparency, independent replication becomes impossible, and clinicians cannot assess the reliability of algorithmic outputs.

To address this, the research community has proposed several reporting standards, including the TRIPOD-AI and CONSORT-AI extensions for clinical trials [50,51]. These frameworks mandate detailed disclosure of model architecture, dataset characteristics, hyperparameter tuning, and validation strategy. Despite these standards of reporting, many trials involving AI are at serious risk of bias and below concordant with guideline [52]. Implementation of such standards in Immuno-AI publications is highly recommended for ensuring reproducibility.

Open-source pipelines offer another solution by allowing external researchers to audit algorithms for bias or technical flaws. Projects such as The Cancer Imaging Archive (TCIA) and the International Cancer Genome Consortium (ICGC) exemplify the benefits of transparent data-sharing platforms that accelerate cross-institutional validation [53,54]. Transparency not only strengthens reproducibility but also helps to build the collective scientific confidence necessary for clinical acceptance.

6.5. *Trust Calibration: The human dimension*

Trust in technology cannot be engineered solely through technical design; it must be earned through consistent performance and perceived reliability. In clinical practice, "trust calibration" refers to aligning clinician expectations with the actual capabilities and limitations of AI tools [55]. Over-reliance can lead to automation bias, whereas under-reliance negates potential benefits.

Building appropriate trust requires transparent communication about uncertainty. Rather than presenting binary predictions, Immuno-AI models should convey probabilistic outputs accompanied by confidence intervals and contextual explanations. Clinicians should be trained to interpret these outputs as decision-support aids rather than determinants. Furthermore, embedding feedback mechanisms that allow users to flag suspicious or incorrect predictions reinforces accountability and continual learning.

Ultimately, clinical trust emerges when AI systems demonstrate not only accuracy but also reliability, interpretability and alignment with human values. Trust is therefore both a technical outcome and a social process built through sustained interaction, mutual learning and

Table 2
Translational barriers and potential enablers for immune-AI deployment.

Barrier Category	Underlying Issue	Clinical Implication	Proposed Solution / Enabler
Validation Deficit [26,27]	Lack of external or prospective validation	Overestimated accuracy, loss of generalisability	Multi-institutional trials, shared validation datasets
Interpretability Gap [15,28,29]	"Black-box" models with limited clinical meaning	Low clinician confidence	Explainable AI (SHAP, LIME) and human-in-the-loop review
Regulatory Fragmentation [30,31]	Inconsistent global frameworks for SaMD	Delays in approval, uncertainty for developers	Harmonised FDA–EMA–MHRA guidance
Ethical and Bias Risks [32]	Dataset imbalance and demographic skew	Potential diagnostic inequity	Fairness audits, diverse data sampling
Infrastructure Gap [33]	Incompatible EMR systems, lack of interoperability	Workflow disruption, data silos	Adoption of FHIR and GA4GH standards
Cultural Resistance [34,35]	Limited clinician literacy, fear of liability	Poor adoption and acceptance	AI literacy programs, co-development frameworks

institutional transparency. Table 2 summarises major barriers and facilitators to immune-AI deployment.

7. Clinical implementation and health system integration

The practical integration of AI models into oncology practice is arguably the most challenging and least developed stage of the Immuno-AI pipeline. While algorithmic development and validation dominate the academic literature, comparatively little attention has been paid to how these systems are embedded into everyday healthcare operations. The move from algorithmic accuracy to clinical implementation involves not only technical readiness but also workflow alignment, economic sustainability, and organisational acceptance.

7.1. Integration into clinical workflow

One of the defining features of successful digital innovation in healthcare is its seamless incorporation into existing clinical workflows. For Immuno-AI models, this integration is particularly complex because decision-making in oncology involves multidisciplinary teams, variable data formats, and time-sensitive treatment windows. Predictive algorithms for checkpoint inhibitor response, for example, must interface with pathology results, radiology images, and genomic reports, often stored across disparate hospital systems.

The lack of standardised data infrastructure significantly hinders deployment. Many institutions rely on incompatible electronic medical record (EMR) systems that cannot easily communicate with other similar platforms or AI platforms [56]. Even when integration is technically possible, delays in data transfer or mismatched variable definitions can result in incomplete or inaccurate inputs. For clinicians, the additional burden of manually exporting or reformatting data undermines workflow efficiency and reduces the likelihood of adoption.

Successful implementation requires user-centred design, where AI outputs are delivered at the point of decision-making rather than as separate analytical reports. Clinical dashboards that summarise predictions, confidence intervals, and relevant biomarkers in intuitive visual formats can improve usability and encourage adoption. Furthermore, decision support systems must remain advisory rather than prescriptive, allowing oncologists to maintain final authority over treatment decisions. Maintaining this balance is critical for preserving both clinical autonomy and patient safety.

7.2. Data security and patient privacy

The ethical management of patient data represents another critical dimension of Immuno-AI implementation. The predictive power of AI systems depends on access to large and diverse datasets, yet such data sharing raises legitimate concerns about privacy and consent. Compliance with data protection regulations such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States is mandatory but often difficult to achieve across borders.

Emerging strategies such as federated learning offer promising solutions. Federated frameworks allow institutions to train shared AI models without exchanging raw data, thereby maintaining patient privacy while pooling model insights across diverse populations [57]. For instance, collaborative projects between European cancer centres have demonstrated that federated approaches can achieve comparable accuracy to centralised models while maintaining strict data security [58].

