Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prognostic value of deep-learning computational pathology in colorectal cancer: a systematic review of haematoxylin and eosin whole-slide image studies

Yinkai Xu
The First Affiliated Hospital of Soochow University
Liping Zhang
The Affiliated Suzhou Hospital of Nanjing Medical University (Suzhou Municipal Hospital)
Chenao Wang
The First Affiliated Hospital of Soochow University
Yizhou Yao
The First Affiliated Hospital of Soochow University
Chen Shen
The Fourth Affiliated Hospital of Soochow University
Wen Gu
wen981194@gmail.com

The First Affiliated Hospital of Soochow University

Systematic Review

Keywords: colorectal cancer, deep learning, computational pathology, whole-slide imaging, prognosis, hazard ratio, systematic review

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11297991/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Deep-learning computational pathology applied to routine haematoxylin and eosin whole-slide images may provide prognostic information in colorectal cancer. This systematic review evaluated the association between image-derived risk groups and survival.

Methods

PubMed and Web of Science Core Collection were searched from inception to 21 June 2026. Studies of colorectal cancer reporting a high-versus-low-risk hazard ratio derived from haematoxylin and eosin images were included. The most-adjusted estimate was selected, with preference for external validation. Random-effects meta-analysis, subgroup and leave-one-out analyses were performed. The recovered risk-of-bias panel reports Quality in Prognosis Studies domain judgements. The review was registered in PROSPERO (CRD420261429467).

Results

Across 27 studies, the pooled hazard ratio was 2.49 (95% confidence interval 2.12–2.93), with substantial heterogeneity (I^2 = 77%) and a 95% prediction interval of 1.25–4.97. Pooled estimates were 2.36 in externally validated studies (17 studies) and 2.74 in internally validated studies (10 studies); the difference was not statistically significant. Leave-one-out estimates ranged from 2.35 to 2.57.

Conclusions

Image-derived risk scores were associated with survival in colorectal cancer. Heterogeneity, differences in risk definitions and possible overlap between validation cohorts limit interpretation of a single pooled estimate. Independent prospective validation and evaluation of clinical utility are needed before these methods can guide treatment decisions.

Background

Prognostic assessment informs treatment and surveillance decisions in colorectal cancer (CRC). Deep-learning (DL) computational pathology has been investigated as a means of obtaining additional prognostic information from routinely prepared haematoxylin and eosin (H&E) tissue sections [1–4].

Published approaches quantify tumour–stroma composition, immune infiltration and tissue architecture or learn image representations directly from whole-slide images. Their reported associations with outcome have motivated evaluation across different cohorts and clinical settings [1–11].

The expanding range of computational-pathology models creates a need to synthesise their reported prognostic associations and examine differences in validation practice. We conducted a systematic review with quantitative synthesis of hazard ratios (HRs) for high versus low image-derived risk in CRC. The objectives were to estimate the pooled association with survival, explore heterogeneity by survival endpoint and validation type, and assess risk of bias. A risk-group HR describes relative prognosis; it does not, by itself, establish individual predictive accuracy or the benefit of treatment guided by that risk score.

Methods

Protocol and registration

The review was registered in PROSPERO (CRD420261429467) and was prepared with reference to PRISMA 2020 [12]. Reporting considerations for prediction models were informed by TRIPOD+AI [13]. The reporting checklist accompanies this manuscript.

Eligibility criteria

Using a PICOTS framework, we included studies of patients with CRC in which a DL or computational-pathology risk score derived from H&E WSIs stratified patients by risk and reported a hazard ratio for survival (OS, DSS, DFS, RFS or PFS). We excluded radiomics-only (CT/MRI) and molecular-only models, multistain models for the primary analysis, and studies without an estimable high-versus-low HR (e.g., reporting only c-index, AUC or log-rank). No date or language restriction was applied.

Information sources and search

PubMed and Web of Science Core Collection were searched from inception to 21 June 2026, combining CRC, DL/computational pathology and prognosis/survival concepts, with supplementary citation searching. Embase and Cochrane CENTRAL were not accessible to the review team. The recovered search strings for the two databases are provided in Additional file 1 (Section S1).

Study selection and data extraction

Records were screened against the eligibility criteria, and candidate reports were assessed in full text. The study-level data comprised the most-adjusted HR and 95% CI, survival endpoint, validation setting and model type. Where estimates were available for several cohorts, an external-validation estimate was preferred. Comparisons were standardised as high versus low image-derived risk.

Risk of bias

Risk of bias was assessed with QUIPS [14] and, for prediction-model studies, PROBAST [15], with attention to patient-level train/test separation, external validation and pre-specification of the risk cut-off.

