

AI in Cancer Prognosis: A Systematic Review of Multimodal Models Combining Pathology Images and High-Throughput Omics

Cancer Informatics
Volume 25: 1–23
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769351261434523
journals.sagepub.com/home/cix



Charlotte Jennings^{1,2} , Andrew Broad^{1,2} , Lucy Godson^{1,2} , Emily Clarke^{1,2} , David Westhead² and Darren Treanor^{1,2,3,4}

Abstract

This study systematically reviews and evaluates published research on machine learning models that integrate histopathology whole slide images and high-throughput -omic data to predict overall survival in cancer. A comprehensive search of PubMed, EMBASE, and Cochrane CENTRAL was conducted through August 12, 2024, with citation screening for additional studies. Eligible studies applied machine learning or deep learning methods to multimodal data combining pathology images and -omics. Data extraction followed the CHARMS checklist, and risk of bias was assessed using the PROBAST + AI tool. Narrative synthesis was conducted in line with PRISMA 2020 guidelines. Forty-eight studies published since 2017 met inclusion criteria, spanning 19 cancer types. All relied on The Cancer Genome Atlas dataset. Modelling approaches included regularised Cox regression (n=4), classical machine learning (n=13), and deep learning (n=31). Reported concordance indices ranged from 0.550 to 0.857, with most multimodal models outperforming unimodal counterparts. However, all studies were assessed as having high or unclear risk of bias—most often due to limited external validation, insufficient reporting, and minimal assessment of clinical utility. This review highlights a rapidly evolving yet methodologically underdeveloped field. While model performance is promising, improvements in data standardisation, reporting practices, and real-world contextualisation are critical for clinical translation. This work was funded by the National Pathology Imaging Cooperative (NPIC), supported by UK Research and Innovation (Project no. 104687).

Keywords

multimodal machine learning, deep learning, artificial intelligence, survival prediction, prognostic models, cancer, histopathology, whole slide images, omics, systematic review, clinical utility, reporting standards

Received: 6 August 2025; accepted: 2 March 2026

Introduction

Cancer remains a key health priority globally, with an estimated 18.1 million new cases of cancer worldwide in 2020 and a projected increase to an annual incidence of 28 million by 2040.¹ Prognostic models that accurately stratify patients by survival risk are essential for guiding treatment decisions, clinical trial design, and for health resource planning making them valuable to patients, clinicians, researchers and policy makers.^{2,3}

Routine cancer care involves the integration of multi-modal data—including clinical, radiological, pathological, and molecular inputs—to guide decision making. Digitisation of pathology slides into whole slide images (WSI) and advances in -omics now enable the computational analysis of these data, supporting precision oncology. Clinical prediction models are tools that estimate a patient's risk of an outcome by combining multiple covariates and have

traditionally relied on regression-based methods. With the increasing scale and complexity of multi-modal cancer data available, there has been increasing interest in applying machine learning (ML) and deep learning (DL) approaches to build more flexible, data-driven prognostic models.⁴

ML models have already demonstrated improved survival prediction performance in single-modality settings,

¹National Pathology Imaging Cooperative, Leeds Teaching Hospitals NHS Trust, UK

²University of Leeds, UK

³Department of Clinical Pathology, Linköping University, Sweden

⁴Centre for Medical Image Science and Visualisation, Linköping University, Sweden

Corresponding author:

Charlotte Jennings, National Pathology Imaging Cooperative, Leeds Teaching Hospitals NHS Trust, Leeds, West Yorkshire LS9 7TF, UK.
Email: charlotte.jennings1@nhs.net



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License

(<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further

permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

including WSI data,^{5,6} and multiomics.⁷ More recent studies report enhanced prognostic predictions by the integration of these data, assumed to be a result of capturing complementary biological signals. To date, clinical translation of ML approaches in other domains has been very low.⁸ There is a need to understand the maturity of this field, whether there is the potential for real-world utility and if the performance gain of these models is sufficient to justify the high costs of laboratory data generation and computation.

In related work, Schneider et al. evaluated studies integrating pathology image and genomic data by DL (searches up to June 2021).⁹ This review identified 11 studies, including 7 predictive of survival (4 cancer-specific survival, 3 overall survival) and 4 related to other tasks, including; prediction of microsatellite instability (MSI) status, malignant versus benign differentiation and subtype prediction. The multimodal approach was shown to be superior to unimodal methods in all studies, however the heterogeneity of studies limited further conclusions. A more general review by Cui et al. across the medical domain also determined that multimodal DL prediction models typically surpass unimodal models for disease diagnosis and prognosis tasks.¹⁰ This selective review evaluated 34 studies using feature-level multimodal DL-based fusion to integrate image and non-image data, and outlined emerging multimodal frameworks and fusion approaches. Data availability and lack of explainability were identified as key limitations to these studies.

This review provides the first focussed review on prognostic prediction in the cancer domain through high-throughput -omic and WSI integration. In line with best practice methodology, we systematically review all literature in which these data are combined as predictors for the prediction of overall survival (OS) in cancer patients using ML or DL techniques. We focussed on OS as a universally applicable, clinically meaningful outcome that is consistently reported across cancer types and public datasets, enabling standardised comparison across studies. While OS reflects factors beyond tumour biology, its broad availability and relevance made it the most practical and informative endpoint for this review.

This review specifically addresses the following questions;

- What is the prediction performance for OS?
- What cancer domains have these techniques been applied to?
- What types and scale of data sources are used?
- What methods are used for data processing and integration of data?
- What is the quality of these studies and their reporting?

Methods

This systematic review without meta-analysis was conducted in accordance with the guidelines for “Preferred Reporting Items for Systematic Reviews and Meta-Analyses” (PRISMA).¹¹ Due to the highly heterogeneous nature of included studies a narrative synthesis approach was used in line with the “Synthesis without meta-analysis” (SWiM) guidelines.¹² The protocol for this review is available at https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=594745 and was approved before the search results were screened for inclusion (Registration CRD42024594745).

php?RecordID=594745 and was approved before the search results were screened for inclusion (Registration CRD42024594745).

Eligibility Criteria

Primary, peer-reviewed research studies reporting the development and/or validation of a prediction model developed using ML methods and integrating pathology WSI data with high-throughput -omic data were sought. Only peer-reviewed journal or conference papers were included.

The population included human participants with any cancer diagnosis. The outcome of interest was OS predicted at any time point. Predictive models for other outcomes, including cancer-specific survival, were excluded. The search strategy was geared towards selection of typical ML or DL approaches. However, to recognise ML methodologies which build on traditional statistical models, studies using regression approaches were accepted as ML models if they were defined as such by the authors of a study—an approach used previously.^{8,13} Studies focussing solely on prognostic factor identification or lacking image and -omic data integration were excluded. Only WSIs of surgical pathology specimens with conventional haematoxylin and eosin (H&E) or immunohistochemical staining were included, and studies relating to cytology, forensic or other material were not included, nor were studies using other imaging techniques. Studies in which all models also integrated additional radiology imaging were not included.

Studies not in English were excluded. No exclusions were applied for date of publication.

Data Sources and Search Strategy

Searches were conducted in 2 research databases (EMBASE and PubMed) and 1 trial registry (Cochrane Central Register of Controlled Trials (CENTRAL)) from inception to 12th August 2024. Searches were restricted to human studies and English language.

The search strategy was composed of terms related to pathology whole slide images, -omic data, machine learning, and cancer. The final search was informed by piloting and preliminary scoping searches using alternative terms and combinations, which were appraised against known key publications in the field. For each category multiple terms were combined with the *OR* operator, before combining categories with the *AND* operator. The full search strategy is available in Supplemental material. Citation checking was also conducted (October 2024) using the selected studies from the abstract screening processes as seed references in a forward and back seed approach using *citation_chaser shiny app* as a complementary strategy to address any indexing limitations.^{14,15}

Study Selection

All studies from the literature searches were imported to the *Rayyan* software to manage the screening stages of the review across the team.¹⁶ First, duplicates were manually removed by C.J. supported by the duplicate detection tool within *Rayyan*. Studies were then screened according to

predefined algorithms. One investigator (C.J.) screened all titles and abstracts, with a second independent screen performed by either E.C. or L.G. Disagreements were resolved by discussion with the third investigator. Full text articles were screened in the same way. The screening algorithms are available in the Supplemental material.

Data Extraction

Data extraction for each study was performed independently by 2 reviewers using a predefined data extraction spreadsheet, which was adapted from a previously developed template.¹⁷ The template was designed with reference to the “Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies” (CHARMS) checklist and “Prediction model Risk Of Bias ASsessment Tool” (PROBAST) and was updated to align with PROBAST + AI on its release.^{18–20} Additional fields specific to the genomic, digital pathology and multimodal ML research domains were added by the authors, for example data processing details and fusion methods. C.J. reviewed all studies, while the second independent review was performed by either A.B. or L.G. Disagreements were resolved by discussion with the third investigator.

Study level information extracted from papers, included: study demographics, cancer domain, inclusion and exclusion criteria, dataset details and participant characteristics. Models meeting the inclusion criteria within these studies were appraised. Detailed information was extracted as below for the best performing OS prediction model in each study meeting the inclusion criteria (defined as our model of interest), with information about any additional models in the paper extracted in a summarised format. Model specific data fields included: approaches to feature generation and selection, model architecture and fusion approaches, and performance measures. Several fields were added or clarified during data extraction with the agreement of all researchers. Any changes were retroactively applied to all previously extracted studies. The final data extraction template is summarised in Supplemental Material.

Information was sought from the full-text articles, as well as Supplemental Materials where appropriate. Inferences were only made where both researchers were confident to do so and labelled as *unclear* where this was not possible. The well-known nature of The Cancer Genome Atlas (TCGA) dataset meant we were able to deduce some characteristics of the data used even when not explicitly provided, such as H&E staining and fixation characteristics. In these situations, assumed data is marked by an asterisk to help understand completeness of reporting of these fields.

Risk of Bias Assessment. The consensus developed PROBAST + AI tool was used to assess the models of interest in this study, providing structured assessment of methodological quality, risk of bias and applicability.²⁰ The tool assesses the likelihood of results being affected by the study design, conduct or analysis. Reporting quality was considered within the interpretation of these assessments; however, formal application of reporting checklists (eg, TRIPOD-AI) was outside the scope of this review. The

model development and evaluation process are assessed by signalling questions across 4 domains (participants, predictors, outcome and analysis), which probe issues including appropriate cohort selection, predictor definition and measurement, outcome clarity, handling of missing data, overfitting and validation methods. Through these questions reviewers are guided towards judging whether each domain is of high, low or unclear concern for quality and risk of bias. Unclear ratings are used where there is not enough information to make a full assessment. The overall quality and risk of bias ratings are determined by the worst domain score. An overall low-concern rating requires low scores in every domain, whereas a high-concern rating in a single domain will lead to high-concern overall.

The PROBAST + AI tool also assesses the applicability of prediction models to the specific criteria of the systematic review across the predictor, participant and outcome domains. Studies are also determined to be of low concern or high concern, or unclear where there is not enough information to make a full assessment.

Owing to the potential subjective interpretation of the signalling questions in the PROBAST + AI tool, 2 independent researchers completed this process for each model, with disagreements resolved as previous by the third. Each paper was assessed by a pathologist and a computer scientist.

Data Synthesis

The results of the literature search and screening process were summarised in a PRISMA flow chart generated by *PRISMA Flow Diagram Shiny app*.²¹ All extracted data were summarised in 2 tables, covering study level and model characteristics.

The data synthesis did not include any meta-analysis due to the diversity of the methods used and was designed with reference to the SWiM protocol.¹² For model characteristics, studies were grouped by the modelling method used and further ordered by year of publication to enable appreciation of any method shift over time. The concordance-index (c-index) was chosen as the comparator metric between studies because it was anticipated to be the most consistently reported evaluation metric based on preliminary searches. Two graphs were generated using the c-index, where it was available, to show (a) comparison of performance between the multimodal model of interest and unimodal model performances explored in the study, and (b) performance variation across different cancer types. Area under the curve (AUC) and time-specific AUC values were also extracted where available. Summary metrics used for calibration, overall performance and clinical utility were also tabulated.

The results of the PROBAST + AI assessments were tabulated by study and presented as summary graphs for quality, risk of bias and applicability.