However, the technical and legal challenges remain significant. Federated systems require consistent data schema and encryption standards, and network latency or institutional firewalls can disrupt training synchronisation. Furthermore, ensuring that privacy-preserving mechanisms do not compromise model performance is a continuing research priority. Addressing these challenges will be essential for the sustainable scaling of Immuno-AI across international healthcare networks.

7.3. Economic and infrastructural challenges

The deployment of AI in oncology is capital-intensive, requiring substantial investment in computational infrastructure, cloud storage, and staff training. These costs create disparities between well-resourced academic centres and smaller community hospitals. While the initial development of Immuno-AI systems is typically supported by research grants, long-term maintenance and software updates require sustainable funding models.

Economic evaluation is therefore necessary to justify adoption. Cost-effectiveness analyses should quantify not only direct financial savings, such as reduced diagnostic turnaround times, but also indirect benefits like improved treatment outcomes and reduced adverse events. In many cases, the high upfront costs of implementation are offset by downstream savings from improved patient selection for expensive immunotherapies. For example, if a predictive model accurately identifies non-responders to checkpoint inhibitors, it can prevent unnecessary treatment costs and toxicity management.

Nevertheless, demonstrating such economic benefits requires robust real-world evidence. To date, most Immuno-AI studies have been conducted retrospectively, with limited data on resource utilisation or long-term outcomes. Without prospective economic evaluation, health systems may hesitate to invest in AI deployment despite promising technical results. Policymakers and funding agencies should therefore prioritise implementation research that integrates health economics into AI validation pipelines.

7.4. Education, literacy and Cultural adoption

Clinician acceptance of AI tools depends strongly on digital literacy and organisational culture. While many oncologists recognise the potential of AI to improve precision medicine, studies show persistent scepticism regarding its reliability and medico-legal implications [59]. Training programs must therefore extend beyond technical instruction to include ethical reasoning, bias awareness, and interpretive skills.

Professional societies such as the European Society for Medical Oncology (ESMO) and the American Society of Clinical Oncology (ASCO) have begun expanding AI-focused educational resources and professional-development initiatives for oncology clinicians. ESMO now maintains an AI & Digital Oncology Hub and a dedicated AI & Digital Oncology Congress, while ASCO has launched an AI in Oncology initiative and associated educational content for oncology professionals [60,61]. These programs emphasise the interpretation of AI-derived probabilities, understanding of model limitations, and the importance of clinician oversight. Hospitals implementing Immuno-AI should also establish cross-disciplinary “AI governance committees” to oversee model performance, manage updates, and provide transparent feedback channels.

Cultural transformation is equally important. For AI to gain acceptance, clinicians must perceive it as an aid rather than a threat to professional judgement. Early studies suggest that co-development and co-validation approaches, where clinicians are actively involved in model design, significantly enhance adoption rates [62]. Involving frontline users in design processes not only improves usability but also fosters trust by aligning algorithmic logic with clinical reasoning.

7.5. Interoperability and systems integration

True clinical utility requires that AI models function as interoperable components within broader healthcare ecosystems. The lack of common data standards and interoperability frameworks remains a persistent obstacle. Variability in file formats, nomenclature, and metadata conventions impedes the seamless transfer of data between laboratories, imaging departments, and clinical AI systems.

Recent initiatives, such as the Fast Healthcare Interoperability Resources (FHIR) protocol and the Global Alliance for Genomics and Health (GA4GH) standards, provide a foundation for addressing these challenges [33]. FHIR enables structured data exchange between EMRs and external applications, while GA4GH promotes harmonisation of genomic data sharing. Adoption of these frameworks, however, remains inconsistent, and many institutions still operate in technological silos.

To achieve genuine systems integration, national health authorities must incentivise adherence to interoperability standards and support investment in infrastructure upgrades. Furthermore, vendors developing Immuno-AI software should adopt open APIs and modular architectures that allow interoperability across platforms. Interconnected systems will not only enhance scalability but also enable the aggregation of larger, more representative datasets essential for model retraining and bias reduction.

8. Future Directions: Towards a clinically trustworthy Immuno-AI

The future of Immuno-AI will be defined not by the sophistication of algorithms but by their ability to demonstrate consistent, transparent, and equitable clinical benefit. To move beyond the current translational bottleneck, future development must prioritise reproducibility, standardisation, clinician engagement, and ethical accountability. The following subsections outline key strategic directions that can accelerate the evolution of Immuno-AI from laboratory prototypes to clinically trustworthy decision-support systems.

8.1. Multi-Institutional and prospective validation frameworks

One of the strongest predictors of translational success is rigorous validation across independent, diverse populations. Most Immuno-AI studies remain retrospective, using convenience datasets that limit generalisability. Prospective, multi-institutional trials are essential to assess real-world performance, model calibration, and clinical impact. These trials should adhere to established guidelines such as CONSORT-AI and SPIRIT-AI, ensuring transparent reporting of data handling, patient inclusion, and endpoint definitions [51].

Collaborative validation consortia can accelerate this process. Examples such as the TCIA and the National Health Service (NHS) AI lab programme have demonstrated that cross-institutional data pooling enhances both statistical robustness and public confidence [53,63]. Importantly, future frameworks must integrate continuous post-deployment monitoring to identify performance drift, ensuring models remain reliable as population characteristics and clinical practices evolve.