Statistical analysis

Log hazard ratios were pooled by random-effects meta-analysis using inverse-variance weighting [16] and reported as pooled HR with 95% confidence and prediction intervals; heterogeneity was quantified by τ^2 and I^2. Pre-specified subgroups (outcome, validation type) and leave-one-out diagnostics were performed, and the external-versus-internal difference was tested on the log scale. Small-study effects were assessed with Egger's test [17]. Analyses were performed in Python using a DerSimonian–Laird random-effects implementation, with an R/metafor script for cross-checking [18].

Results

Study selection and characteristics

The two-database search returned 1,116 records (PubMed 473; Web of Science Core Collection 643); after removing duplicates, 1,025 titles/abstracts were screened and candidate reports were assessed in full text; 27 studies met the eligibility criteria for the primary analysis (Fig. 1 and Table 1). The most frequent reason for exclusion at full text was the absence of an estimable high-versus-low hazard ratio (studies reporting only a concordance index, AUC or log-rank statistic), followed by non-H&E or multimodal inputs and overlapping cohorts. Studies were published between 2018 and 2026 and originated from Europe, Asia, North America and Australia; outcomes included overall (OS), disease-specific (DSS/CSS), disease-free (DFS), recurrence-free (RFS) and progression-free survival (PFS), and seventeen studies provided externally validated estimates. Models ranged from convolutional neural networks and attention-based multiple-instance learning to tumour–stroma-ratio and foundation-model approaches.

Pooled prognostic effect

Across 27 studies, the pooled HR for high versus low DL-defined risk was 2.49 (95% CI 2.12–2.93), with high heterogeneity (I^2 = 77%) and a 95% prediction interval of 1.25–4.97 (Fig. 2; Table 2).

Heterogeneity, subgroups and small-study effects

In pre-specified subgroups (Fig. 4; Table 2), the pooled effect was numerically lower under external validation (HR 2.36, 95% CI 1.96–2.85; k = 17) than under internal/training validation (HR 2.74, 95% CI 2.02–3.72; k = 10), but this difference did not reach statistical significance (secondary test statistics require reconciliation). Effects were similar across endpoints—OS 2.27 (k = 13), CSS/DSS 2.76 (k = 6)

and DFS/RFS/PFS 2.68 (k = 8). Leave-one-out analysis confirmed robustness, with the pooled HR remaining between 2.35 and 2.57. Egger's regression test showed no significant funnel asymmetry (intercept 1.02; p value requires reconciliation; Fig. 3), although power to detect small-study effects remained limited.

Risk of bias

Risk of bias was reported (QUIPS, with PROBAST for prediction-model studies; Fig. 5). Overall risk of bias was low in 6 studies, moderate in 18 and high in 3. The statistical-analysis and study-participation domains were rated moderate most frequently, driven by data-driven risk-cut-off optimisation, internal-only validation and small event counts (e.g., very wide confidence intervals in two studies), while study-attrition reporting was often incomplete; the prognostic-factor (DL score) and outcome-measurement domains were generally low risk. Under PROBAST, prediction-model studies were more often rated high risk in the analysis domain, reflecting optimism from cut-off selection and limited external validation.

Discussion

Across 27 studies, higher risk assigned by DL/computational-pathology methods was associated with poorer survival (pooled HR 2.49, 95% CI 2.12–2.93). The included approaches ranged from automated tumour–stroma quantification and gland morphology to multiple-instance learning and foundation models. Most individual HRs were approximately 1.4–3.9, although two estimates exceeded 8. These findings support an association between image-derived risk and outcome, while the diversity of models and endpoints limits claims that their prognostic performance is equivalent.

Several studies evaluated stromal features or the tumour–stroma ratio. These interpretable measures provide a plausible link between tissue morphology and outcome and may aid pathological interpretation of model outputs. The present synthesis did not compare algorithms on a shared cohort or establish which tissue features account for the pooled association. It therefore cannot demonstrate that stromal information explains all DL predictions or exclude confounding by staining, scanning or cohort characteristics.

The pooled HR was 2.36 in studies with external validation and 2.74 in studies with internal or training validation. A non-significant between-group comparison does not establish equivalence or demonstrate that internal optimism has been eliminated. Reuse of public cohorts, including TCGA, MCO and PLCO, may also introduce dependence across study estimates. The wide prediction interval (1.25–4.97) further indicates that the magnitude of association may vary considerably in a new setting.

Heterogeneity was substantial overall (I^2 = 77%). It was lower for cancer-/disease-specific survival (I^2 = 52%) and higher for recurrence/progression endpoints (I^2 = 87%). Differences in patient populations, survival endpoints, adjustment variables, risk thresholds and validation designs may contribute to this variation. The two largest HRs in Table 1 belong to different endpoint groups (RFS and DSS), illustrating

the need to interpret extreme estimates within their endpoint-specific contexts. Leave-one-out estimates remained between 2.35 and 2.57, but this stability does not resolve dependence between overlapping cohorts or bias shared by multiple studies.