Results

Literature searches returned a total of 457 studies, of which 132 were duplicates. Two hundred eighty-four records were excluded during abstract and title screening and 21 were

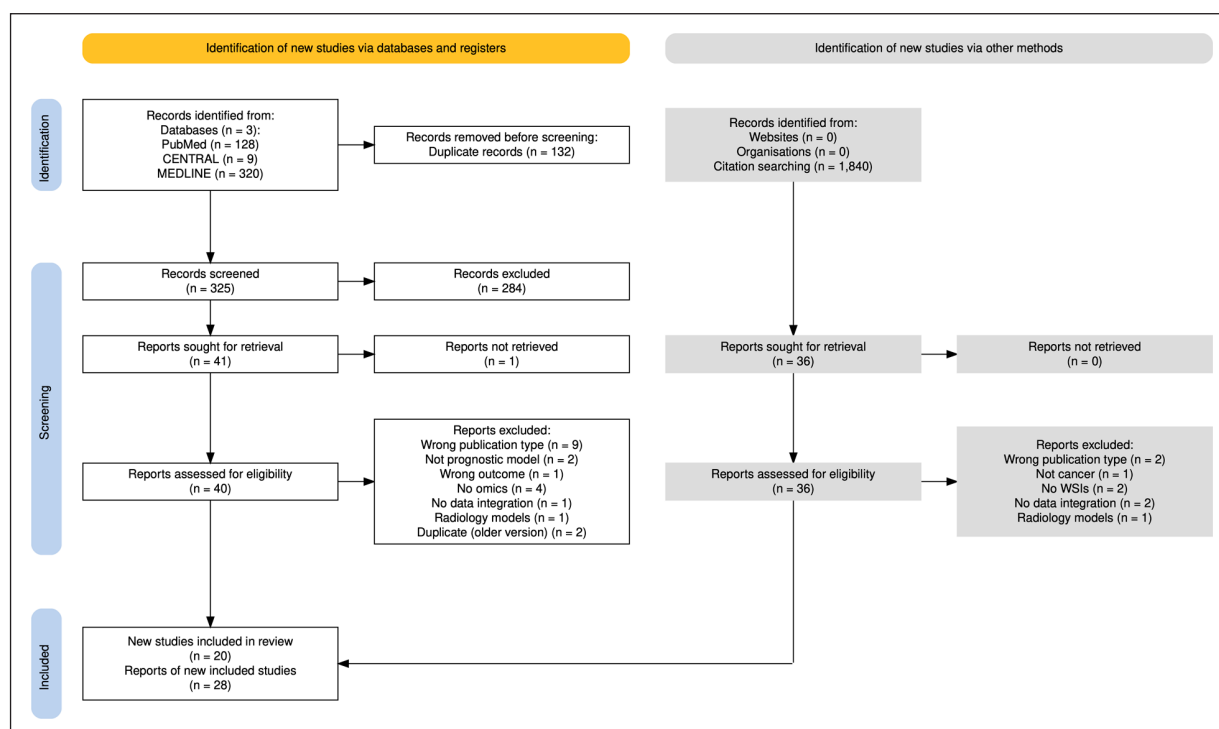


Figure 1. PRISMA 2020 flow diagram.

PRISMA 2020 flowchart of the study identification and selection process for the systematic review. Records were screened on titles and abstracts, and reports were assessed based on the full-text content; CENTRAL register of controlled trials.

excluded by the full text screening process. An additional 1840 titles and abstracts were screened as part of citation searches, of which 36 new studies were screened in full text and 29 included in the review Figure 1. The study characteristics are shown in Table 1. Details of the models evaluated are shown in Table 2.

Of the 48 included studies,^{22–69} 10 of these were conference papers with the remaining 38 being journal papers. All included studies were published since 2017. Studies appeared in 31 different publications, mainly in computer science and computational biology focussed journals (n=22), but also in scientific, medical and medical imaging domains. Eight of the studies were in cancer specific journals. Lead authors were distributed across 5 countries; China (n=33), USA (n=10), Germany (n=2), Hong Kong (n=2) and Spain (n=1).

Risk of Bias Assessments

The results of the PROBAST + AI assessments for each model of interest are shown in Table 3 and risk of bias and applicability across the review are summarised in Figure 2. Due to the limited use of external datasets for validation, assessments across the participant, predictor and outcome domains for quality of model development and risk of bias of evaluation are usually the same.

Participant. All studies had an unclear (n=47) or high concern (n=1) of bias in the participant domain. This was predominantly due to a lack of information provided about the TCGA datasets used, including recruitment methods, setting, and inclusion and exclusion criteria, reflecting limitations in

reporting at the level of the prediction model studies rather than deficiencies in the underlying source datasets. Only a couple of studies included additional detail, such as recruitment dates²⁴ and number of sites.³³ Where additional datasets were used, a greater level of information was usually provided, but this still provided an incomplete picture of the source of data in all studies.^{22,39,42,48,59,65,67}

Individual study criteria for inclusion or exclusion, where stated, were usually data based; defined by presence in a given TCGA data subset, by data quality and/or by completeness of matched data across modalities. Only 2 studies had clinically orientated inclusion criteria.^{46,54} Presentation of participant characteristics was poor, with 18 studies (38%) providing no description of the cohort used, see Table 1. The remaining studies presented variable coverage of participant demographics, disease and treatment information. In 1 study, this was presented comprehensively enough to assess the representativeness of the dataset.⁵⁴ In this study of oral squamous cell carcinoma, participant demographics were broadly representative of the USA population,⁷⁰ however were skewed towards a population with a higher pathologic stage of disease.

Predictor. Studies generally provided no information about data acquisition processes making it difficult to assess similarity of predictor assessment. Within the TCGA dataset, there are known to be differences in protocols and processes between contributing institutions which are sufficient to cause batch effect in both sequencing and imaging data.^{71–75} Whilst the variation was not well described, a few of the papers outlined pre-processing steps in one or both modalities designed to address these issues. Only one study

Table 1. Characteristics of the Studies Included in the Systematic Review.

Author, year	Internal data					Ext. data		Participant characteristics provided in study:						
	Location	Model of interest	Cancer types	Name	Origin	Name (origin)	WSI data	Omic data	Clinical data	Participant characteristics provided in study:				
										Fixation stain resolution	Data types	Variables in model	Number of participants	Demo-graphic information
Boehm et al, 2022 ²²	USA	GHC*	(Ovary) HGSOc	MSKCC; TCGA-OV	USA, USA*	-	FFPE H&E	CNV, somatic mutation, SNP	RD status, PARP inhibitor administration	283 (444)	Age, gender, race	Stage	Partially	Yes
Cheerla and Gevaert, 2019 ²³	USA	Multimodal (Clin + miRNA + mRNA + WSI)	(Multi - 20) Pan cancer	TCGA-PanCanAtlas	USA*	-	-	mRNA, miRNA	Race, age, gender, grade	11160	-	-	All open source*	Yes
Chen et al, 2021 ²⁴	China	HTRS	(Head & Neck) SCC	TCGA	-	*	-	mRNA	-	212	Age, gender	Tumour site, grade, stage	On request	-
Chen et al, 2021 ²⁵	China	Multimomics model	(Lung) Adenocarcinoma	TCGA (LUAD)	USA*	*	FFPE H&E	mRNA, genomics, protein expression	-	470	Age, gender	Stage	All open source	-
Chen et al, 2022 ²⁶	USA	MMF	(Multi - 14) Pan cancer	TCGA (BLCA, BRCA, COADREAD, HNSC, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, PAAD, SKCM, STAD, UCEC)	USA*	-	FFPE H&E	CNV, somatic mutation, mRNA	-	5720	Age, gender, ethnicity	Grade	All open source	Yes
Chen et al, 2022 ²⁷	USA	Pathomic Fusion	(Multi - 2) Glioma	TCGA (LGG, GBM, KIRC)	USA*	-	FFPE* H&E	CNV, somatic mutation, mRNA	-	769 (1186)	-	-	All open source*	Yes
Cheng et al, 2017 ²⁸	China	Lasso-Cox	Renal (CCRCC)	TCGA	USA*	-	-	mRNA	-	410	Age, gender	Grade, Stage	Yes	Yes*
Hao et al, 2020 ²⁹	USA	PAGE-NET	(Brain) Glioblastoma	TCGA-GBM	USA*	-	20x, 40x	mRNA	Age	447	Age	-	All open source*	Yes
Hou et al, 2022 ³⁰	China	Multi-modality	(Liver) HCC	TCGA	USA*	-	Frozen H&E* 20x	mRNA	Age, Sex	346	Age, Gender	T group, N group, M group and overall TNM stage	All open source	Yes
Hou et al, 2023 ³¹	China	HGCN	(Multi - 6) KIRC	TCGA	USA*	-	FFPE* H&E*	"Genomic profile data"	Age, Gender, Therapy, BMI	385 (2102)	-	-	All open source*	Yes
Ji et al, 2024 ³²	China	HGRS	(Renal) CCRCC	TCGA	USA*	-	FFPE H&E	mRNA	-	281	Age, gender	TNM stage, Grade	All open source	-
Li et al, 2020 ³³	China	DeepHit	(Breast) -	TCGA-BRCA	USA*	-	-	mRNA	-	826	-	-	All open source	-
Li et al, 2021 ³⁴	China	HGPF	(Colon) Adenocarcinoma	TCGA-COAD	USA*	-	FFPE H&E*	mRNA	-	199	Age, gender	Stage/TNM classification	All open source	-

(continued)

Table 1. (continued)

Author, year	Location	Model of interest	Cancer types	Internal data			Ext. data		Participant characteristics provided in study:						
				Name	Origin	Name (origin)	Fixation stain resolution	Omic data	Clinical data	Number of participants	Demo-graphic information	Disease information	Data available	Code available	
Li et al, 2022 ³⁵	China	HFBSurv	(Multi - s10) Breast	TCGA	USA*	-	-	mRNA, CNV	-	-	1015 (-)	-	-	All open source*	Yes
Liu et al, 2023 ³⁶	China	MGCT	(Multi - 5) Pan cancer	TCGA (LUAD, BRCA, BLCA, GBMLGG, UCEC)	USA*	-	-	CNV, somatic mutation, mRNA	-	-	3557	-	-	All open source*	Yes
Liu et al, 2024 ³⁷	China	IntraSA-InterCA	(Breast) -	TCGA	USA*	-	-	mRNA	-	-	345	-	-	All open source*	-
Lv et al, 2021 ³⁸	China	PG-TFNet	(Colorectum) -	TCGA	USA*	-	-	mRNA	Age, gender, pathologic stage	-	522	-	-	All open source*	-
Lv et al, 2023 ³⁹	China	Trans-Surv	(Colorectum) -	TCGA; (NCT-CRC-HE-100K; CRC-VAL-HE-7K tissue classifier used in model)	USA*	-	-	CNV, mRNA	Age, gender, pathologic stage	-	520	-	-	All open source*	-
Ning et al, 2020 ⁴⁰	China	Gene + His*	(Renal) CCRCC	TCGA	USA*	-	-	mRNA	-	-	209	Age, gender	Grade, stage	All open source*	Yes
Ning et al, 2023 ⁴¹	China	McLR Framework	(Multi - 33) LIHC	TCGA (KIRC, LIHC, LUAD)	USA*	-	-	mRNA	-	-	334 (1177)	Age, gender	Stage, pT stage	All open source*	Yes*
Perez-Herrera et al, 2024 ⁴²	Spain	Model 3	(Breast) -	TCGA (BRCA) & TIGER; (segmentation training only); METABRIC	USA*; Netherlands & Belgium	METABRIC (Canada-UK*)	-	SNP, CNV, mRNA	Age, subtype	-	3593	Age	Subtype	All open source*	-
Qiu et al, 2024 ⁴³	China	DDM-net	(Brain) Glioma	TCGA (GBMLGG)	USA*	-	-	CNV, somatic mutation, mRNA	-	-	954	-	WHO grading	All open source	-
Shao et al, 2020 ⁴⁴	China	OMMF5	(Multi - 3) LUSC	TCGA (KIRC, KIRP, LUSC)	USA*	-	-	CNV, DNAm, mRNA	-	-	365 (787)	-	Stage	All open source*	-
Shao et al, 2023 ⁴⁵	China	FAM3L	(Multi - 3) Multi (KIRP)	TCGA (BRCA, KIRP, KIRC)	USA*	-	-	mRNA	-	-	251 (1824)	Age	-	All open source*	-
Shao et al, 2020 ⁴⁶	China	M2DP	(Multi - 3) BRCA	TCGA (LUSC, BRCA, LIHC)	USA*	-	-	mRNA	Stage	-	790 (1324)	-	Stage	All open source*	-
Shao et al, 2023 ⁴⁷	China	IMO-TILS	(Multi - 3) Breast	TCGA	USA*	-	-	mRNA, miRNA	-	-	365 (920)	Age	-	All open source*	-
Steyaert et al, 2023 ⁴⁸	USA	Late fusion	(Brain - Multi) Adult Glioblastoma	TCGA; Paediatric Brain Tumour Atlas (PBTA) Kids First cohort	USA*	CPTAC-GBM (USA*)	-	mRNA	-	-	783 (1088)	-	Low grade, high grade	All open source	Yes

(continued)

Table 1. (continued)

Author, year	Location	Model of interest	Cancer types	Internal data		Ext. data		WSI data		Omic data		Clinical data		Participant characteristics provided in study:			
				Name	Origin	Name (origin)	Fixation stain resolution	Data types	Variables in model	Number of participants	Demo-graphic information	Disease information	Data available	Code available			
Subramanian et al, 2021 ⁴⁹	USA	GCCA	(Breast) -	TCGA (BRCA)	USA*	-	H&E*	mRNA	-	974	-	-	All open source	Yes			
Subramanian et al, 2024 ⁵⁰	USA	SCCA	(Breast) -	TCGA (BRCA)	USA*	-	H&E	mRNA	-	974	Age, gender, ethnicity	-	Yes	Yes			
Sun et al, 2018 ⁵¹	China	GPML	(Breast) -	TCGA (BRCA)	USA*	-	H&E	CNV, DNAm, mRNA protein expression	-	578	Age	-	All open source*	-			
Tan et al, 2022 ⁵²	China	MultiCoFusion	(Brain) GBMLGG	TCGA (GBM, LGG)	USA*	-	H&E	mRNA	Grade	469	-	Grade	All open source*	-			
Vale-Silva et al, 2021 ⁵³	Germany	MultiSurv	(Multi - 33) Multi	TCGA	USA*	-	H&E	CNV, DNAm, mRNA, miRNA, DNAm	Age, gender, race, cancer type, tumour stage, prior malignancy, synchronous malignancy, prior treatment, pharmaceutical treatment, radiation treatment.	11 167	-	-	All open source*	Yes			
Vollmer et al, 2024 ⁵⁴	Germany	Random Survival Forest	(Oral) SCC	TCGA (HNSC)	USA*	-	H&E 40x	mRNA, somatic/germline DNA	Yes, but unclear what	406	Age, gender, race, ethnicity	Stage	All open source	Yes			
Wang et al, 2021 ⁵⁵	China	GPDBN	(Breast) -	TCGA (BRCA)	USA*	-	H&E*	mRNA	-	1015	-	-	All open source*	Yes			
Wang et al, 2023 ⁵⁶	China	HC-MAE	(Multi - 6) LGG	TCGA (LIHC, BRCA, LUAD, COAD, LGG, STAD)	USA*	-	H&E*	DNAm, mRNA, miRNA	-	451 (2420)	-	-	All open source*	Yes			
Wei et al, 2023 ⁵⁷	China	MultiDeepCox-SC	(Stomach) -	TCGA (STAD)	USA*	*	FFPE H&E 20x	mRNA	Age	357	Age, gender, race	Stage, grade	All open source	-			
Wu et al, 2023 ⁵⁸	China	CAMR	(Multi - 3) LGG	TCGA (BRCA, LUSC, LGG)	USA*	-	H&E*	CNV, mRNA	-	629 (2135)	-	-	All open source*	Yes			