8.2. Adaptive and continually learning systems

Traditional AI models in oncology are static, trained once and rarely updated. However, clinical data are dynamic. Treatment regimens, biomarkers, and population genetics evolve over time. Adaptive or continually learning systems in form of digital twins that update as new data are introduced, offer a promising solution to this limitation [64–66].

Such systems require strong governance to balance flexibility with safety. Continuous learning must occur within predefined regulatory boundaries, using locked base models that are retrained only after pre-specified review intervals. Regulatory bodies including the FDA have begun exploring "Predetermined Change Control Plans," allowing developers to propose clear procedures for safe algorithmic evolution [30].

In Immuno-AI, adaptive frameworks could enable ongoing refinement of response predictors as new biomarkers are discovered or as treatment combinations evolve. For example, adaptive models could incorporate novel immune signatures or spatial transcriptomic data to improve response prediction for next-generation checkpoint inhibitors. The clinical relevance of spatial transcriptomics integration is substantiated by recent methodological advances. SpaIM [18], a style-transfer learning model evaluated across 53 spatial transcriptomics datasets, demonstrated that imputing unmeasured gene expressions from single-cell RNA sequencing data substantially enhances downstream biological analyses, including differential gene expression profiling and ligand–receptor interaction mapping. When applied to lung tumour tissues profiled using NanoString CosMx, SpaIM recovered 92 lymphocyte-specific differentially expressed genes compared with 59 in raw data, while identifying biologically significant pairs such as VEGFA–ITGB1 and LTF–TFRC that are exclusively detectable in imputed data. These findings suggest that spatial multi-omic enrichment pipelines could broaden the biomarker space available to future Immuno-AI systems, particularly as spatial-omics methods become more standardized, scalable, and clinically translatable in oncology. The challenge lies in developing mechanisms for auditing and versioning adaptive models to ensure reproducibility and traceability across clinical updates.

8.3. Digital twins and Real-Time immunotherapy monitoring

The convergence of AI, bioinformatics, and systems biology has enabled the emergence of "digital twins", a computational replica of individual patients that simulate biological and treatment responses in silico. Digital twins can integrate multi-omic data, including imaging, genomics, proteomics, and clinical history, to model treatment outcomes dynamically [64,65,67].

In immuno-oncology, this approach could enable personalised treatment optimisation, allowing oncologists to test hypothetical scenarios, such as dosing modifications or combination therapies, before applying them clinically. Pilot studies have shown that digital twin models can predict tumour progression and immune response trajectories with encouraging accuracy [68].

However, implementing digital twin frameworks for immunotherapy requires large, standardised datasets and computational resources that few institutions currently possess. Ethical challenges also arise, particularly regarding consent, data ownership, and the interpretability of simulated predictions. Despite these hurdles, digital twins represent a frontier for integrating AI-driven prediction into continuous, patient-specific care.

8.4. Ethics-Embedded design and fairness auditing

Ethical design must become a structural component of Immuno-AI development rather than an afterthought. Algorithmic fairness requires that predictive performance be consistent across demographic subgroups, including race, sex, socioeconomic status, and tumour subtype. Bias audits using stratified performance metrics and counterfactual testing can identify imbalances early in the design process [69].

In addition, transparency in data provenance and model documentation through frameworks such as model cards and datasheets for datasets can facilitate accountability [70]. These tools provide standardised documentation of data sources, intended use cases, and known limitations. Incorporating ethics at the design stage also involves establishing oversight committees with multidisciplinary representation, including clinicians, ethicists, data scientists, and patient advocates.

Furthermore, future models should prioritise explainability by design, embedding interpretability directly into algorithmic architecture rather than relying on post hoc explanations. Integrating ethical review into the technical workflow will not only prevent harm but also enhance credibility among both clinicians and patients.

8.5. Regulatory-Grade AI: Harmonising global standards

The global regulation of AI-based medical tools remains fragmented, with varying requirements across jurisdictions. Harmonising

international standards is crucial for scaling Immuno-AI systems across healthcare systems. Initiatives such as the International Medical Device Regulators Forum (IMDRF) are promoting common terminology and assessment pathways for SaMD [71].

In oncology, regulatory-grade AI must meet not only performance benchmarks but also demonstrate transparent lifecycle management. Developers should be required to submit documentation detailing data sources, validation procedures, and post-market surveillance strategies. Additionally, real-time registries for AI models, such as analogous to clinical trial registries could increase accountability and public trust by allowing stakeholders to track approved algorithms and updates.

A harmonised regulatory framework would also reduce duplication of effort, enabling more efficient global collaboration. As Immuno-AI continues to expand, alignment between the FDA, EMA, and MHRA on principles of validation, transparency, and adaptive monitoring will be essential for enabling safe and scalable clinical translation.

8.6. The Road Ahead: A systems View of clinical trust

The translation of Immuno-AI into practice should be conceptualised as a systems problem rather than a technological one. Building clinical trust requires simultaneous progress across multiple dimensions, such as scientific, regulatory, ethical, and organisational. Models must be transparent enough to be trusted, adaptable enough to remain relevant, and rigorously validated enough to ensure safety.

Collaborative initiatives that integrate academia, healthcare providers, and industry will be key to this transformation. The future of Immuno-AI lies in ecosystem thinking creating interconnected infrastructures that facilitate continual validation, open data sharing, and equitable access. Only through sustained interdisciplinary collaboration can Immuno-AI transition from algorithmic accuracy to enduring clinical trust.