Requiring an estimable high-versus-low-risk HR also shapes the evidence included in this synthesis. Studies reporting only discrimination measures could not contribute to the primary analysis. HRs, discrimination, calibration and clinical net benefit answer different questions, and no single measure establishes clinical usefulness. Reporting prespecified risk-group contrasts alongside discrimination and calibration would improve interpretation and future synthesis; the risk threshold, adjustment covariates and validation setting should be reported transparently.

For diagnostic pathology, the potential value lies in extracting reproducible prognostic information from routinely acquired tissue images. Implementation nevertheless requires slide digitisation, computational infrastructure, quality control and external validation; this review does not establish cost-effectiveness. Prospective studies should determine whether a model adds useful information to conventional clinicopathological assessment and whether decisions guided by the model improve outcomes. Retrospective prognostic associations alone cannot identify patients who benefit from treatment intensification.

The review has several limitations. Only two bibliographic databases were searched, so relevant studies indexed elsewhere may have been missed. Differences in endpoints and risk thresholds complicate the interpretation of a single pooled HR. Shared public cohorts may violate the independence assumed by conventional pooling methods. The recovered reporting materials do not include a formal certainty-of-evidence assessment. Prognostic association should therefore be distinguished from individual prediction, clinical utility and treatment benefit.

Conclusions

DL and computational-pathology risk scores derived from H&E images were associated with survival in CRC across the included studies. The pooled estimate should be interpreted in the context of substantial heterogeneity, variable validation practices and possible cohort overlap. Standardised reporting, independent prospective validation and evaluation of incremental clinical utility are needed before adoption in routine pathological risk assessment.

Abbreviations

AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; CI, confidence interval; CNN, convolutional neural network; CRC, colorectal cancer; CSS, cancer-specific survival; DFS, disease-free survival; DL, deep learning; DSS, disease-specific survival; H&E, haematoxylin and eosin; HR, hazard ratio; OS, overall survival; PFS, progression-free survival; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROBAST, Prediction model Risk Of Bias ASsessment Tool;

QUIPS, Quality In Prognosis Studies; RFS, recurrence-free survival; TSR, tumour–stroma ratio; WSI, whole-slide image.

Declarations

Availability of data and materials

The study-level HRs and confidence intervals recovered from the original submission are provided in Additional file 1 (Table S1), together with the reported database search strings. The analysis uses published aggregate study data and does not contain a newly collected institutional patient cohort. The review registration is PROSPERO CRD420261429467.

Authors’ contributions

YX, LZ and CW wrote the original draft, checked the study materials and results, and performed the statistical analyses. YY, CS and WG designed the study and critically reviewed the research and manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable. This study is a systematic review of previously published aggregate data; no new human participants, human data or animal data were collected, so ethical approval and informed consent were not required.

Consent for publication

Not applicable.

Acknowledgements

Not applicable.

References

1. Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. The Lancet. 2020;395(10221):350-360. doi:10.1016/s0140-6736(19)32998-8.
2. Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis CA, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLOS Medicine. 2019;16(1):e1002730. doi:10.1371/journal.pmed.1002730.
3. Zhao K, Li Z, Yao S, Wang Y, Wu X, Xu Z, et al. Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. EBioMedicine. 2020;61:103054. doi:10.1016/j.ebiom.2020.103054.
4. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific Reports. 2018;8(1):3395. doi:10.1038/s41598-018-21758-3.
5. Jiang W, Wang H, Dong X, Yu X, Zhao Y, Chen D, et al. Pathomics Signature for Prognosis and Chemotherapy Benefits in Stage III Colon Cancer. JAMA Surgery. 2024;159(5):519. doi:10.1001/jamasurg.2023.8015.
6. Sajjad U, Akbar AR, Su Z, Leyva A, Knight D, Frankel WL, et al. Morphology-Aware Prognostic Model for Five-Year Survival Prediction in Colorectal Cancer from H&E Whole-Slide Images: A Study Using Multi-Center Clinical Trial Cohort. Cancers. 2026;18(7):1150. doi:10.3390/cancers18071150.
7. Xiao H, Weng Z, Sun K, Shen J, Lin J, Chen S, et al. Predicting 5-year recurrence risk in colorectal cancer: development and validation of a histology-based deep learning approach. British Journal of Cancer. 2024;130(6):951-960. doi:10.1038/s41416-024-02573-2.
8. Zhao K, Wu L, Huang Y, Yao S, Xu Z, Lin H, et al. Deep learning quantified mucus-tumor ratio predicting survival of patients with colorectal cancer using whole-slide images. Precision Clinical Medicine. 2021;4(1):17-24. doi:10.1093/pcmedi/pbab002.
9. Geessink OGF, Baidoshvili A, Klaase JM, Ehteshami Bejnordi B, Litjens GJS, van Pelt GW, et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. Cellular Oncology. 2019;42(3):331-341. doi:10.1007/s13402-019-00429-z.
10. Zhou X, Lu Y, Wu Y, Yu Y, Liu Y, Wang C, et al. Construction and validation of a deep learning prognostic model based on digital pathology images of stage III colorectal cancer. European Journal of Surgical Oncology. 2024;50(7):108369. doi:10.1016/j.ejso.2024.108369.
11. Jakab A, Patai ÁV, Micsik T. Digital image analysis provides robust tissue microenvironment-based prognosticators in patients with stage I-IV colorectal cancer. Human Pathology. 2022;128:141-151. doi:10.1016/j.humpath.2022.07.003.

12. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
13. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
14. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing Bias in Studies of Prognostic Factors. Annals of Internal Medicine. 2013;158(4):280-286. doi:10.7326/0003-4819-158-4-201302190-00009.
15. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal Medicine. 2019;170(1):51-58. doi:10.7326/m18-1376.
16. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986;7(3):177-188. doi:10.1016/0197-2456(86)90046-2.
17. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-634. doi:10.1136/bmj.315.7109.629.
18. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software. 2010;36(3):1-48. doi:10.18637/jss.v036.i03.
19. Wei B, Li L, Feng Y, Liu S, Fu P, Tian L. Exploring prognostic biomarkers in pathological images of colorectal cancer patients via deep learning. The Journal of Pathology: Clinical Research. 2024;10(6):e70003. doi:10.1002/2056-4538.70003.
20. Lou S, Huang Y, Du F, Xue J, Mo G, Li H, et al. Development and validation of a deep learning-based pathomics signature for prognosis and chemotherapy benefits in colorectal cancer: a retrospective multicenter cohort study. Frontiers in Immunology. 2025;16:1602909. doi:10.3389/fimmu.2025.1602909.
21. Millward J, He Z, Nibali A, Mouradov D, Mielke LA, Tran K, et al. Automated deep learning-based assessment of tumour-infiltrating lymphocyte density determines prognosis in colorectal cancer. Journal of Translational Medicine. 2025;23(1):298. doi:10.1186/s12967-025-06254-3.
22. Loeffler CML, Bando H, Sainath S, Muti HS, Jiang X, van Treeck M, et al. HIBRID: histology-based risk-stratification with deep learning and ctDNA in colorectal cancer. Nature Communications. 2025;16(1):7561. doi:10.1038/s41467-025-62910-8.
23. Li X, Jonnagaddala J, Yang S, Zhang H, Xu XS. A retrospective analysis using deep-learning models for prediction of survival outcome and benefit of adjuvant chemotherapy in stage II/III colorectal cancer. Journal of Cancer Research and Clinical Oncology. 2022;148(8):1955-1963. doi:10.1007/s00432-022-03976-5.
24. Jiang X, Hoffmeister M, Brenner H, Muti HS, Yuan T, Foersch S, et al. End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. The Lancet Digital Health. 2024;6(1):e33-e43. doi:10.1016/s2589-7500(23)00208-x.