(continued)

Table 1. (continued)

Author, year	Location	Model of interest	Cancer types	Internal data		Ext. data		WSI data		Omic data		Clinical data		Participant characteristics provided in study:				
				Name	Origin	Name (origin)	Fixation stain resolution	Data types	Variables in model	Number of participants	Deno-graphic information	Disease information	Data available	Code available				
Xie et al, 2024 ⁵⁹	China	GaCaMML	(Stomach)	Ruijin Hospital (Shanghai); GSE54129 (Public Chinese cohort, Ruijin)	China, China	TCGA-LUAD - (USA*)	-	H&E 40x	-	mRNA	63 (95)	Age, sex	Histological type, grade, TNM stage, Bormann classification, tumour position	On request	-			
Zeng et al, 2020 ⁶⁰	China	Multi-omics model	(Head & Neck) SCC	TCGA (HNSCC)	USA*	-	-	H&E 20x, 40x	-	mRNA, somatic mutation, protein expression	216	Age, gender	Anatomic site, stage	All open source	-			
Zeng et al, 2021 ⁶¹	China	Multi-omics model	(Ovary) HGSOc	TCGA	USA*	*	-	H&E	-	Genomics, mRNA, protein expression	229	Age	Stage	All open source*	-			
Zhan et al, 2021 ⁶²	China	Two-stage Cox-nnet	(Liver) HCC	TCGA	USA*	-	-	FFPE H&E	-	mRNA	290	-	-	All open source*	Yes			
Zhang et al, 2020 ⁶³	China	HI-MKL	(Brain) Glioblastoma	TCGA (GBM)	USA*	-	-	H&E* 20x, 40x	-	CNV, DNAm, mRNA	251	Age, gender	-	All open source*	-			
Zhao et al, 2023 ⁶⁴	China	Ada-RSIS	(Multi - 3) UCEC	TCGA (BLCA, UCEC, LGG)	USA*	TCGA (GBM, UCS; USA*)	-	H&E	-	mRNA	539 (1437)	Age, gender	Stage	All open source*	Yes*			
Zheng et al, 2024 ⁶⁵	USA	FSM	(Lung - Multi) LUAD	TCGA; National Lung Screening Trial (generating features)	USA*	CPTAC (LUAD, LUSC; USA*)	-	H&E*	-	mRNA	444 (1259)	Age, gender, race	Stage	All open source*	Yes			
Zhou et al, 2023 ⁶⁶	Hong Kong	CMTA	(Multi - 5) GBMLGG	TCGA (BLCA, BRCA, GBMLGG, LUAD, UCEC)	USA*	-	-	H&E 40x	-	CNV, mRNA, SNP	569 (2831)	-	-	All open source*	Yes			
Zhou et al, 2023 ⁶⁷	USA	Integrative CNN	(Colon) Adenocarcinoma	TCGA (COAD); Wayne State University	USA*, USA	TCGA-READ (USA*)	-	FFPE H&E	-	Somatic mutation, MSI signature	Unclear	Age, sex	Stage	Yes	Yes			
Zhou et al, 2024 ⁶⁸	Hong Kong	MSEN	(Multi - 5) GBM-LGG	TCGA (BLCA, BRCA, LUAD, UCEC, GBMLGG)	USA*	-	-	H&E*	-	CNV, Somatic mutation, mRNA	492 (2670)	-	-	All open source*	Yes*			
Zhu et al, 2023 ⁶⁹	China	SAMMS	(Multi - 2) LGG	TCGA (LGG, KIRC)	USA*	-	-	H&E* 20x	-	CNV, mRNA, miRNA	500 (806)	-	-	All open source*	-			

Asterisk generally used to denote fields where answer reasonably assumed by both investigators; - used where information not provided or unclear; **Model of interest (asterisk)** – indicates where an ineligible multimodal model (eg, incorporating radiology) achieved a higher result than our model of interest; **External data (asterisk)** where external validation datasets are used within the study but not for the multimodal model. **Code availability (asterisk)** – indicates where code links not provided or where link did not lead to associated code despite described as being available.

Abbreviations: BMI, body mass index; CNV, copy number variation; CPTAC, clinical proteomic tumour analysis consortium; DNAm, DNA methylation; FFPE, formalin fixed paraffin embedded; H&E, haematoxylin and eosin; mRNA, messenger RNA; miRNA, micro-RNA; MSKCC, memorial Sloan Kettering cancer centre; SNP, Single nucleotide polymorphisms; TCGA, the cancer genome atlas.

Table 2. Characteristics of the Models of Interest Included in the Systematic Review.

Method	Author, year (ref)	Modelling description	Study size		Events	Predictors		WSI Feature	Omic feature		Performance measures				Internal results					
			n	n (%)		Cand.	Final		Approach (named method)	Named method (s)	Fusion level	Validation	Calibration	Discrimination	Overall	Clin. Utility	Variability (metric)	AUC	Variability (metric)	
ML regularised regression	Cheng et al, 2017 ²⁸	Lasso-Cox Image features and eigengenes into lasso-regularised Cox Proportional Hazards model	410	135 (33)	-	165	13	Hand crafted	ImQCM	Feature	Cross-validation	-	-	Log-rank test/Risk group curves	-	-	-	-		
	Shao et al, 2020 ⁴⁴	OMIFS Feature selection via ordinal multi-modality feature selection model, then features into Cox Proportional Hazards model	365	145 (40)	-	2011	16	Hand crafted	WGCNA	Feature	Random split data	-	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	0.773	x	0.793	x	
	Shao et al, 2020 ⁴⁶	MZDP Multi-task, multi-modal feature selection algorithm with Cox Proportional Hazards model	790	82 (10)	-	-	155	Hand crafted	ImQCM	Feature	Cross-validation	-	-	Brier score	-	0.726	±0.06 (unclear)	-	-	
	Shao et al, 2023 ⁴⁵	FAM3L Feature-aware multi-modal metric learning method learns survival associated features which are combined into Cox Proportional Hazards model	251	38 (15)	-	-	163	Hand crafted	ImQCM	Feature	Cross-validation	-	-	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	0.752	±0.061 (unclear)	0.791	0.069 (unclear)
Machine learning	Sun et al, 2018 ⁵¹	GPMKL Selected features integrated by multiple kernel learning, with a separate kernel per data modality	578	445 (77)	-	59428	350	Hand crafted cellprofiler	-	Feature	Cross-validation	-	-	C-Statistic/AUC graph	-	0.643	x	0.828	±0.034 (St Err)	
	Zeng et al, 2020 ⁶⁰	Multi-omics model Extracted image and omic features integrated directly into random forest model	216	106 (49)	-	35302	954	Hand crafted cellprofiler	-	Feature	Cross-validation	-	-	AUC graph/Log-rank test/Risk group curves	-	DCA	-	0.929	x (5-y)	
	Zhang et al, 2020 ⁶³	HI-MKL Histopathological image feature integrating multiple kernel learning method - multiple kernels derived from selected features are combined within each data type, before combining across modalities for a combined kernel for MKL	251	204 (81)	-	66650	350	Hand crafted cellprofiler	-	Feature	Cross-validation	-	-	AUC graph/Risk group curves	-	-	-	0.932	x	
	Chen et al, 2021 ²⁵	Multionics model Selected histopathology image and multionics features (multiple selection techniques) integrated by Random Forest model	470	169 (36)	-	-	954	Hand crafted cellprofiler	-	Feature	Cross-validation	-	-	AUC graph/Log-rank test/Risk group curves	-	DCA	-	0.938	x (3-y)	
	Chen et al, 2021 ²⁴	HTRS Selected histopathology image and gene features integrated by a Random Forest model	212	-	-	5532	259	Hand crafted cellprofiler	WGCNA	Feature	Cross-validation	-	Calib. plot	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	DCA	0.768	0.715-0.820 (95%CI)	0.826	x (5-y)
	Li et al, 2021 ²⁴	HGPF Direct integration of features selected by SVM-RFE and LASSO-COX (images), and WGCNA (hub genes) into Random Forest	199	32 (16)	-	20344	377	Hand crafted cellprofiler	WGCNA	Feature	Cross-validation	-	Calib. plot	AUC graph/Log-rank test/Risk group curves	-	DCA	-	0.924	x (5-y)	
	Subramanian et al, 2021 ⁴⁹	GCCA Graph-structured variant of sparse canonical correlation analysis, capturing intra and intermodality correlations	974	-	-	-	3213	Hand crafted cellprofiler	-	Feature	Cross-validation	-	-	F1 score	-	-	-	-	-	
	Zeng et al, 2021 ⁶¹	Multi-omics model Extracted image features and selected omics features integrated directly into Random Forest model	229	133 (58)	-	-	970	Hand crafted cellprofiler	-	Feature	Cross-validation	-	-	AUC graph/Log-rank test/Risk group curves	-	DCA	-	0.911	x (5-y)	
	Ning et al, 2023 ⁴¹	McLR Framework Unimodal features concatenated into multimodal feature matrix and used as input to multi-constraint latent representation framework	334	116 (35)	-	20003	256	Hand crafted PFTAS	-	Feature	Cross-validation	-	-	C-Statistic/Log-rank test/Risk group curves	-	0.72	0.045 (StD)	-	-	
	Zhao et al, 2023 ⁶⁴	Ada-RSIS Adaptive risk-aware sharable and individual subspace learning - subspace learning of modality specific and modality sharable representations passed through function to capture intra and inter modality terms. Representations integrated via grouping co-expression constraint and gaussian-based weighting strategy.	539	91 (17)	-	20003	256	Hand crafted PFTAS	-	Feature	Cross-validation	-	-	C-Statistic/Risk group curves	-	0.711	±0.048 (StD)	-	-	
	Ji et al, 2024 ²²	HGRS Direct integration of features selected by SVM-RFE and LASSO-COX (images), and WGCNA (hub genes) into Random Forest survival model	281	76 (27)	-	-	8	Hand crafted cellprofiler	WGCNA	Hybrid	Cross-validation	-	Calib. plot	AUC graph/Log-rank test/Risk group curves	-	DCA	-	0.780	x (5-y)	

(continued)

Table 2. (continued)

Method	Author, year (ref)	Modelling description	Study size		Events		Predictors		WSI Feature		Omic feature		Performance measures				Internal results		
			n	n	n (%)	Cand.	Final	Approach (named method)	Named method (s)	Fusion level	Validation	Calibration	Discrimination	Overall	Clin. Utility	C-index	Variability (metric)	AUC	Variability (metric)
Deep learning	Subramanian et al., 2024 ⁵⁰	SCCA Probabilistic graph modelling by penalised canonical correlation analysis-based joint embeddings and Cox Proportional Hazards model	974	-	21 075	2075	-	Hand crafted cellprofiler	-	Feature	Cross-validation	-	C-Statistic/Risk group curves	-	-	0.683	±0.0669 (unclear)	-	-
	Vollmer et al., 2024 ⁵⁴	Random Survival Forest Feature selection via then Random Forest model for survival prediction.	406	177 (44)	-	242	-	Hand crafted cellprofiler	-	Feature	Cross-validation	-	C-Statistic	-	-	0.834	x	-	-
	Cheerla et al., 2019 ²³	Multimodal (Clin + miRNA + mRNA + WSI) Modality specific features generated via multiple CNNs compressed by unsupervised encoder into single vector for survival prediction by cox loss function. Multimodal dropout approach used in training.	11 160	-	62 308	2048	-	Learned SqueezeNet CNN	Deep Highway Neural Network	Feature	Random split data	-	C-Statistic	-	-	0.78	x (average result)	-	-
	Hao et al., 2020 ²⁵	PAGE-NET Pathology (CNN), genome and clinical (adapted CoxPASNet) layers fused in Cox Proportional Hazards model	447	-	6405	81	-	Learned Pre-trained CNN	CoxPASNET	Feature	Cross-validation	-	C-Statistic/Wilcoxon rank-sum test	-	-	0.702	±0.0294 (unclear)	-	-
Deep learning	Li et al., 2020 ²³	DeepHit Fully connected neural network with image and omic feature fusion via conditional autoencoder.	826	-	-	-	-	Learned ResNet, DCGMM	-	Feature	Cross-validation	-	C-Statistic	-	-	0.757	0.713-0.801 (95%CI)	-	-
	Ning et al., 2020 ⁴⁶	Gene + His Cross-modal feature-based integrative framework. Deep features extracted from images (CNNs), combined with eigengenes (WGCNA) and fed into cox model for prediction	209	60 (29)	-	138	-	Learned CNN	WGCNA	Feature	Cross-validation	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.638	0.527-0.749 (unclear)	-	-
	Lv et al., 2021 ³⁸	PG-TFNet Output of transformer-based multiscale pathological feature fusion module (images) and multilayer perceptron features (genes) integrated by cross attention transformer-based multimodal feature fusion module and fed into cox layer (with clinical data) for survival prediction.	522	-	-	-	-	Learned ResNet (multiscale)	MLP	Feature	Cross-validation	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.816	±0.016 (unclear)	-	-
	Vale-Silva et al., 2021 ⁵³	MultiSurv Dedicated deep learning sub models per data modality (n=6), fed into data fusion layer then final network takes fused feature representation and outputs conditional survival probabilities.	11 081	360 (32)	108 437	3072	-	Learned ResNet50 CNN (ImageNet)	CNN	Feature	Bootstrap	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.787	0.769-0.806 (95%CI)	-	-
Deep learning	Wang et al., 2021 ⁵⁵	GPDBM Outputs of modality specific bilinear feature encoding modules (BFEM) and an inter-modality BFEM fused and fed into fully connected neural network for survival prediction	1015	-	20 446	64	-	Hand crafted cellprofiler	-	Feature	Cross-validation	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	-	0.725	±0.066 (unclear)	0.808	±0.053 (unclear)
	Zhan et al., 2021 ⁶²	Two-stage Cox-nnet Outputs of modality specific cox-nnet models integrated as nodes to a second stage cox-nnet with an output layer for cox regression.	290	-	-	-	-	Hand crafted cellprofiler	-	Feature	Cross-validation	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.75	x	-	-
	Boehm et al., 2022 ²²	GHC Fusion of negative log partial hazards from unimodal cox models with learned image features into a multimodal linear cox proportional hazards model	283	-	-	5	-	Combo ResNet18	HRD status inferred	Decision	Random split data	-	C-Statistic/Log-rank test/Risk group curves/Kendall rank correlation	-	-	0.55	0.531-0.563 (95%CI)	-	-
	Chen et al., 2022 ²⁶	MMF Kronecker product fusion of unimodality subnetworks to generate multimodal representation feature and predict survival via log likelihood function	5720	1651 (29)	-	64	-	Learned ResNet50 (ImageNet), attention module	SNN	Feature	Cross-validation	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	-	0.644	x (average result)	0.662	x (average result)