9. Conclusion

The rapid expansion of AI in immuno-oncology has generated significant enthusiasm, but clinical adoption remains limited. The discrepancy between algorithmic accuracy and practical applicability defines the core translational challenge of Immuno-AI. Despite models achieving high AUC and predictive precision, the absence of multicentre validation, interpretability, and regulatory clarity has constrained clinical deployment. Accuracy alone does not confer trust; reproducibility, transparency, and ethical integrity are the true determinants of clinical readiness.

The critical analysis presented in this review highlights three recurring themes that define the translational gap. First, methodological issues, particularly single-institution training, lack of external validation, and inconsistent data preprocessing undermine reproducibility. Second, cultural and institutional barriers, including limited clinician involvement and inadequate AI literacy, hinder confidence and acceptance. Third, regulatory and ethical frameworks have not yet evolved to accommodate adaptive algorithms capable of continuous learning. Collectively, these factors reinforce the conclusion that technological capability has outpaced implementation infrastructure.

The path forward requires a fundamental shift in priorities. The next generation of Immuno-AI must be designed not only for accuracy but for accountability. Models should undergo prospective validation across multiple healthcare systems, using harmonised datasets compliant with international standards such as GA4GH and FHIR. Transparency in data sources, feature selection, and decision logic must become mandatory to enable replication and regulatory oversight. Explainability must transition from a desirable attribute to a clinical requirement, supported by interpretable architectures that align algorithmic reasoning with established clinical knowledge.

Clinician engagement must also move from consultation to co-development. Human-in-the-loop frameworks, combined with cross-disciplinary governance committees, can ensure that algorithmic recommendations are clinically relevant, ethically sound, and continuously refined through feedback. Simultaneously, regulatory bodies should establish dynamic approval mechanisms capable of monitoring adaptive models while safeguarding patient safety.

Ultimately, achieving clinical trust in Immuno-AI will depend on the ability of researchers, clinicians, regulators, and patients to collaborate around shared standards of transparency, fairness, and interpretability. The translation of Immuno-AI is not merely a technical task but a societal commitment to responsible innovation. Only through this collective effort can the field transition from isolated demonstrations of algorithmic success to widespread, equitable, and trustworthy clinical impact.

CRediT authorship contribution statement

Emmanuel O. Oisakede: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Raphael Igbarumah Ayo Daniel:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Olabanke Florence Olawuyi:** Writing – review & editing, Methodology, Investigation. **John Oluwatosin Alabi:** Writing – original draft, Methodology, Investigation. **Claret Chinenyenwa Analikwu:** Writing – review & editing, Methodology, Investigation. **David B. Olawade:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation.

References

[1] A.S. Zamani, A. Motwakel Eltayeb, A. Alluhayb, Md. Akhtar, Mobin., Ayub, R., Abdelrahim, M.A.A., Mohamed, S.S.I., Ahmad, N., Application of machine learning in predicting cancer complications using longitudinal Data: a systematic review and Meta-Analysis, Int. J. Med. Inf. 208 (2026) 106217, https://doi.org/10.1016/j.ijmedinf.2025.106217.

[2] A. Yahya, P. Lobelo, A. Eram, S.A. Hussain, I.A. Badruddin, L. Liza, D.O. Albina, Predicting adverse drug reactions in oncology: a critical review of machine learning approaches and future directions, Results Eng. 27 (2025) 106002, https://doi.org/10.1016/j.rineng.2025.106002.

[3] S. Macheka, P.Y. Ng, O. Ginsburg, A. Hope, R. Sullivan, A. Aggarwal, Prospective evaluation of artificial intelligence (AI) applications for use in cancer pathways following diagnosis: a systematic review, BMJ Oncol. 3 (1) (2024) e000255, https://doi.org/10.1136/bmjonc-2023-000255.

[4] E.O. Oisakede, O. Akinro, O.J. Bello, C.C. Analikwu, E. Egbon, D.B. Olawade, Predictive Models for Checkpoint Inhibitor Response in Cancer: a Review of Current Approaches and Future Directions, Crit. Rev. Oncol. Hematol. 104980–104980 (2025), https://doi.org/10.1016/j.critrevonc.2025.104980.

[5] J. Li, Q. Guo, X. Tan, Multi-modal feature integration for thyroid nodule prediction: Combining clinical data with ultrasound-based deep features, J. Radiat. Res. Appl. Sci. 18 (2025) 101217, https://doi.org/10.1016/j.jrras.2024.101217.

[6] Goyal, M., Marotti, J.D., Workman, A.A., Tooker, G.M., Ramin, S.K., Kuhn, E.P., Chamberlin, M.D., diFlorio-Alexander, R.M., Saeed Hassanpour, 2024. A multi-model approach integrating whole-slide imaging and clinicopathologic features to predict breast cancer recurrence risk. npj Breast Cancer 10. doi: 10.1038/s41523-024-00700-z.

[7] W. Liu, Z. Feng, M. Zhang, R. Mao, J. Li, Predicting neoadjuvant immunotherapy efficacy with machine learning models in non-small cell lung cancer: a systematic review and meta analysis, Int. J. Med. Inf. 212 (2026) 106345, https://doi.org/10.1016/j.ijmedinf.2026.106345.

[8] K. Hiwase, P. Verma, N. Zade, P. Kumar, I.N. Weerarathna, S. Gundewar, A Review on the applications and Implications of Artificial Intelligence and Machine Learning in Oncology, International Journal of Computational Intelligence Systems 19 (2026), https://doi.org/10.1007/s44196-026-01164-8.