25. Zhao Z, Chen D, Wang R, Zhang X, Wen X, Zheng X, et al. Multiview deep-learning-enabled histopathology for prognostic and therapeutic stratification in stage II colorectal cancer: A retrospective multicenter study. PLOS Medicine. 2026;23(1):e1004614. doi:10.1371/journal.pmed.1004614.
26. Ji MY, Yuan L, Lu SM, Gao MT, Zeng Z, Zhan N, et al. Glandular orientation and shape determined by computational pathology could identify aggressive tumor for early colon carcinoma: a triple-center study. Journal of Translational Medicine. 2020;18(1):129. doi:10.1186/s12967-020-02297-w.
27. Ye H, Zhao K, Cui Y, Li Z, Zhang H, Zhong ME, et al. Artificial intelligence-based tumor-stroma ratio quantification reveals prognostic value and stromal-driven immunosuppression in colorectal cancer: an international validation study. Journal of Translational Medicine. 2026;24(1):269. doi:10.1186/s12967-026-07681-6.
28. Jiao Y, Li J, Qian C, Fei S. Deep learning-based tumor microenvironment analysis in colon adenocarcinoma histopathological whole-slide images. Computer Methods and Programs in Biomedicine. 2021;204:106047. doi:10.1016/j.cmpb.2021.106047.
29. Öztürk SK, Bokhorst JM, Baumann E, Sheahan K, van de Velde CJH, Marijnen CAM, et al. Exploring Intratumoral Budding in Colorectal Cancer Using Computational Pathology: A Biopsy-Based Evaluation. Modern Pathology. 2025;38(2):100655. doi:10.1016/j.modpat.2024.100655.
30. Carvalho R, Zander T, Barroso VM, Bekisoglu A, Zerbe N, Klein S, et al. AI-based tumor-stroma ratio quantification algorithm: comprehensive evaluation of prognostic role in primary colorectal cancer. Virchows Archiv. 2025;487(4):799-813. doi:10.1007/s00428-025-04048-y.
31. Heilijgers F, Polack M, Roodvoets AGH, Meershoek - Klein Kranenbarg E, Peeters KCMJ, Buettner R, et al. Automating tumor–stroma ratio quantification in colon cancer patients from the UNITED study. ESMO Open. 2026;11(1):105934. doi:10.1016/j.esmoop.2025.105934.
32. Wu C, Pai RK, Kosiorek H, Banerjee I, Pfeiffer A, Hagen CE, et al. Improved Risk-Stratification Scheme for Mismatch-Repair Proficient Stage II Colorectal Cancers Using the Digital Pathology Biomarker QuantCRC. Clinical Cancer Research. 2024;30(9):1811-1821. doi:10.1158/1078-0432.ccr-23-3211.
33. Sinicrope FA, Nelson GD, Saberzadeh-Ardestani B, Segovia DI, Graham RP, Wu C, et al. Use of Deep Learning to Evaluate Tumor Microenvironmental Features for Prediction of Colon Cancer Recurrence. Cancer Research Communications. 2024;4(5):1344-1350. doi:10.1158/2767-9764.crc-24-0031.
34. Wang X, Jiang Y, Yang S, Wang F, Zhang X, Wang W, et al. Foundation Model for Predicting Prognosis and Adjuvant Therapy Benefit From Digital Pathology in GI Cancers. Journal of Clinical Oncology. 2025;43(32):3468-3481. doi:10.1200/jco-24-01501.

Tables

Table 1. Characteristics of the 27 included studies. HR, hazard ratio; CI, confidence interval; Ext, external validation; Int, internal validation. Survival endpoint abbreviations are defined in the text.

Study (author, year)	Journal	DL approach	Valid.	Endpoint	HR (95% CI)
Skrede, 2020 [1]	Lancet	DoMore-v1 (CNN)	Ext	CSS	3.04 (2.07–4.47)
Kather, 2019 [2]	PLoS Med	deep stroma score	Ext	OS	1.63 (1.14–2.33)
Zhao K, 2020 [3]	eBioMedicine	AI tumour-stroma ratio	Ext	OS	2.08 (1.26–3.42)
Bychkov, 2018 [4]	Sci Rep	CNN–RNN tiles	Int	DSS	1.89 (1.41–2.53)
Jiang W, 2024 [5]	JAMA Surg	pathomics signature (stage III)	Int	OS	2.78 (2.34–3.31)
Wei B, 2024 [19]	J Pathol Clin Res	CRCRS	Ext	OS	2.36 (1.47–3.79)
Lou, 2025 [20]	Front Immunol	pathomics signature	Ext	OS	3.26 (2.36–4.51)
Millward, 2025 [21]	J Transl Med	TIL density (DL)	Ext	OS	1.37 (1.13–1.66)
Loeffler, 2025 [22]	Nat Commun	HIBRID (DL)	Ext	DFS	2.31 (1.86–2.86)
Li X, 2022 [23]	J Cancer Res Clin Oncol	CRCNet	Ext	OS	2.76 (1.17–6.50)
Jiang X, 2024 [24]	Lancet Digit Health	end-to-end DL	Ext	DSS	3.08 (2.44–3.89)
Zhao Z, 2026 [25]	PLoS Med	SurvFinder (multiview)	Int	RFS	8.23 (5.43–12.47)
Sajjad, 2026 [6]	Cancers	PRISM	Int	OS	3.21 (2.18–4.72)