(continued)

Table 2. (continued)

Method	Author: year (ref)	Modelling description	Study size		Events	Predictors		WSI Feature	Omic feature		Performance measures				Internal results			
			n	n	n (%)	Cand.	Final	Approach (named method)	Named method (s)	Fusion level	Validation	Calibration	Discrimination	Overall	Clin. Utility	C-index	Variability (metric)	Variability (metric)
	Chen et al, 2022 ²⁷	Pathomic Fusion Unimodal networks as feature extractors (CNN and GCN - images, Feed-forward SNN - genomic) fused by gating attention mechanism and Kronecker product to model modality interactions. Network supervised by cox partial likelihood loss function.	769	-	-	-	-	Combo VGG19 (ImageNet), GCN	SNN	Feature	Cross-validation	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.826	±0.009 (unclear)	-
	Hou et al, 2022 ³⁰	Multi-modality Pathology patch level risk score (via multi-instance fully convolutional network aggregated to patient level by attention mechanism) integrated with hub genes into Cox Proportional Hazards model.	346	123 (36)	5005	8	Learned VGG19	WGNA	Hybrid	Cross-validation	-	Calibration plot	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	DCA	0.746	±0.077 (95%CI)	0.816 x (1-y)
	Li et al, 2022 ³⁵	HFBSurv Fusion of modality-specific and cross-modality attentional factorised bilinear modules with cox partial likelihood loss function for survival prediction	1015	-	46 125	240	Hand crafted cellprofiler	-	Feature	Cross-validation	-	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	-	0.766	±0.024 (unclear)	0.806 ±0.025 (unclear)
	Tan et al, 2022 ⁵²	MultiCoFusion Features concatenated then fused via three-layer FCNN before output fed to multitask output.	469	-	62 437	2000	Learned ResNet-152 (ImageNet)	SGCN	Feature	Cross-validation	-	-	C-Statistic/AUC graph/Risk group curves	-	-	0.857	±0.015 (unclear)	0.915 ±0.016 (unclear)
	Hou et al, 2023 ³¹	HGCN Distinct modality graph representations fed into Graph Convolutional Network layers then hypergraph convolutional layer with hyperedge mixing module (intermodal interaction) before decision via cox loss. Jointly trained with online masked autoencoder (to simulate missing data)	385	-	21 504	3	Learned KiniaNet (CNN)	GSEA	Decision	Cross-validation	-	-	C-Statistic/Log-rank test/Risk group curves	Brier score	-	0.747	±0.007 (95%CI)	-
	Liu et al, 2023 ³⁶	MGCT Modality specific embeddings integrated with two-stage mutual guided cross-modality transformer framework	3557	-	-	256	Learned ResNet50 (ImageNet)	SNN	Feature	Cross-validation	-	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	-	0.663	x (average result)	-
	Lv et al, 2023 ³⁹	Trans-Surv Multi Scale Fusion Transformer (image feature extraction), integration of image and omic data by cross-modal fusion transformer, combined with clinical data into Cox Layer	520	-	-	-	Learned Multiple CNNs	-	Feature	Cross-validation	-	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.822	±0.023 (unclear)	-
	Shao et al, 2023 ⁴⁷	IMO-TILS Deep generalised canonical correlation analysis with attention layer to fuse image and multi-omic data	365	37 (9)	272	272	Learned Unet++ , ResNet-101 (ImageNet), KNN	ImQCM	Feature	Cross-validation	-	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	-	0.709	±0.030 (unclear)	0.737 ±0.047 (unclear)
	Steyart et al, 2023 ⁴⁸	Late Fusion Modality specific cox proportional hazards models using model derived image (CNN) and omic (MLP) features. Outputs integrated into a final cox regression module (late fusion)	783	-	12 879	4096	Learned ResNet50 CNN	MLP	Decision	Cross-validation	-	-	C-Statistic/Log-rank test/Risk group curves	Brier score, Composite score	-	0.797	±0.019 (Std)	-
	Wang et al, 2023 ⁵⁶	HC-MAE Hierarchical cross-attention masked autoencoder - multiscale WSI representations (guided by omic cross attention) and omic features fed through separate transformers then fused via concatenation. Survival function negative log partial likelihood	451	-	-	2052	Learned vision transformer-based masked autoencoder	MLP	Feature	Cross-validation	-	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.851	±0.033 (unclear)	-
	Wei et al, 2023 ⁵⁷	MultiDeepCox-SC Integration of age and gene predictors with risk score of deep learning-based image model (DeepCoxSC) using a cox model	357	142 (40)	61 584	12	Hand crafted cellprofiler	-	Hybrid	Cross-validation	-	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	-	0.744	±0.070 (unclear)	0.833 ±0.055 (unclear) (2-y)

(continued)

Table 2. (continued)

Method	Author: year (ref)	Modelling description	Study size		Events		Predictors		WSI Feature	Omic feature		Performance measures				Internal results			
			n	n	n (%)	Cand.	Final	Approach (named method)		Named method (s)	Fusion level	Validation	Calibration	Discrimination	Overall	Clin. Utility	C-index	Variability (metric)	Variability (metric)
	Wu et al., 2023 ⁵⁸	CAMR Cross-aligned multimodal representation learning - modality specific representations and modality invariant representations learnt by separate networks and fused by a gating attention mechanism and concatenation. Survival function is cox partial likelihood loss.	629	-	-	-	240	Hand crafted cellprofiler	-	-	Feature	Cross-validation	-	C-Statistic/AUC graph/Risk group curves	-	-	0.841	±0.20 (unclear)	0.889 ±0.017 (unclear)
	Zhou et al., 2023 ⁶⁶	CMTA Cross-modal translation alignment - two parallel encoder-decoder structures integrate intra-modal representations and cross-modal representations generated by an attention mechanism.	569	-	-	-	-	Learned ResNet-50 (ImageNet)	-	FCNN	Feature	Cross-validation	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.853	±0.0116 (unclear)	-
	Zhou et al., 2023 ⁶⁷	Integrative CNN Tile level CNN image features concatenated with clinical variables and mutation status and/or signature. Final feature vector into MLP to predict risk.	-	-	-	-	-	Learned Inception V3 (ImageNet)	-	MSI status	Feature	Cross-validation	-	C-Statistic/AUC graph/Risk group curves	-	-	0.69	±0.19 (unclear)	0.8 x
	Zhu et al., 2023 ⁶⁹	SAMMS Pre-trained SAM for images + omic data into modality specific and cross-modality common subnetworks. Outputs concatenated with clinical data and fusion representation used for cox prediction layer.	500	-	-	-	102	Learned SAM (pretrained)	-	-	Feature	Cross-validation	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	-	0.843	x	0.782 x
	Liu et al., 2024 ²⁷	IntraSA-InterCA Modality specific and intramodality modules for generation of feature embeddings integrated by an adaptive fusion block into a combined feature. Survival prediction via multilayer deep neural network.	345	-	-	22 779	64	Hand crafted cellprofiler	-	-	Feature	Cross-validation	-	C-Statistic/AUC graph/Log-rank test/Risk group curves	-	-	0.742	±0.039 (unclear)	0.841 ±0.032
	Perez-Herrera et al., 2024 ⁴²	Model 3 Integration of image derived variables (UNet), omic features (multiple models) with clinical features via Cox Proportional Hazards model	3593	-	-	-	12	Learned -	-	-	Feature	Cross-validation	-	Hazard ratios	-	-	-	-	-
	Qiu et al., 2024 ⁴³	DDM-Net Dual-space disentangled-multimodal adversarial autoencoder comprising two visual autoencoders and a fusion module input to a cox proportional hazards model for survival prediction. Two imputation networks designed to handle missing data.	954	-	-	121 415	2048	Learned ResNet50 (ImageNet)	-	SNN	Feature	Cross-validation	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.824	±0.10 (unclear)	-
	Xie et al., 2024 ⁶⁹	GaCaMML Extracted features fed through cross-modal attention layer and MIL aggregation layer of model before fusion of gene and image features by concatenation. The log-likelihood loss function is used for survival prediction.	63	12 (19)	30 000	256	Learned ResNet50 (ImageNet)	-	-	FCNN	Feature	Cross-validation	-	C-Statistic/Log-rank test/Risk group curves	-	-	0.613	x	-
	Zheng et al., 2024 ⁶⁵	FSM Undirected graph embeddings combined using an attention-based mechanism	444	156 (35)	-	-	-	Learned CNN (pretrained)	-	FCNN	Feature	Cross-validation	-	C-Statistic/AUC graph	-	-	0.703	0.017 (StD)	0.679 (StD)
	Zhou et al., 2024 ⁶⁸	MSEN Multimodal Survival Ensemble Network - unimodal bag representations of WSIs and omic data are passed through feature encoders and attention-based modules, then integrated via an ensemble strategy and input to Cox Proportional Hazard model	492	-	-	-	-	Learned HIPT	-	Ensemble based encoder	Feature	Cross-validation	-	C-Statistic/Risk group curves	-	-	0.827	±0.018 (StD)	-

Abbreviations: Calib, calibration; CI, confidence interval; Clin., clinical; CNN, convolutional neural network; DCA, decision curve analysis; DCGMM, deep conditional gaussian mixture model; FC, fully connected; GCN, graph convolutional network; GSEA, gene set enrichment analysis; HIPT, hierarchical image pyramid transformer; HRD, homologous recombination deficiency; KNN, K nearest neighbour; ImQCM, local maximal quasi-clique merger; MLP, multi-layer perceptron; PFTAS, parameter free threshold adjacency statistics; SAM, segment anything model; SGCN, sparse graph convolutional network; SNN, self-normalising network; StD, standard deviation; StErr, standard error; WGCNA, weighted gene co-expression network analysis.

Table 3. Risk of Bias and Applicability Assessment.