[9] E.U. Alum, C.K. Egwu, V.S. Manjula, P.O. Ekpang, J.E. Ekpang, D.A. Echegu, B. N. Alum, D.E. Uti, Overcoming the Black Box Challenge: Building Trust in Artificial Intelligence Algorithms in Oncology, Technol. Cancer Res. Treat. 25 (2026), https://doi.org/10.1177/15330338261434649.

[10] J.Y. Cheng, J.T. Abel, U.G.J. Balis, D.S. McClintock, L. Pantanowitz, Challenges in the Development, Deployment, and Regulation of Artificial Intelligence in Anatomic Pathology, Am. J. Pathol. 191 (2021) 1684–1692, https://doi.org/10.1016/j.ajpath.2020.10.018.

[11] J.E.M. Swillens, I.D. Nagtegaal, S. Engels, A. Lugli, R.P.M.G. Hermens, J.A.W. M. van der Laak, Pathologists' first opinions on barriers and facilitators of computational pathology adoption in oncological pathology: an international study, Oncogene 42 (2023) 2816–2827, https://doi.org/10.1038/s41388-023-02797-1.

[12] A.A. Davis, V.G. Patel, The role of PD-L1 expression as a predictive biomarker: an analysis of all US food and drug administration (FDA) approvals of immune