Study (author, year)	Journal	DL approach	Valid.	Endpoint	HR (95% CI)
Xiao H, 2024 [7]	Br J Cancer	DL-RRS	Ext	RFS	3.89 (2.51–6.03)
Zhao K, 2021 [8]	Precis Clin Med	mucus-tumour ratio	Ext	OS	1.96 (1.12–3.42)
Geessink, 2019 [9]	Cell Oncol	automated TSR	Int	DSS	2.48 (1.29–4.78)
Zhou X, 2024 [10]	Eur J Surg Oncol	tumour risk signature	Ext	PFS	3.63 (1.13–11.66)
Jakab, 2022 [11]	Hum Pathol	automated TSR	Int	OS	2.00 (1.15–3.51)
Ji M, 2020 [26]	J Transl Med	gland morphology (computational)	Int	DSS	9.65 (2.15–43.12)
Ye H, 2026 [27]	J Transl Med	AI tumour-stroma ratio	Ext	OS	2.27 (1.57–3.28)
Jiao Y, 2021 [28]	Comput Methods Programs Biomed	DL tumour-microenvironment	Int	PFS	1.67 (1.11–2.50)
Öztürk, 2025 [29]	Mod Pathol	DL tumour budding	Ext	OS	3.50 (1.10–10.80)
Carvalho, 2025 [30]	Virchows Arch	AI tumour-stroma ratio	Int	OS	2.30 (1.10–5.10)
Heilijgers, 2025 [31]	ESMO Open	automated TSR (UNITED)	Ext	DFS	1.45 (1.07–1.95)
Wu, 2024 [32]	Clin Cancer Res	QuantCRC (digital pathology)	Ext	RFS	2.35 (1.15–4.82)
Sinicrope, 2024 [33]	Cancer Res Commun	DL tumour-microenvironment (TSR)	Int	RFS	2.02 (1.14–3.57)

Study (author, year)	Journal	DL approach	Valid.	Endpoint	HR (95% CI)
Jiang, 2025 [34]	J Clin Oncol	pan-cancer foundation model (DL)	Ext	DSS	3.15 (1.91–5.21)

Study labels reproduce the original extraction. The label “Jiang, 2025” appears to correspond to Wang et al. [34] and requires source-level confirmation; “Heilijgers, 2025” refers to the article published in ESMO Open volume 11 (2026) [31]. These labels are retained here to allow comparison with the original figures.

Table 2. Pooled hazard ratios, overall and by subgroup.

Analysis	k	Pooled HR (95% CI)	I^2 (%)
Overall	27	2.49 (2.12–2.93)	77
Overall survival (OS)	13	2.27 (1.84–2.80)	73
Cancer-/disease-specific (CSS/DSS)	6	2.76 (2.14–3.55)	52
DFS / RFS / PFS	8	2.68 (1.77–4.04)	87
External validation	17	2.36 (1.96–2.85)	74
Internal / training	10	2.74 (2.02–3.72)	80