Author, year (ref)	Development							Evaluation							Overall			
	Quality				Applicability			Risk of bias				Applicability			Dev		Eval	
	1. Participants	2. Predictors	3. Outcome	4. Analysis	1. Participants	2. Predictors	3. Outcome	1. Participants	2. Predictors	3. Outcome	4. Analysis	1. Participants	2. Predictors	3. Outcome	Quality	Applicability	Risk of Bias	Applicability
Boehm et al, 2022 ²²	?	-	+	?	+	?	+	?	-	+	-	+	?	+	-	?	-	?
Cheerla and Gevaert, 2019 ²³	?	?	+	?	+	+	+	?	?	+	-	+	+	+	?	+	-	+
Chen et al, 2021 ²⁴	?	?	+	?	+	+	+	?	?	+	?	+	+	+	?	+	?	+
Chen et al, 2021 ²⁵	?	?	+	?	+	+	+	?	?	+	-	+	+	+	?	+	-	+
Chen et al, 2022 ²⁶	?	?	?	-	+	+	?	?	?	?	-	+	+	?	-	?	-	?
Chen et al, 2022 ²⁷	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Cheng et al, 2017 ²⁸	?	?	+	-	+	+	+	?	?	+	-	+	+	+	-	+	-	+
Hao et al, 2020 ²⁹	?	?	?	-	+	+	?	?	?	?	-	+	+	?	-	?	-	?
Hou et al, 2022 ³⁰	?	?	+	-	+	+	+	?	?	+	-	+	+	+	-	+	-	+
Hou et al, 2023 ³¹	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Ji et al, 2024 ³²	?	?	+	-	+	+	+	?	?	+	?	+	+	+	-	+	?	+
Li et al, 2020 ³³	?	?	?	?	+	+	?	?	?	?	-	+	+	?	?	?	-	?
Li et al, 2021 ³⁴	?	?	+	?	+	+	+	?	?	+	?	+	+	+	?	+	?	+
Li et al, 2022 ³⁵	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Liu et al, 2023 ³⁶	?	?	+	?	+	+	+	?	?	+	-	+	+	+	?	+	-	+
Liu et al, 2024 ³⁷	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Lv et al, 2021 ³⁸	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Lv et al, 2023 ³⁹	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Ning et al, 2020 ⁴⁰	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Ning et al, 2023 ⁴¹	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Perez-Herrera et al, 2024 ⁴²	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Qiu et al, 2024 ⁴³	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Shao et al, 2020 ⁴⁴	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Shao et al, 2023 ⁴⁵	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Shao et al, 2020 ⁴⁶	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Shao et al, 2023 ⁴⁷	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Steyaert et al, 2023 ⁴⁸	?	+	+	-	+	+	+	?	+	+	+	+	+	+	-	+	-	+
Subramanian et al, 2021 ⁴⁹	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Subramanian et al, 2024 ⁵⁰	?	?	+	-	+	+	+	?	?	+	-	+	+	+	-	+	-	+
Sun et al, 2018 ⁵¹	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Tan et al, 2022 ⁵²	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Vale-Silva et al, 2021 ⁵³	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Vollmer et al, 2024 ⁵⁴	-	?	+	?	+	+	+	-	?	+	-	+	+	+	-	+	-	+
Wang et al, 2021 ⁵⁵	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Wang et al, 2023 ⁵⁶	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Wei et al, 2023 ⁵⁷	?	?	+	?	+	+	+	?	?	+	-	+	+	+	?	+	-	+
Wu et al, 2023 ⁵⁸	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Xie et al, 2024 ⁵⁹	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Zeng et al, 2020 ⁶⁰	?	?	+	?	+	+	+	?	?	+	-	+	+	+	?	+	-	+
Zeng et al, 2021 ⁶¹	?	?	+	-	+	+	+	?	?	+	-	+	+	+	-	+	-	+
Zhan et al, 2021 ⁶²	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Zhang et al, 2020 ⁶³	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Zhao et al, 2023 ⁶⁴	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?
Zheng et al, 2024 ⁶⁵	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Zhou et al, 2023 ⁶⁶	?	?	+	?	+	+	+	?	?	+	-	+	+	+	?	+	-	+
Zhou et al, 2023 ⁶⁷	?	-	+	?	+	+	+	?	-	+	-	+	+	+	-	+	-	+
Zhou et al, 2024 ⁶⁸	?	?	+	?	+	+	?	?	?	+	-	+	+	?	?	?	-	?
Zhu et al, 2023 ⁶⁹	?	?	+	-	+	+	?	?	?	+	-	+	+	?	-	?	-	?

Source: High concern (-, shaded red), Unclear (?), shaded yellow), Low concern (+, shaded green).

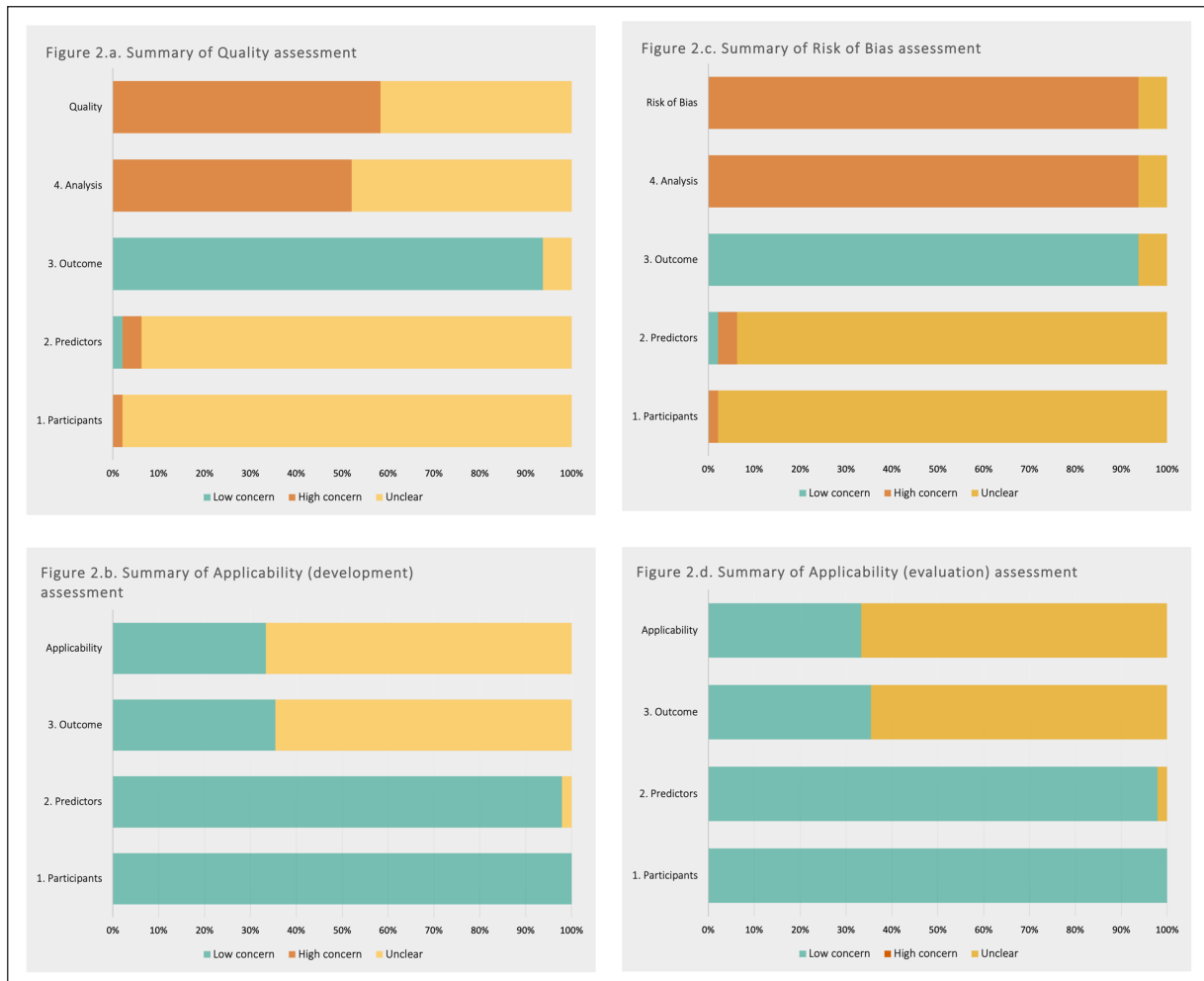


Figure 2. PROBAST + AI summary results. Summarised results for the assessments of (a) quality and (b) applicability (study development), and (c) risk of bias and (d) applicability (study evaluation) for the 48 studies in this review.

described performing these across both modalities and was rated as low concern.⁴⁸

In one of the studies, participants from one contributing site in the TCGA dataset were removed after persistent variation in stain post processing, which raised concern for selection bias.⁶⁷ Another study integrated -omic data from different datasets, generated by different methods without clear discussion about the comparability of these methods and any consequent differences when used to infer the homologous recombination (HRD) status from them.²²

Outcome. Assessments of quality and RoB were generally low due to nature of the outcome (death), which is not so sensitive to bias. However, studies which did not provide enough information to appreciate the intended outcome prediction time horizon or assess the length of follow up of the patients used in the study were rated as unclear.^{26,29,33}

Analysis. In this domain, there was high concern of quality in 25 studies and of risk of bias in 45 studies. Common factors contributed to these decisions as outlined below.

Assessment of adequacy of sample size in ML and DL approaches is difficult due to the influence of regularisation

techniques, shrinkage methods, and hyperparameter tuning strategies aimed to prevent overfitting. The application of the traditional “rule of thumb” 10 events per variable are inappropriate in this setting and there are no clearly defined minimum criteria. As shown by Collins et al. (2021), model complexity and tuning parameters substantially increase data demands, and the number of outcome events may be more influential than total sample size alone.⁷⁶ The use of cross validation, which increases the effective size of the test set thereby reducing over optimism in the results further complicates assessment. Recent guidance encourages authors to justify their sample size for the reader.⁷⁷ However, none of the studies presented any calculation or justification of their sample size and many were missing key data such as sample size ($n=1$), number of events ($n=26$), number of candidate ($n=24$) or final predictors ($n=9$) used in the models making it difficult to assess event per variable.

Another key consideration was the common use of a complete case analysis approach, with subjects with missing data modalities excluded and variables with missing values in included patients often removed. This approach can introduce bias if the missingness is not at random. However, it is also appreciated to make sense as part of

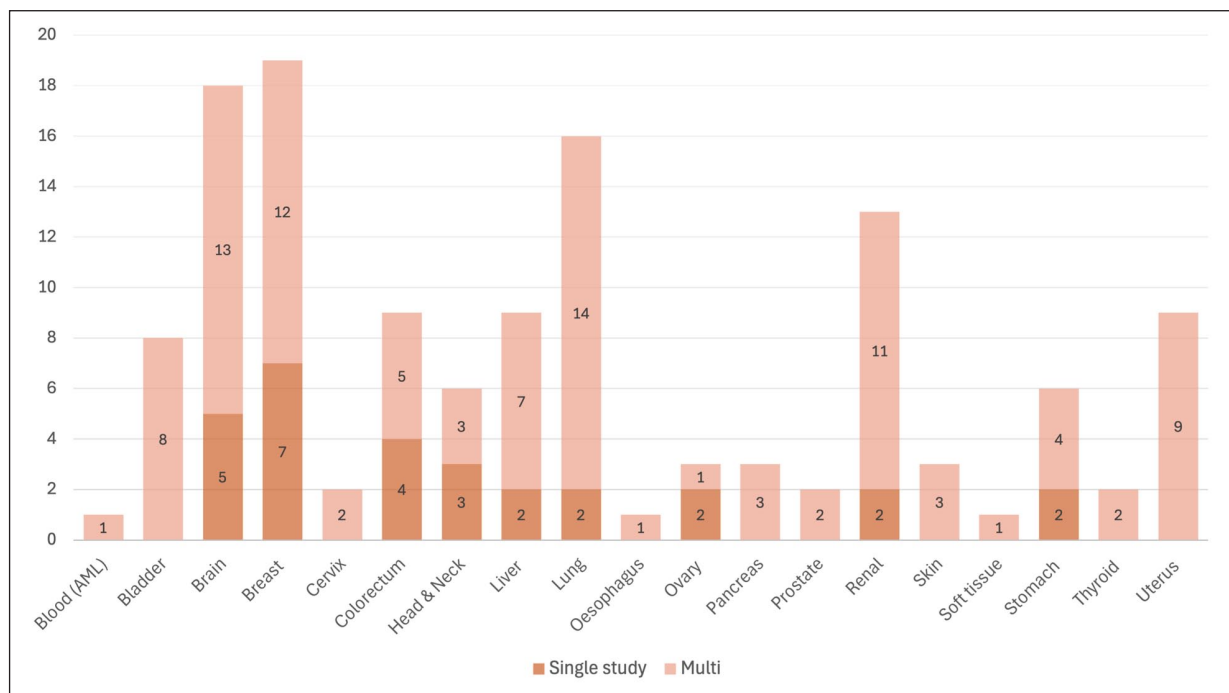


Figure 3. Cancer types represented across the studies in this review.

The number of studies (y-axis) are presented by organ systems (x-axis), showing a differentiation between whether they were part of multi- or single cancer studies. Nineteen studies developed models across multiple cancer subtypes and organs, with the remaining 29 being single cancer type studies. The most commonly studied cancers were those of the brain, breast and lung.

predictor reduction strategy. Three studies described a proactive approach to data missingness and are discussed in the model section below.

All studies reported measures of discrimination, which test the model's ability to correctly distinguish between patients who experience the event (eg, death) earlier versus later. However, only 4 evaluated calibration, which refers to how closely a survival prediction model's estimated probabilities of survival match the observed survival outcomes. These are complementary aspects of model performance and evaluating only 1 can miss serious model flaws—for example, a model may rank patients correctly (good discrimination) but still systematically over- or under-predict risk (poor calibration), leading to biased clinical decisions. The limited reporting of calibration substantially restricts assessment of clinical reliability and undermines the interpretability of reported performance gains, particularly in the context of clinical translation where absolute risk estimates are required for decision-making.

Additionally, the presentation of results did not include metrics of variability in 6 studies and in a further 19 the metric used was unclear.

Overall Quality and Risk of Bias. None of the studies had high or unclear ratings in just a single domain, indicating literature in this area faces broad methodological and reporting issues rather than isolated methodological flaws. In particular, incomplete participant description and missing validation detail limit confidence in the reported performance metrics and suggest current results should be treated as exploratory.

Overall Applicability. This systematic review was designed to be broad in scope. All studies were assessed as low concern in the participant domain as all were studies of cancer patients. In the predictor domain, all studies were rated as low concern, clearly using high-throughput -omic data and pathology whole slide images in their models. In the outcome domain, all studies were predictive of survival. However, 31 studies were not explicit about an outcome of OS and used ambiguous phrasing such as “cancer survival” or just “survival.” The most used dataset in this study (TCGA) is known to have associated cancer specific and OS data and as such specific phrasing was sought or the applicability of the study was assessed as unclear.⁷⁸

Data Synthesis Results

Datasets. Twenty-nine studies focussed on a cancer of a single organ/anatomical region, while 19 were multi-tissue studies. Nineteen different organs were studied, with the most studied being brain, breast, lung and kidney. Figure 3 summarises frequencies of the cancer domains studied.