checkpoint inhibitors, J. Immunother. Cancer 7 (1) (2019) 278, https://doi.org/10.1186/s40425-019-0768-9.
[13] T.G. Chang, Y. Cao, H.J. Sfreddo, S.R. Dhruba, S.H. Lee, C. Valero, S.K. Yoo, D. Chowell, L.G. Morris, E. Ruppin, LORIS robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features, Nat. Cancer 5 (8) (2024) 1158–1175, https://doi.org/10.1038/s43018-024-00772-7.
[14] S.K. Yoo, C.W. Fitzgerald, B.A. Cho, B.G. Fitzgerald, C. Han, E.S. Koh, A. Pandey, H. Sfreddo, F. Crowley, M.R. Korostin, N. Debnath, Prediction of checkpoint inhibitor immunotherapy efficacy for cancer using routine blood tests and clinical data, Nat. Med. 31 (3) (2025) 869–880, https://doi.org/10.1038/s41591-024-03398-5.
[15] A. Budhkar, Q. Song, J. Su, X. Zhang, Demystifying the black box: a survey on explainable artificial intelligence (XAI) in bioinformatics, Comput. Struct. Biotechnol. J. 27 (2025) 346–359.
[16] B. Zhang, Z. Wan, Y. Luo, X. Zhao, J. Samayoa, W. Zhao, S. Wu, Multimodal integration strategies for clinical application in oncology, Front. Pharmacol. 16 (2025) 1609079, https://doi.org/10.3389/fphar.2025.1609079.
[17] A. Budhkar, Z. Tang, X. Liu, X. Zhang, J. Su, Q. Song, xSiGra: explainable model for single-cell spatial data elucidation, Brief. Bioinform. 25 (5) (2024) bbae388, https://doi.org/10.1093/bib/bbae388.
[18] B. Li, Z. Tang, A. Budhkar, X. Liu, T. Zhang, B. Yang, J. Su, Q. Song, SpaIM: single-cell spatial transcriptomics imputation via style transfer, Nat. Commun. 16 (2025) 7861, https://doi.org/10.1038/s41467-025-63185-9.
[19] Z. Wang, X. Liu, K. Han, L. Lei, C. Shi, W. Liu, Q. Guo, Multimodal deep learning for immunotherapy response prediction and biomarker discovery in non-small cell lung cancer, J. Am. Med. Inform. Assoc. 32 (2025), https://doi.org/10.1093/jamia/ocaf142.
[20] Y. Li, J. Xia, T. He, Y. Hu, D. Zhou, D. Zou, B. Li, M. Zhang, Z. Huang, M. Zhang, X. Liu, M. Wang, H. Luo, F. Lu, C. Zhang, X. Zhao, S. Su, J. Peng, Development and validation of an interpretable machine learning model for predicting progression-free survival after immunotherapy in patients with non-small cell lung cancer: a multicenter study, Front. Immunol. 16 (2025), https://doi.org/10.3389/fimmu.2025.1686260.
[21] M. Rakaee, M. Tafavvoghi, B. Ricciuti, J.V. Alessi, A. Cortellini, F. Citarella, L. Nibid, G. Perrone, E. Adib, C.A.M. Fulgenzi, M.H. Filho, C., Di Federico, A., Jabar, F., Hashemi, S., Houda, I., Richardsen, E., Rasmussen Busund, L.-T., Donnem, T., Bahce, I., Pinato, D.J., Deep Learning Model for predicting Immunotherapy Response in Advanced Non-Small Cell Lung Cancer, JAMA Oncol. 11 (2024), https://doi.org/10.1001/jamaoncol.2024.5356.
[22] Fa'ak, F., Coudray, N., Jour, G., Ibrahim, M., Illa-Bochaca, I., Qiu, S., Claudio Quiros, A., Yuan, K., Johnson, D.B., Rimm, D.L., Weber, J.S., Tsirigos, A., Osman, I., Artificial Intelligence Algorithm Predicts Response to Immune Checkpoint Inhibitors, Clin. Cancer Res. 31 (2025) 3526–3536, https://doi.org/10.1158/1078-0432.ccr-24-3720.
[23] L. Provenzano, M. Favali, L. Mazzeo, A. Spagnoletti, M. Ruggirello, G. Calareso, F. G. Greco, R. Vigorito, A. Quarta, F. Calimeri, M. Monteleone, G. Baselli, E. De Momi, B. Guirges, A. Di Lello, A. Zec, A. Ferrarin, C. Giani, C. Silvestri, M. Occhipinti, Integrating radiomics and real-world data to predict immune checkpoint inhibitor efficacy in advanced non-small-cell lung cancer, ESMO Real World Data and Digital Oncology 10 (2025) 100182, https://doi.org/10.1016/j.esmorw.2025.100182.
[24] Vanguri, R.S., Luo, J., Aukerman, A., Egger, J.V., Fong, C.J., Horvat, N., Pagano, A., Araujo-Filho, J.J. de A.B., Geneslaw, L., Rizvi, H., Sosa, R.E., Boehm, K.M., Yang, S.-R., Bodd, F.M., Ventura, K., Hollmann, T.J., Ginsberg, M.S., Gao, J., MSK MIND Consortium, Hellmann, M.D., Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer, Nature Cancer 3 (2022) 1151–1164, https://doi.org/10.1038/s43018-022-00416-8.
[25] S. Iivanainen, J. Ekstrom, H. Virtanen, V.V. Kataja, J.P. Koivunen, Electronic patient-reported outcomes and machine learning in predicting immune-related adverse events of immune checkpoint inhibitor therapies, BMC Med. Inf. Decis. Making 21 (2021), https://doi.org/10.1186/s12911-021-01564-0.
[26] S.H. Park, K. Han, Methodologic Guide for evaluating Clinical Performance and effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction, Radiology 286 (2018) 800–809, https://doi.org/10.1148/radiol.2017171920.
[27] C.H. Cheng, S. Shi, Artificial intelligence in cancer: applications, challenges, and future perspectives, Mol. Cancer 24 (2025), https://doi.org/10.1186/s12943-025-02450-3.
[28] D.E. Mathew, D.U. Ebem, A.C. Ikegwu, P.E. Ukeoma, N.F. Dibiaezue, Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human, Neural Process. Lett. 57 (1) (2025) 16, https://doi.org/10.1007/s11063-025-11732-2.
[29] T. Kotsiopoulos, G. Papakostas, T. Vafeiadis, V. Dimitriadis, A. Nizamis, A. Bolzoni, D. Bellinati, D. Ioannidis, K. Votis, D. Tzovaras, P. Sarigiannidis, Revolutionizing defect recognition in hard metal industry through AI explainability, human-in-the-loop approaches and cognitive mechanisms, Expert Syst. Appl. 124839–124839 (2024), https://doi.org/10.1016/j.eswa.2024.124839.
[30] FDA, 2025. Artificial Intelligence in Software [WWW Document]. U.S. Food and Drug Administration. URL https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device. [Accessed 1 November 2025].
[31] European Medicines Agency (EMA). *Review of AI/ML Applications in Medicines Lifecycle (2024) Horizon Scanning Short Report*; 2025. https://www.ema.europa.eu/en/documents/report/review-artificial-intelligence-machine-learning-applications-medicines-lifecycle-2024-horizon-scanning-short-report_en.pdf [Accessed 3 November 2025].