Figures

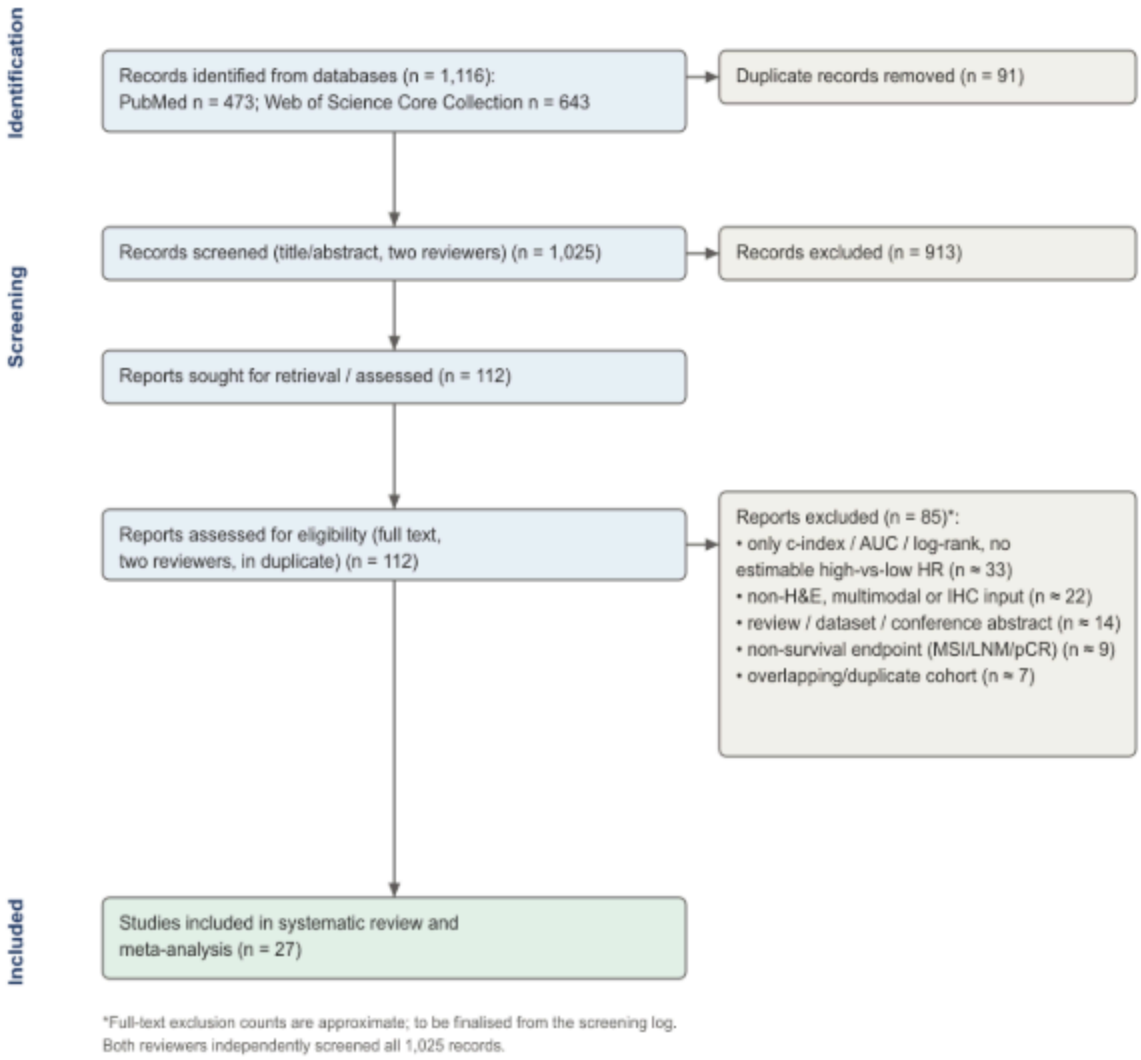


Figure 1

Study selection flow diagram. The original flow diagram is retained in this author-review draft. Full-text exclusion reasons and counts remain approximate and require reconciliation against the final screening log.

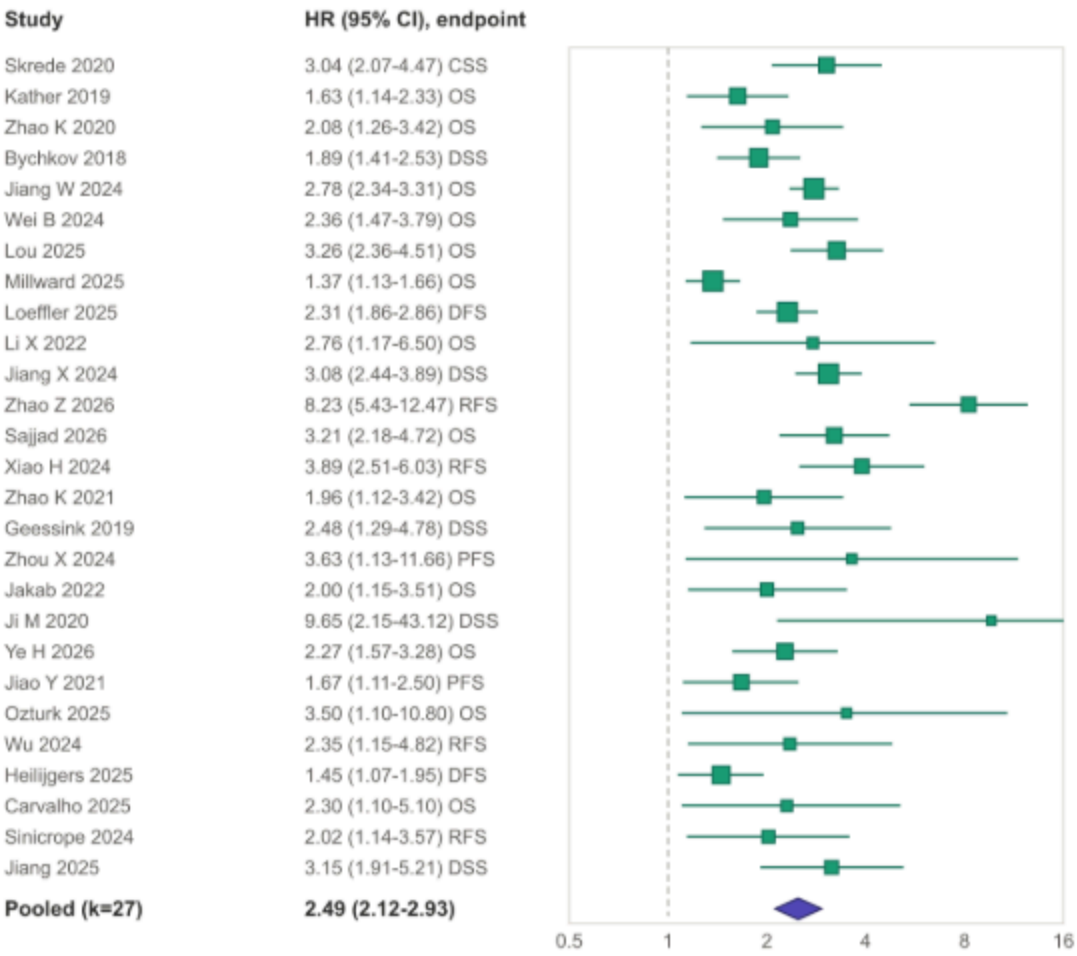


Figure 2

Forest plot of hazard ratios for high versus low DL-defined risk (k = 27; pooled HR 2.49, 95% CI 2.12–2.93; I^2 = 77%).