Alongside WSIs, gene expression (mRNA) data was used in all studies, with additional -omics used in variable combinations; somatic mutation data (n=9), micro-RNA (n=5), copy number variation (CNV; n=16), single nucleotide variation (SNV; n=3), DNA methylation (n=5) and protein expression (n=5). Clinical data was integrated as an additional modality into 15 models and the variables used most commonly were age, gender and stage. The number of participants used per model of interest ranged from 63 to 11 160.

The TCGA dataset was used in some capacity in all studies, 47 for training, and in 3 studies as an independent external dataset. TCGA is a well-known, multi-institutional public dataset comprising mixed -omic data and WSIs for over 11 000 participants from the USA spanning 33 cancer types.

Other datasets used in multimodal model development and/or validation were from Clinical Proteomic Tumour Analysis Consortium (CPTAC),^{48,65} Memorial Sloan Kettering Cancer Centre,²² Molecular Taxonomy of Breast Cancer International Consortium (METABRIC),⁴² NCT-CRC-HE-100,³⁹ Paediatric Brain Tumour Atlas,⁴⁸ National Lung Screening Trial,⁶⁵ Tumour infiltrating lymphocytes in breast cancer (TIGER),⁴² Ruijin Hospital (Shanghai)⁵⁹ and Wayne State University.⁶⁷ The countries from which data were used include USA, China, Belgium and Netherlands. The predominant usage of open access public data repositories meant data from these studies was largely assessed as available, however only 18 studies included specific statements relating to data availability – asterisks in Table 1 indicate where data availability was assumed. Those that did not, all exclusively used the TCGA dataset.

Reliance on the TCGA dataset is a concern for this field. Such dependence increases the risk that models are inadvertently learning dataset-specific artefacts rather than generalisable and true biological patterns.

Modelling Methods. In this review, the model of interest was defined as the best performing model. Many of the studies, created more than one model by either: independently training and evaluating the same model architecture on datasets of multiple cancer types, combining different -omic combinations with image and/or clinical data into the model, or by testing multiple different model architectures or fusion approaches. Where additional models were present in a study, this is indicated in Supplemental Table S1.

Models. Modelling approaches included cox regression models with ML-based regularisation and/or feature selection ($n=4$),^{28,44–46} classical ML methods ($n=13$) and DL methods ($n=31$). ML methods included 7 models based on random forest,^{24,25,32,34,54,60,61} a multi constraint latent representation framework,⁴¹ 2 variants of canonical correlation analysis,^{49,50} two multiple kernel learning models^{51,63} and a multiview subspace learning model.⁶⁴ Of the 31 DL models, 25 used a cox-based loss as the objective function for survival prediction (log partial or negative log partial likelihood), but cross entropy loss^{26,65} and binary cross entropy loss^{37,55} were also used. One study used a custom likelihood and ranking loss function (DeepHit).³³

Within the models, fusion of data at a feature level was more common ($n=42$) than decision level fusion.^{22,31,48} Three studies used a hybrid approach, processing 1 modality to a decision level (risk score) for integration with features of another modality in the prediction model.^{30,32,57} Studies using feature fusion methods leveraged mechanisms to capture cross-modal interactions, most commonly through attention-based approaches, where models compare different parts of the input, assign relevance scores,

and weight them to focus on the most important information (see Supplemental Table S1).

While most models required fully matched datasets, 4 studies implemented explicit strategies to accommodate missing data modalities without excluding patients.^{23,31,43,53} Qiu et al. and Hou et al. used latent-space reconstruction—via variational autoencoders and online masked autoencoders, respectively—to recover missing features.^{31,43} In contrast, Vale-Silva and Rohr used multimodal dropout during training and standard imputation for partial missingness, while Cheerla and Gevaert dynamically reweighted available modalities during fusion.^{23,53} Despite differing mechanisms, all approaches aimed to maintain performance under incomplete input and reported reduced accuracy when modalities were missing. Models lacking such strategies may be vulnerable to overfitting and underperformance in real-world, incomplete datasets.

Model code was described as available in 25 studies, partially available code (as “related core code”) in 1 study⁴¹ and not commented on in the remaining 22 studies. Using links to code repositories, in 4 studies code could not be found,²⁸ was described as “coming soon”⁶⁸ or had been taken down and was now only available on request.^{41,64}

Data Pre-Processing. These complex data require pre-processing before they can be used in a model and reduction in the dimension of the also data helps to reduce model overfitting. Such steps were variably described in these studies and are summarised below and in Supplemental Table S1.

Most researchers performed tissue segmentation and/or stain normalisation of WSIs before dividing the original WSI into patches, ranging from 16×16 to 5000×5000 pixels, most commonly 1000×1000 pixels ($n=12$) or 512×512 pixels ($n=11$). Four studies used multiscale patches.^{27,38,39,56} Further selection of patches for inclusion was commented on in 32 studies and was performed at random ($n=9$), by RGB density ($n=8$), nuclear density ($n=2$), proximity to RGB mean ($n=1$), based on expert defined regions of interest ($n=9$), by exclusion of patches with low tissue coverage ($n=2$) or for not meeting feature extraction model criteria ($n=1$). Image features were generated equally by hand-crafted ($n=23$) or learned approaches ($n=23$) with 2 studies using a combined approach. CellProfiler was the most popular tool for generating hand-crafted features ($n=17$) and ResNet50 the most popular architecture for learned features ($n=7$). Some models incorporated graph-based ML,^{49,50} graph neural network,^{31,65} transformer^{38,39,56} or foundation model⁶⁹ approaches for feature generation.

Pre-processing of -omic data, where described, included removal of missing data, normalisation of read counts, and discretisation of expression or CNV data into categorical variables. The 2 studies which described correction for batch effect, both used ComBATSeq.^{48,65} In some studies, -omic data was used directly, however several performed feature generation, including; knowledge-based approaches, such as gene set enrichment analysis ($n=1$) and inference of known mutational signatures ($n=2$), gene modularisation by data driven methods, including Weighted Gene

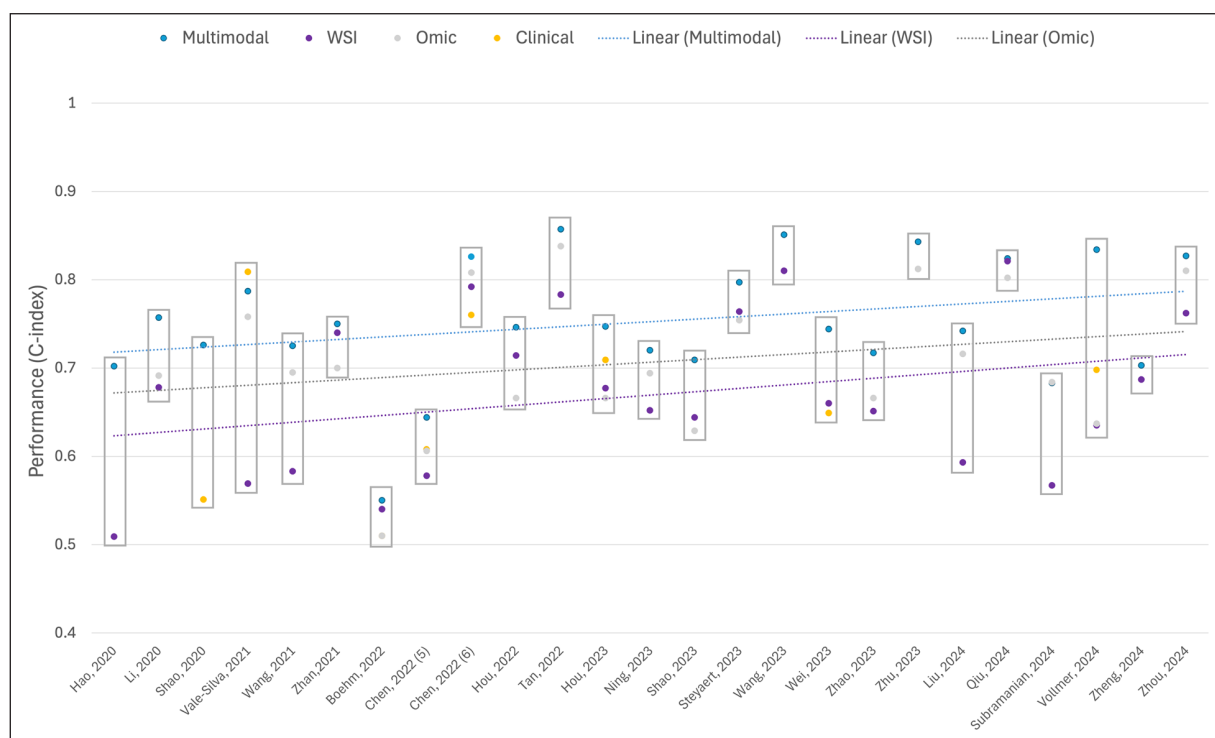


Figure 4. Comparison of multimodal and unimodal model performance using c-index.

The 25 studies which reported the c-index for unimodal comparison are shown above. Studies are ordered by year of publication (x-axis) with regression lines to show a trend of performance (c-index, y-axis) over time. Different studies performed different unimodal comparisons reflected by incomplete data points for several of the studies. Where multiple -omics were evaluated separately, the best performing is charted here. A further 7 studies evaluated unimodal performance against the multimodal model but generated AUC (n=6) or F1 metrics (n=1) so were not included in this comparison. Supplemental Table S1, indicates all studies which undertook unimodal performance comparisons.

Co-expression Network Analysis (n=6) and local maximal Quasi-Clique Merger (n=4), and generation of feature representations by ML methods (n=15).

Steps to reduce number of predictors were described in most studies, see Supplemental Table S1, and used a combination of data and knowledge-based approaches. The heterogeneity of approaches to feature generation (predictors) makes it difficult to compare number of predictors across studies. It was also challenging to determine the predictor number in many of the models. Particularly those where features are generated via DL and have cross-modal interactions resulting in complex representations from the WSI and -omic inputs.

Model Performance and Evaluation. All studies performed some form of internal validation, with most using a cross-validation approach (n=44) and the remaining few using random data splitting (n=3) and bootstrapping (n=1).

As discussed, studies focussed on measures of discrimination for reporting model performance. The c-index was the most commonly reported metric (39/48). It is a measure of the ability of the model to rank patients in terms of their predicted risk of event by indicating the proportion of all usable patient pairs in which the patient with the shorter observed survival time also has the higher predicted risk. A c-index of 1 indicates perfect performance and 0.5 is equivalent to random predictions. Results across these studies ranged from 0.550 to 0.857. Studies which additionally modelled survival as a binary classification task

(ie, predicting longer or shorter survival based on defined thresholds) reported AUC metrics (n=14). An AUC of 1 reflects perfect classification and 0.5 random guessing. Results across studies ranged 0.662 to 0.932. Time-specific AUC, which measures how well the model can distinguish between classes at a set time point, was reported in 8 studies with results across multiple time points but most commonly for 1-, 3- and 5-year survival. Two studies performed neither of these analyses, instead assessing model performance by hazard ratios⁴² and F1 score.⁴⁹

Five studies performed external validation on their multimodal model,^{48,59,64,65,67} with an additional 4 studies undertaking external validation but only of unimodal models in their study.^{24,25,57,61} Results of multimodal external validation ranged from 0.583 to 0.778 (c-index).

Only 7 studies performed clinical utility analyses, all via decision curve analysis.^{24,25,30,32,34,60,61} Four additionally explored end-user presentation of the model via development of nomograms^{24,30,32,34} and a further study generated an online tool.⁵⁷

Unimodal Performance Comparison. Thirty-four out of the 48 studies made a performance comparison between the multimodal model of interest and unimodal models based on clinical, -omic or image data. Where the c-index was available as a metric (n=25), these have been plotted in Figure 4. Studies which made a unimodal comparison but assessed this by another metric (AUC=10, F1=1) could not be included. In all but one study, multimodal models

outperformed unimodal models but the extent of improvement in prediction is highly variable. Vale Silva and Rohr's MultiSurv model showed greater performance on clinical data (c-index 0.809) than with their multimodal model (c-index 0.787, multiomics + WSI), or image (c-index 0.569) and -omic (c-index 0.758, mRNA) based models.⁵³ The relatively modest performance gain observed for most multimodal models over unimodal comparators suggest that WSI and -omic data may carry signals with significant overlap rather than distinct contributions. Indeed, morphological patterns associated with underlying mutational features are already well-established for some tumours. The technical variation of methods within these studies may also partly account for the variability of multimodal improvement observed across studies, with potential for significant noise introduction through patch sampling and processing methods or suboptimal alignment and weighting of different data modes through fusion approaches.

Other Performance Comparisons. Performance across cancer types was compared where there were results presented in greater than 5 studies, see Supplemental Figure S2. This synthesis used data from all datasets evaluated in the studies for the model of interest, including alternative dataset results from multi-cancer studies. In this descriptive comparison, variation in performance across different cancer types is observed. Higher c-index values observed for glioma and renal papillary cell carcinoma may reflect that, in these tumour types, both morphology and key molecular alterations are independently diagnostic and formally incorporated into WHO classification frameworks.^{79,80} Complementary histological features and underlying biology patterns likely provides a stronger signal for multimodal models to exploit. In contrast, in cancers with more heterogeneous genotype-phenotype relationships such as colorectal or breast cancer, the benefit of multimodal fusion may be in inherently limited.