[32] J.E. Alderman, J. Palmer, E. Laws, M.D. McCradden, J. Ordish, M. Ghassemi, S. R. Pfohl, N. Rostamzadeh, H. Cole-Lewis, B. Glocker, M. Calvert, T.J. Pollard, J. Gill, J. Gath, A. Adebajo, J. Beng, C.H. Leung, S. Kuku, L.-A. Farmer, R.N. Matin, Tackling algorithmic bias and promoting transparency in health datasets: the STANDING together consensus recommendations, The Lancet Digital Health 7 (2024), https://doi.org/10.1016/s2589-7500(24)00224-3.
[33] Global Alliance for Genomics and Health, 2025. Framework for responsible sharing of genomic and health-related data. Available at: https://www.ga4gh.org/framework/ [Accessed 3 November 2025].
[34] Olawade, D.B., Ojo, I.O., Oisakede, E.O., Joel-Medewase, V.I., Wada, O.Z., 2025. Artificial Intelligence in Nigerian Oncology Practice: A Qualitative Exploration of Oncologists' Perspectives. Journal of Cancer Policy 45, 100626–100626. .
[35] Lambert, S.I., Madi, M., Sopka, S., Lenes, A., Stange, H., Buszello, C.-P., Stephan, A., 2023. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. npj Digital Medicine 6. doi: 10.1038/s41746-023-00852-5.
[36] G.S. Collins, M. Chester-Jones, S. Gerry, J. Ma, J. Matos, J. Sehjal, B. Tsegaye, P. Dhiman, Clinical prediction models using machine learning in oncology: challenges and recommendations, BMJ Oncology 4 (2025) e000914–e, https://doi.org/10.1136/bmjonc-2025-000914.
[37] Y. She, B. He, F. Wang, Y. Zhong, T. Wang, Z. Liu, M. Yang, B. Yu, J. Deng, X. Sun, C. Wu, L. Hou, Y. Zhu, Y. Yang, H. Hu, D. Dong, C. Chen, J. Tian, Deep Learning for Predicting Major Pathological Response to Neoadjuvant Chemoimmunotherapy in Non-Small Cell Lung Cancer: A Multicentre Study. eBioMedicine 86 (2022) 104364, https://doi.org/10.1016/j.ebiom.2022.104364.
[38] V.P. Veeraraghavan, S. Daniel, R. Manyam, A. Reddy, S.R. Patil, Development and validation of a prediction model for risk stratification and outcome prediction in oral oncology patients, Oral Oncology Reports 13 (2025) 100728, https://doi.org/10.1016/j.oor.2025.100728.
[39] J. Mitra, S. Ghose, R. Thawani, Clinically Explainable Prediction of Immunotherapy Response Integrating Radiomics and Clinico-Pathological Information in Non-Small Cell Lung Cancer, Cancers 17 (2025) 2679, https://doi.org/10.3390/cancers17162679.
[40] M. Craddock, C. Crockett, A. McWilliam, G. Price, M. Sperrin, S.N. van der Veer, C. Faivre-Finn, Evaluation of Prognostic and Predictive Models in the Oncology Clinic, Clin. Oncol. 34 (2022) 102–113, https://doi.org/10.1016/j.clon.2021.11.022.
[41] Yuan, X., Xu, H., Zhu, J., Yang, Z., Pan, B., Wu, L., Chen, H., 2025. Systematic review and meta-analysis of artificial intelligence for image-based lung cancer classification and prognostic evaluation. npj Precision Oncology 9. doi: 10.1038/s41698-025-01095-1.
[42] C.S. Santos, M. Amorim-Lopes, Externally validated and clinically useful machine learning algorithms to support patient-related decision-making in oncology: a scoping review, BMC Med. Res. Method. 25 (2025), https://doi.org/10.1186/s12874-025-02463-y.
[43] E.-M. Froicu, I. Creangă-Murariu, V.-A. Afrăsânie, B. Gafton, T. Alexa-Stratulat, L. Miron, D.M. Pușcașu, V. Poroch, G. Bacoanu, I. Radu, M.-V. Marinca, Artificial Intelligence and Decision-making in Oncology: a Review of Ethical, Legal, and Informed Consent challenges, Curr. Oncol. Rep. (2025), https://doi.org/10.1007/s11912-02501698-8.
[44] European Cancer Organisation, 2025. Harnessing AI for Cancer Care in Europe [WWW Document]. European Cancer Organisation. URL https://www.europeancancer.org/resources/publications/harnessing-ai-for-cancer-care-in-europe.html (accessed 4.16.26).
[45] Z. Sadeghi, R. Alizadehsani, M.A. Cifci, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R.S. Alkhawaldeh, S. Hussain, B. Alatas, A review of explainable artificial intelligence in healthcare, Comput. Electr. Eng. 118 (2024) 109370, https://doi.org/10.1016/j.compeleceng.2024.109370.
[46] S. Choi, S.I. Cho, W. Jung, T. Lee, S.J. Choi, S. Song, G. Park, S. Park, M. Ma, S. Pereira, D. Yoo, Deep learning model improves tumor-infiltrating lymphocyte evaluation and therapeutic response prediction in breast cancer, npj Breast Cancer 9 (1) (2023) 71, https://doi.org/10.1038/s41523-023-00577-4.
[47] Amaliah, N.R., Tjahjono, B., Vasile Palade, 2025. Human-in-the-Loop XAI for Predictive Maintenance: A Systematic Review of Interactive Systems and Their Effectiveness in Maintenance Decision-Making. Electronics 14, 3384–3384. doi: 10.3390/electronics14173384.
[48] H. Kim, S. Kim, S. Choi, C. Park, S. Park, S. Pereira, M. Ma, D. Yoo, K. Paeng, W. Jung, S. Park, Clinical validation of artificial intelligence–powered PD-L1 tumor proportion score interpretation for immune checkpoint inhibitor response prediction in non–small cell lung cancer, JCO Precis. Oncol. 8 (2024) e2300556, https://doi.org/10.1200/po.23.00556.
[49] H. Semmelrock, T. Ross-Hellauer, S. Kopeinik, D. Theiler, A. Haberl, S. Thalmann, D. Kowald, Reproducibility in machine-learning-based research: Overview, barriers, and drivers, AI Mag. 46 (2025), https://doi.org/10.1002/aaai.70002.
[50] G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. van Smeden, A.-L. Boulesteix, J. C. Camaradou, L.A. Celi, S. Denaxas, A.K. Denniston, B. Glocker, R.M. Golub, H. Harvey, G. Heinze, M.M. Hoffman, TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ (2024) e078378–e, https://doi.org/10.1136/bmj-2023-078378.
[51] CONSORT-AI and SPIRIT-AI Steering Group, 2019. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. Nature Medicine 25, 1467–1468. doi: 10.1038/s41591-019-0603-3.