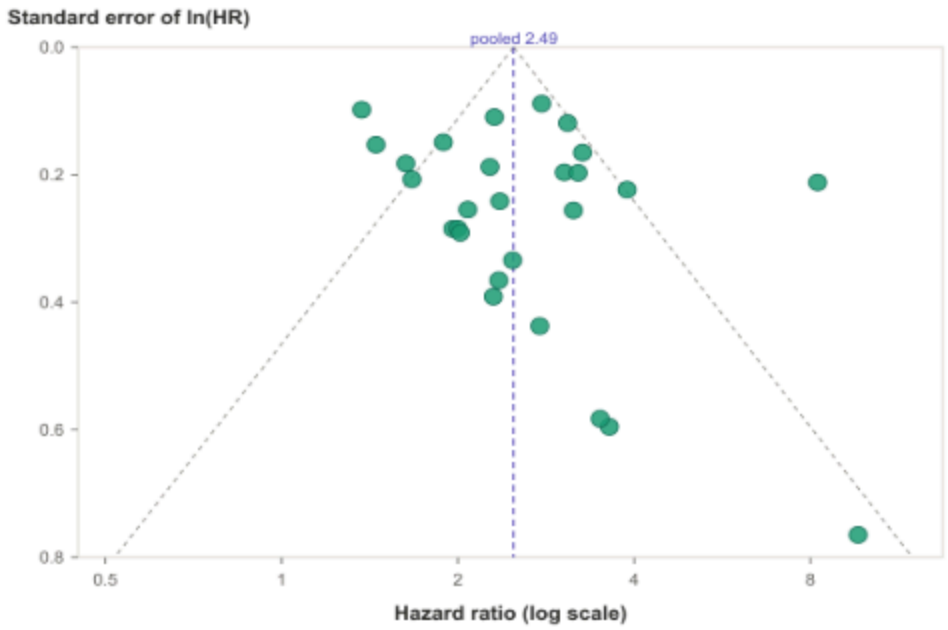


Figure 3

Funnel plot of study-level hazard ratios and their standard errors. The horizontal axis displays HRs on a logarithmic scale.

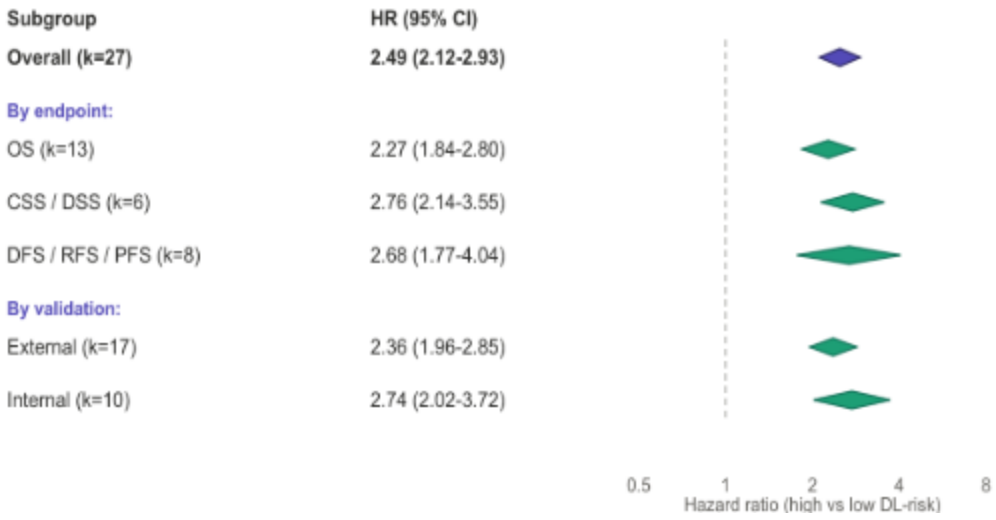


Figure 4

Pooled hazard ratios by validation setting and survival endpoint. The external-versus-internal hypothesis test requires reconciliation with the original analysis code.



Figure 5

Reported risk-of-bias judgements. The recovered figure displays six QUIPS domains and the original overall ratings (low, 6; moderate, 18; high, 3). Separate PROBAST results and the underlying assessment record require confirmation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1DRAFT.docx