The number of patients in the studies showed a weak trend towards increased performance with increased sample size. The presence of 400 patients or more appeared to be sufficient to achieve optimal results on internal test sets. Studies with fewer participants generally performed the worse (Supplemental Figure S3). There was no meaningful association between performance and increasing event rate, although only 15 studies reported data sufficiently to be included in this analysis (Supplemental Figure S4). These comparisons are illustrative only, due to heterogeneity in methods and risk of bias noted in this review.

Discussion

Our review highlights the rapid growth of this research domain in the last 5 years. Models have been developed across a wide spectrum of cancers, representing 19 organ systems and 23 specific diagnostic subtypes. Similar to previous works, we find that multimodal integration of WSI and -omic data, generally leads to improved prognosis prediction. Despite methodological progress, these models are far from ready for clinical evaluation.

All included studies were judged to be at high or unclear risk of bias, highlighting significant issues in study design, reporting, and validation. As a result of the significant heterogeneity and potential bias in these models, we are not able to more rigorously assess these models by meta-analysis. However, the range in reported performance is wide (c-index, 0.550-0.857) and the gains over unimodal models, where evaluated, are often small. The generally small increments in c-index imply that integration alone does not guarantee better prognostic discrimination, likely reflecting a combination of biological redundancy between WSI and -omic features, and methodological issues that may mask a true effect. As such, we remain uncertain about the potential clinical value of these models for the future.

These results also raise the prospect that multimodal data integration will not universally add value across the oncology domain, instead adding value in a cancer-specific way dictated by the underlying coupling of tissue morphology and molecular alterations within each cancer. However, the current evidence base is characterised by substantial heterogeneity in modelling approaches, predictor sets, and outcome definitions, which limits the feasibility of robust subgroup analyses within this single broad review. Future benchmarking across tumour types may help clarify where multimodal approaches offer the most meaningful advantage and should be further developed.

ML methods are an attractive prospect, able to flexibly handle high-dimensional, non-linear data. However, despite these perceived advantages many studies of ML prediction models in other domains have found no additional benefit of ML over traditional models.⁸ Given the expense of data generation, that is not part of routine clinical workflows (eg, transcriptomic data), the gain would need to be substantial to be used clinically. Model complexity is another factor in the cost of implementing these multimodal models and only 1 study in this review evaluated compute requirements of their model.³⁵ Future studies must compare models against standard clinical prediction methods and begin to evaluate the clinical utility and cost benefit of their models, and not just seek to outperform the discrimination metrics of competing models.

These shortcomings seem reflective of the immaturity of this field in which methodological exploration for handling complex image data and innovation of ML approaches in predictive modelling has been the focus so far.

DL-based models predominate in these studies, as does feature fusion of data. Some of the more recent DL models make use of graph networks, transformers and foundation models – keeping pace with DL developments in other fields. Whilst many studies develop approaches to capture cross-modal interactions and maximise the benefit of this complementary data, relatively fewer have explored models which are robust to missing data modalities – a common real-world scenario. Of concern, most studies lacked clinical contextualisation, with populations often defined by data availability rather than intended use. Important data issues, including assessment of the representativeness of the data for demographic and clinical features, and removal of missing data without consideration and evaluation of

whether this is missing at random or not, were almost universally overlooked. The well-known nature of TCGA used by most studies may be leading authors to assume knowledge from readers. This assumption, however, directly limits transparent reporting at the level of the prediction model studies themselves, constraining clear assessment of participant and predictor domains within risk-of-bias frameworks, regardless of the quality or documentation of the underlying source datasets. Yet, there are well-documented batch effects and participant shifts versus the general population for age, race and stage in this dataset, and more detail is important to the interpretation of these models.^{71-74,81-83} Despite reliance on public datasets, open science practices were inconsistently adopted, with limited code and data availability described.

The heavy dependence on the TCGA dataset, used in all studies, may be an indicator that we don't yet have the necessary data to build and test these models. The whole field may be at risk of overfitting to the features of a single source and creating an illusion of reproducibility. The oncology community need to build more easily accessible datasets with detailed meta data and appropriate longitudinal follow up to provide clear and independent data sets on which to validate models and establish their capacity to generalise beyond TCGA.

Limitations of the Review

The review protocol was designed to limit bias and maximise inclusion of eligible studies. While controlled vocabulary terms are available for several of the search concepts, the strategy relied primarily on free-text searching. Despite being supplemented by extensive citation searching, it is possible that some relevant studies were not retrieved. The review was restricted to studies published in English and as such may have missed studies in other languages. All stages of screening and data extraction were performed by 2 independent researchers, except for the initial duplicate screening which was performed by a single researcher (with the aid of *Rayyan* software¹⁶), raising the possibility of incorrectly excluding studies in error.

The significant heterogeneity between the studies in this review precluded more detailed data synthesis and meta-analysis which limited the conclusions that can be made. Due to the rapid expansion of this research area, a future review focussed on a specific cancer types or specific modelling methods may provide the opportunity for a more rigorous analysis.

Finally, this review focussed on models which directly integrated pathology and -omic data, overlooking studies outside this domain. Most relevant for future consideration may be the evaluation of models which additionally incorporate radiology data,^{84,85} replicating the current clinical paradigm, and models which used -omic and image data in a pipeline. Such approaches generally first generate an -omic prediction from the image, then generate an image-based prediction model.^{86,87} The rising availability and use of spatial -omic data may propel this field forward, bridging the gap between morphology and bulk -omic data for greater biological and interpretable insights.^{88,89}

Current Limitations and Future Recommendations

Our review finds that many aspects of methodology and reporting were suboptimal in these studies, in keeping with findings of other systematic reviews of prognostic models in the oncology domain.^{2,8,90} Whilst there is rapid proliferation of prognostic models in cancer, the translation rate from academic generation to clinical implementation is very low.⁸ The next phase of model development in this domain will benefit from reflections on lessons learnt about how to design clinically useful and usable models.⁷³ Future research in this area should adhere to reporting guidance, recently updated to better reflect complexities of ML-based modelling. TRIPOD + AI is a 27-item checklist designed to promote transparent reporting of studies developing, validating or updating prediction models.⁹¹ The PROBAST + AI tool used in this review also outlines best practice approaches to data selection and handling, modelling, and reporting.²⁰ Together, these frameworks highlight that methodological quality and transparent reporting are tightly linked and both are necessary for meaningful evaluation and clinical translation of prognostic models. Without better reporting, apparent progress in multimodal cancer prognosis prediction may be overstated.

Even for well-developed and validated models, clinical implementation is not realistic in the near future. Multimodal models are being developed in advance of the widespread clinical availability of high-throughput -omic and WSI data. In the UK, for example, approximately 10% of the molecular workload in the National Health Service is whole genome sequencing with the majority of the work centred on targeted mutation testing and smaller gene panels – there is no routine RNA sequencing.⁹² In parallel, digitisation of pathology services across the UK is not yet complete.⁹³ Understanding of where and when these models may provide additional benefit to patients from well-conducted research may inform how such services develop.

Finally, research in this field will not progress without wider use and availability of additional datasets. So far, generation of these complex data has largely been the reserve of publicly funded initiatives. Other pan cancer projects such as the Applied Proteogenomic Organisational Learning Outcomes (APOLLO) network⁹⁴ and the Cancer Moonshot Biobank,⁹⁵ are increasing the availability of publicly accessible -omic data with matched pathology and radiological imaging data. Genomics England's 100 000 Genomes Project Multimodal Programme is a large apply to access dataset also on the horizon.⁹⁶ However, existing "apply to access" datasets, such as the Children's Brain Tumour Tissue Consortium⁹⁷ and METABRIC (breast)⁹⁸ seem under used in this field so far. The reason for this is unclear and may be multifactorial. Critical to clinical implementation will be increased availability and use of datasets from more diverse populations (genetic ancestry, geographic, environmental) and datasets which are representative of both the medical diversity of a condition, and the technical diversity of the data generated.⁹⁹

Conclusion

This review of machine learning-based multimodal predictive models found 48 eligible studies predicting overall survival in cancer participants. This demonstrates significant growth in this research field since first works were identified in 2017. Progress has centred on methodological innovation, particularly through deep learning approaches. However, studies are limited by poor reporting, limited validation and lack of clinical contextualisation. To advance towards clinical impact, future research must prioritise transparent reporting, large diverse datasets, meaningful comparison to clinical standards, and a clear demonstration of how and where these models might be used in practice. Striking the right balance between model complexity, computational demands, and meaningful gains in predictive performance will be essential for real-world implementation—particularly in resource-constrained healthcare settings where cost-effectiveness is critical.

ORCID iDs

Charlotte Jennings  <https://orcid.org/0000-0003-1584-4057>
 Andrew Broad  <https://orcid.org/0000-0001-7131-6860>
 Lucy Godson  <https://orcid.org/0000-0002-3419-7628>
 Emily Clarke  <https://orcid.org/0000-0001-7106-9898>
 David Westhead  <https://orcid.org/0000-0002-0519-3820>
 Darren Treanor  <https://orcid.org/0000-0002-4579-484X>

Ethical Considerations

No ethical approval was required for this systematic review as all data used in the analysis were derived from publicly available articles.

Consent for Publication

Not applicable. The publication contains no individual participant data.

Author Contributions

Authors C.J., D.T. and D.W. planned study. C.J. conducted the searches. Abstracts were screened by C.J., E.C., L.G. Full texts articles were screened by C.J., E.C., L.G. Data extraction was performed by A.B., C.J., L.G. C.J. analysed the data and wrote the manuscript, which was revised by A.B., D.T., D.W., E.C., L.G. AI tools were not used in the generation or refinement of this manuscript. All authors approved the manuscript for publication.

Funding

This work was funded by the National Pathology Imaging Co-operative (NPIC). NPIC is supported by a £50m investment from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by UK Research and Innovation (UKRI, Project no. 104687). David R. Westhead is supported in part by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (BRC; NIHR203331). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

All data generated or analysed during this study are included in this published article and its supplementary information files. Extra data may be available from the corresponding author on request. No scientific data has been generated or modified using AI.

Supplemental Material

Supplemental material for this article is available online.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209-249.
2. Collins GS, Whittle R, Bullock GS, et al. Open science practices need substantial improvement in prognostic model studies in oncology using machine learning. *J Clin Epidemiol.* 2024;165:111199.
3. Mariotto AB, Noone AM, Howlader N, et al. Cancer survival: an overview of measures, uses, and interpretation. *J Natl Canc Inst Monogr.* 2014;2014(49):145-186.
4. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med.* 2019;25(1):24-29.
5. Jiang S, Suriawinata AA, Hassanpour S. MHAttnSurv: multi-head attention for survival prediction using whole-slide pathology images. *Comput Biol Med.* 2023;158:106883.
6. Yang Z, Wei T, Liang Y, et al. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *Nat Commun.* 2025;16(1):2366.
7. Tran D, Nguyen H, Pham VD, et al. A comprehensive review of cancer survival prediction using multi-omics integration and clinical variables. *Brief Bioinform.* 2025;26(2):bbaf150. doi:10.1093/bib/bbaf150
8. Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol.* 2022;22(1):101.
9. Schneider L, Laiouar-Pedari S, Kuntz S, et al. Integration of deep learning-based image analysis and genomic data in cancer pathology: a systematic review. *Eur J Cancer.* 2022;160:80-91.
10. Cui C, Yang H, Wang Y, et al. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Prog Biomed Eng.* 2023;5(2):022001.
11. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71. doi:10.1136/bmj.n71
12. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ.* 2020;368:l6890. doi:10.1136/bmj.l6890
13. Breen J, Allen K, Zucker K, et al. Artificial intelligence in ovarian cancer histopathology: a systematic review. *NPJ Precis Oncol.* 2023;7(1):83.
14. Haddaway NR, Grainger MJ, Gray CT. An R package and Shiny app for forward and backward citations chasing in academic searching. 2021. Accessed February 28, 2025. doi:10.5281/zenodo.4543513