[52] D. Chen, K. Arnold, R. Sukhdeo, J. Farag Alla, S. Raman, Concordance with CONSORT-AI guidelines in reporting of randomised controlled trials investigating artificial intelligence in oncology: a systematic review, BMJ Oncology 4 (2025) e000733, https://doi.org/10.1136/bmjonc-2025-000733.
[53] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The Cancer Imaging Archive (TCIA): Maintaining and operating a Public Information Repository, J. Digit. Imaging 26 (2013) 1045–1057, https://doi.org/10.1007/s10278-013-9622-7.
[54] Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., Wong-Erasmus, M., Yao, L., Kasprzyk, A., 2011. International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. Database 2011, bar026–bar026. doi: 10.1093/database/bar026.
[55] G.M. Lucas, B. Becerik-Gerber, S.C. Roll, Calibrating workers' trust in intelligent automated systems, Patterns 101045–101045 (2024), https://doi.org/10.1016/j.patter.2024.101045.
[56] M.M. Bertagnolli, B. Anderson, A. Quina, S. Piantadosi, The electronic health record as a clinical trials tool: Opportunities and challenges, Clin. Trials 17 (2020) 237–242, https://doi.org/10.1177/1740774520913819.
[57] M.J. Sheller, B. Edwards, G.A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R.R. Colen, S. Bakas, Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Sci. Rep. 10 (2020) 12598, https://doi.org/10.1038/s41598-020-69250-1.
[58] A. Gyrard, S. Abedian, P. Gribbon, G. Manias, R. van Nuland, K. Zatloukal, I. E. Nicolae, G. Danciu, S. Nechifor, L. Marti-Bonmati, P. Mallol, S. Dalmiani, S. Autexier, M. Jendrossek, I. Avramidis, E. Garcia Alvarez, P. Holub, I. Blanquer, A. Boden, R. Hussein, Lessons Learned from European Health Data Projects with Cancer Use cases: Implementation of Health Standards and internet of things Semantic Interoperability, J. Med. Internet Res. 27 (2025) e66273, https://doi.org/10.2196/66273.
[59] A. Hantel, T.P. Walsh, J.M. Marron, K.L. Kehl, R. Sharp, E. Van Allen, G.A. Abel, Perspectives of Oncologists on the Ethical Implications of using Artificial Intelligence for Cancer Care, JAMA Netw. Open 7 (2024) e244077, https://doi.org/10.1001/jamanetworkopen.2024.4077.
[60] ESMO, 2026. Esmo.org. URL https://www.esmo.org/newsroom/esmo-ai-digital-oncology-hub/educational-resources. (accessed 4.17.26).
[61] ASCO, 2024. AI and Oncology. Asco.org. URL https://www.asco.org/news-initiatives/current-initiatives/ai-oncology. (accessed 4.17.26).
[62] J.M. Schwartz, A.J. Moy, S.C. Rossetti, N. Elhadad, K.D. Cato, Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: a scoping review, J. Am. Med. Inform. Assoc. 28 (2021) 653–663, https://doi.org/10.1093/jamia/ocaa296.
[63] Cresswell, K., Williams, R., Dungey, S., Anderson, S., Bernabeu, M.O., Hajar Mozaffar, Yang, X., Sai, V., Bea, S., Eason, S., 2025. A mixed methods formative evaluation of the United Kingdom National Health Service Artificial Intelligence Lab. PubMed 8, 448–448. doi: 10.1038/s41746-025-01805-w.
[64] D.B. Olawade, E.O. Oisakede, O.J. Bello, C.C. Analikwu, E. Egbon, A. Ojo, Digital twins in oncology: from predictive modelling to personalised treatment strategies, Crit. Rev. Oncol. Hematol. 220 (2026) 105171, https://doi.org/10.1016/j.critrevonc.2026.105171.
[65] D.B. Olawade, O. Akinro, E.O. Oisakede, O.J. Bello, C.C. Analikwu, E. Egbon, Digital twin applications in radiology and radiotherapy: applications, challenges, and future perspectives, Eur. J. Radiol. 112865 (2026), https://doi.org/10.1016/j.ejrad.2026.112865.
[66] D. Jones, C. Snider, A. Nassehi, J. Yon, B. Hicks, Characterising the Digital Twin: a systematic literature review, CIRP J. Manuf. Sci. Technol. 29 (2020) 36–52, https://doi.org/10.1016/j.cirpj.2020.02.002.
[67] A. Osipov, O. Nikolic, A. Gertych, S. Parker, A. Hendifar, P. Singh, D. Filippova, G. Dagliyan, C.R. Ferrone, L. Zheng, J.H. Moore, W. Tourtellotte, J.E. Van Eyk, D. Theodorescu, The Molecular Twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients, Nature Cancer 5 (2024) 299–314, https://doi.org/10.1038/s43018-023-00697-7.
[68] S. Shen, W. Qi, X. Liu, J. Zeng, S. Li, X. Zhu, C. Dong, B. Wang, Y. Shi, J. Yao, B. Wang, L. Jing, S. Cao, G. Liang, From virtual to reality: innovative practices of digital twins in tumor therapy, J. Transl. Med. 23 (2025), https://doi.org/10.1186/s12967-025-06371-z.
[69] G. Cornacchia, V.W. Anelli, G.M. Biancofiore, F. Narducci, C. Pomo, A. Ragone, E. Di Sciascio, Auditing fairness under unawareness through counterfactual reasoning, Inf. Process. Manag. 60 (2023) 103224, https://doi.org/10.1016/j.ipm.2022.103224.
[70] M.M. Ahmed, O.J. Okesanya, M. Oweidat, Z.K. Othman, S.S. Musa, D.E. Lucero-Prisno III, The ethics of data mining in healthcare: challenges, frameworks, and future directions, Biodata Min. 18 (2025), https://doi.org/10.1186/s13040-025-00461-w.
[71] International Medical Device Regulators Forum (IMDRF), 2025. Software as a Medical Device | International Medical Device Regulators Forum [WWW Document]. www.imdrf.org. URL https://www.imdrf.org/working-groups/software-medical-device. [Accessed 3 November 2025].