15. Hirt J, Nordhausen T, Fuerst T, Ewald H, Appenzeller-Herzog C. Guidance on terminology, application, and reporting of citation searching: the TARCiS statement. *BMJ*. 2024; 385:e078384. doi:10.1136/bmj-2023-078384
16. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210.
17. Fernandez-Felix BM, López-Alcalde J, Roqué M, Muriel A, Zamora J. CHARMS and PROBAST at your fingertips: a template for data extraction and risk of bias assessment in systematic reviews of predictive models. *BMC Med Res Methodol*. 2023;23(1):44.
18. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
19. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58.
20. Moons KGM, Damen JAA, Kaul T, 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
21. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Sys Rev*. 2022;18(2):e1230.
22. Boehm KM, Aherne EA, Ellenson L, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer*. 2022; 3(6):723-733.
23. Cheerla A, Gevaert O. Deep learning with multimodal representation for pan cancer prognosis prediction. *Bioinformatics*. 2019;35(14):i446-i454.
24. Chen L, Zeng H, Zhang M, Luo Y, Ma X. Histopathological image and gene expression pattern analysis for predicting molecular features and prognosis of head and neck squamous cell carcinoma. *Cancer Med*. 2021;10(13):4615-4628.
25. Chen L, Zeng H, Xiang Y, Huang Y, Luo Y, Ma X. Histopathological images and multi-omics integration predict molecular characteristics and survival in lung adenocarcinoma. *Front Cell Dev Biol*. 2021;9:720110.
26. Chen RJ, Lu MY, Williamson DFK, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*. 2022;40(8):865-878.e6.
27. Chen RJ, Lu MY, Wang J, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging*. 2022;41(4):757-770.
28. Cheng J, Zhang J, Han Y, et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer Res*. 2017;77(21):e91-e100.
29. Hao J, Kosaraju SC, Tsaku NZ, Song DH, Kang M. PAGE-Net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. *Pac Symp Biocomput*. 2020;25:355-366.
30. Hou J, Jia X, Xie Y, Qin W. Integrative histology-genomic analysis predicts hepatocellular carcinoma prognosis using deep learning. *Genes*. 2022;13(10):1770.
31. Hou W, Lin C, Yu L, Qin J, Yu R, Wang L. Hybrid graph convolutional network with online masked autoencoder for robust multimodal cancer survival prediction. *IEEE Trans Med Imaging*. 2023;42(8):2462-2473.
32. Ji J, Liu Y, Bao Y, Men Y, Hui Z. Network analysis of histopathological image features and genomics data improving prognosis performance in clear cell renal cell carcinoma. *Urol Oncol Semin Original Investig*. 2024;42(8):249.
33. Li S, Shi H, Sui D, Hao A, Qin H. A novel pathological images and genomic data fusion framework for breast cancer survival prediction. In: *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, 2020, pp.1384-1387. IEEE. doi:10.1109/EMBC44109.2020.9176360
34. Li H, Chen L, Zeng H, Liao Q, Ji J, Ma X. Integrative analysis of histopathological images and genomic data in colon adenocarcinoma. *Front Oncol*. 2021;11:636451. doi:10.3389/fonc.2021.636451
35. Li R, Wu X, Li A, Wang M. HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics*. 2022;38(9):2587-2594.
36. Liu M, Liu Y, Cui H, Li C, Ma J. MGCT: mutual-guided cross-modality transformer for survival outcome prediction using integrative histopathology-genomic features. In: *2023 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2023, pp.1306-1312. IEEE. doi:10.1109/BIBM58861.2023.10385897
37. Liu H, Shi Y, Li A, Wang M. Multi-modal fusion network with intra- and inter-modality attention for prognosis prediction in breast cancer. *Comput Biol Med*. 2024;168:107796.
38. Lv Z, Lin Y, Yan R, Yang Z, Wang Y, Zhang F. PG-TFNet: transformer-based fusion network integrating pathological images and genomic data for cancer survival analysis. In: *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2021, pp.491-496. IEEE. doi:10.1109/BIBM52615.2021.9669445
39. Lv Z, Lin Y, Yan R, Wang Y, Zhang F. TransSurv: transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer. *IEEE/ACM Trans Comput Biol Bioinform*. 2023;20(6):3411-3420.
40. Ning Z, Pan W, Chen Y, et al. Integrative analysis of cross-modal features for the prognosis prediction of clear cell renal cell carcinoma. *Bioinformatics*. 2020;36(9):2888-2895.
41. Ning Z, Lin Z, Xiao Q, et al. Multi-constraint latent representation learning for prognosis analysis using multi-modal data. *IEEE Trans Neural Netw Learn Syst*. 2023;34(7):3737-3750.
42. Perez-Herrera LV, Sucre A, Sanchez XC, et al. AI-driven integration of histological and genomic data for understanding breast cancer survival. In: *2024 IEEE international symposium on biomedical imaging (ISBI)*, 2024, pp.1-5. IEEE. doi:10.1109/ISBI56570.2024.10635774
43. Qiu L, Zhao L, Zhao W, Zhao J. Dual-space disentangled-multimodal network (DDM-net) for glioma diagnosis and prognosis with incomplete pathology and genomic data. *Phys Med Biol*. 2024;69(8):085028.
44. Shao W, Han Z, Cheng J, et al. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. *IEEE Trans Med Imaging*. 2020;39(1):99-110.
45. Shao W, Liu J, Zuo Y, et al. FAM3L: feature-aware multimodal metric learning for integrative survival analysis of human cancers. *IEEE Trans Med Imaging*. 2023;42(9):2552-2565.
46. Shao W, Wang T, Sun L, et al. Multi-task multi-modal learning for joint diagnosis and prognosis of human cancers. *Med Image Anal*. 2020;65:101795.
47. Shao W, Zuo Y, Shi Y, et al. Characterizing the survival-associated interactions between tumor-infiltrating lymphocytes

- and tumors from pathological images and multi-omics data. *IEEE Trans Med Imaging*. 2023;42(10):3025-3035.
48. Steyaert S, Qiu YL, Zheng Y, Mukherjee P, Vogel H, Gevaert O. Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. *Commun Med*. 2023;3(1):44.
 49. Subramanian V, Syeda-Mahmood T, Do MN. Multimodal fusion using sparse CCA for breast cancer survival prediction. In: *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, 2021, pp.1429-1432. IEEE. doi:10.1109/ISBI48211.2021.9434033
 50. Subramanian V, Syeda-Mahmood T, Do MN. Modelling-based joint embedding of histology and genomics using canonical correlation analysis for breast cancer survival prediction. *Artif Intell Med*. 2024;149:102787.
 51. Sun D, Li A, Tang B, Wang M. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Comput Methods Programs Biomed*. 2018;161:45-53.
 52. Tan K, Huang W, Liu X, Hu J, Dong S. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. *Artif Intell Med*. 2022;126:102260.
 53. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. *Sci Rep*. 2021;11(1):13505.
 54. Vollmer A, Hartmann S, Vollmer M, et al. Multimodal artificial intelligence-based pathogenomics improves survival prediction in oral squamous cell carcinoma. *Sci Rep*. 2024;14(1):5687.
 55. Wang Z, Li R, Wang M, Li A. GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics*. 2021;37(18):2963-2970.
 56. Wang S, Hu X, Zhang Q. HC-MAE: hierarchical cross-attention masked autoencoder integrating histopathological images and multi-omics for cancer survival prediction. In: *2023 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2023, pp. 642-647. IEEE. doi:10.1109/BIBM58861.2023.10385635
 57. Wei T, Yuan X, Gao R, et al. Survival prediction of stomach cancer using expression data and deep learning models with histopathological images. *Cancer Sci*. 2023;114(2):690-701.
 58. Wu X, Shi Y, Wang M, Li A. CAMR: cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics*. 2023;39(1):btad025. doi:10.1093/bioinformatics/btad025
 59. Xie Y, Sang Q, Da Q, et al. Improving diagnosis and outcome prediction of gastric cancer via multimodal learning using whole slide pathological images and gene expression. *Artif Intell Med*. 2024;152:102871.
 60. Zeng H, Chen L, Huang Y, Luo Y, Ma X. Integrative models of histopathological image features and omics data predict survival in head and neck squamous cell carcinoma. *Front Cell Dev Biol*. 2020;8:553099.
 61. Zeng H, Chen L, Zhang M, Luo Y, Ma X. Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer. *Gynecol Oncol*. 2021;163(1):171-180.
 62. Zhan Z, Jing Z, He B, et al. Two-stage Cox-nnet: biologically interpretable neural-network model for prognosis prediction and its application in liver cancer survival using histopathology and transcriptomic data. *NAR Genom Bioinform*. 2021;3(1):015.
 63. Zhang Y, Li A, He J, Wang M. A novel MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data. *IEEE J Biomed Health Inform*. 2020;24(1):171-179.
 64. Zhao Z, Feng Q, Zhang Y, Ning Z. Adaptive risk-aware sharable and individual subspace learning for cancer survival analysis with multi-modality data. *Brief Bioinform*. 2023;24(1):bbac489. pii: bbac489. doi:10.1093/bib/bbac489
 65. Zheng Y, Conrad RD, Green EJ, et al. Graph attention-based fusion of pathology images and gene expression for prediction of cancer survival. *IEEE Trans Med Imaging*. 2024;43(9):3085-3097.
 66. Zhou F, Chen H. Cross-modal translation and alignment for survival analysis. In: *2023 IEEE/CVF international conference on computer vision (ICCV)*, 2023, pp.21428-21437. IEEE. doi:10.1109/ICCV51070.2023.01964
 67. Zhou J, Foroughi-Pour A, Deirawan H, et al. Integrative deep learning analysis improves colon adenocarcinoma patient stratification at risk for mortality. *EBioMedicine*. 2023;94:104726.
 68. Zhou C, Wang H, Li X, Liu W, Liu Z. Multimodal survival ensemble network: integrating genomic and histopathological insights for enhanced cancer prognosis. In: *ICASSP 2024 - 2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2024, pp.2330-2334. IEEE. doi:10.1109/ICASSP48485.2024.10447524
 69. Zhu W, Chen Y, Nie S, Yang H. SAMMS: multi-modality deep learning with the foundation model for the prediction of cancer patient survival. In: *2023 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2023, pp.3662-3668. IEEE. doi:10.1109/BIBM58861.2023.10385661
 70. NIDCR. Oral cancer incidence (new cases) by age, race and sex. National Institute of Dental and Craniofacial Research. 2025. Accessed June 19, 2025. <https://www.nidcr.nih.gov/research/data-statistics/oral-cancer/incidence>
 71. Dehkharghanian T, Bidgoli AA, Riasatian A, et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagn Pathol*. 2023;18(1):67.
 72. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun*. 2021;12(1):4423.
 73. Rasnic R, Brandes N, Zuk O, Linial M. Substantial batch effects in TCGA exome sequences undermine pan-cancer analysis of germline variants. *BMC Cancer*. 2019;19(1):783.
 74. Ibing S, Michels BE, Mosdzien M, Meyer HR, Feuerbach L, Körner C. On the impact of batch effect correction in TCGA isomiR expression data. *NAR Cancer*. 2021;3(1):007.
 75. Choi JH, Hong SE, Woo HG. Pan-cancer analysis of systematic batch effects on somatic sequence variations. *BMC Bioinformatics*. 2017;18(1):211.
 76. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008.
 77. Riley RD, Ensor J, Snell KIE, et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *Lancet Digit Health*. 2025;7:100857. doi:10.1016/j.landig.2025.01.013
 78. Liu J, Lichtenberg T, Hoadley KA, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cells*. 2018;173(2):400-416.
 79. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neurooncol*. 2021;23(8):1231-1251.
 80. Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM. The 2016 WHO classification of tumours of the urinary

- system and male genital organs-part A: renal, penile, and testicular tumours. *Eur Urol.* 2016;70(1):93-105.
81. Spratt DE, Chan T, Waldron L, et al. Racial/ethnic disparities in genomic sequencing. *JAMA Oncol.* 2016;2(8):1070-1074.
 82. Wang X, Steensma JT, Bailey MH, Feng Q, Padda H, Johnson KJ. Characteristics of the cancer genome atlas cases relative to U.S. general population cancer cases. *Br J Cancer.* 2018;119(7):885-892.
 83. Schmitt M, Maron RC, Hekler A, et al. Hidden variables in deep learning digital pathology and their potential to cause batch effects: prediction model study. *J Med Internet Res.* 2021;23(2):e23436.
 84. Schulz S, Woerl AC, Jungmann F, et al. Multimodal deep learning for prognosis prediction in renal cancer. *Front Oncol.* 2021;11:788740.
 85. Yuan Y, Zhang X, Wang Y, et al. Multimodal data integration using deep learning predicts overall survival of patients with glioma. *VIEW.* 2024;5(5):20240001. doi:10.1002/VIW.20240001
 86. Levy-Jurgenson A, Tekpli X, Kristensen VN, Yakhini Z. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Sci Rep.* 2020;10(1):18802.
 87. Gao R, Yuan X, Ma Y, et al. Harnessing TME depicted by histological images to improve cancer prognosis through a deep learning system. *Cell Rep Med.* 2024;5(5):101536.
 88. Marx V. Method of the year: spatially resolved transcriptomics. *Nat Methods.* 2021;18(1):9-14.
 89. Nature Methods Editorial. Method of the year 2024: spatial proteomics. *Nat Methods.* 2024;21(12):2195-2196. doi:10.1038/s41592-024-02565-3
 90. Dhiman P, Ma J, Andaur Navarro CL, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. *Diagn Prognost Res.* 2022;6(1):13.
 91. Collins GS, Moons KGM, Dhiman P, 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
 92. NHS England. Accelerating genomic medicine in the NHS. Updated October 12, 2022. Accessed June 26, 2025. <https://www.england.nhs.uk/long-read/accelerating-genomic-medicine-in-the-nhs/>
 93. Moxley-Wyles B, Colling R. Artificial intelligence and digital pathology: where are we now and what are the implementation barriers? *Diagn Histopathol.* 2024;30(11):597-603.
 94. Office of Cancer Clinical Proteomics Research, CBIIT. The applied proteogenomics organizational learning and outcomes (APOLLO) network. Updated February 18, 2025. Accessed June 26, 2025. https://datascience.cancer.gov/collaborations/applied-proteogenomics-organizational-learning-and-outcomes-apollo-network?utm_source=chatgpt.com
 95. National Cancer Institute Cancer Moonshot Biobank. The cancer moonshot biobank. Accessed June 26, 2025. <https://moonshotbiobank.cancer.gov>
 96. Jennings CN, Humphries MP, Wood S, et al. Bridging the gap with the UK genomics pathology imaging collection. *Nat Med.* 2022;28(6):1107-1108.
 97. Lilly JV, Rokita JL, Mason JL, et al. The children's brain tumor network (CBTN) - accelerating research in pediatric central nervous system tumors through collaboration and open science. *Neoplasia.* 2023;35:100846.
 98. European Genome Phenome Archive. Molecular taxonomy of breast cancer international consortium (METABRIC). 2012. Accessed June 26, 2025. <https://ega-archive.org/studies/EGAS000000000098>
 99. Raven-Adams MC, Hernandez-Boussard T, Joly Y, et al. Defining and pursuing diversity in human genetic studies. *Nat Genet.* 2024;56(10):1985-1988.