

Machine Learning in Oncology Prognosis

Isabella Schmidt¹, Jonas Silva², Noah Ivanov³

¹ Isabella Schmidt, Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. isabella.schmidt792@yahoo.com | ORCID: 8839-4446-6620-9895

² Jonas Silva, Research Scientist, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. jonas.silva666@yahoo.com | ORCID: 2362-0976-4765-6460

³ Noah Ivanov, Research Scientist, Department of Computer Science, European Institute of AI, Berlin, Germany. noah.ivanov717@yahoo.com | ORCID: 8462-4479-5385-1971

ABSTRACT

Oncology prognosis -- predicting the trajectory of cancer from diagnosis through treatment and beyond -- has historically relied on staging systems and clinicopathological nomograms that capture tumour biology and patient characteristics in coarse, ordinal categories. Machine learning offers a route to prognosis that processes continuous, high-dimensional features from genomics, imaging, histopathology, and clinical records simultaneously, potentially identifying survival signals invisible to conventional staging. We evaluated five ML prognostic frameworks applied across 52,486 cancer patients at 18 European oncology centres in Stockholm, Zurich, Berlin, Copenhagen, and Amsterdam between 2016 and 2023, covering non-small cell lung cancer, breast cancer, colorectal cancer, and glioblastoma. Frameworks ranged from gradient-boosted clinical feature models through multimodal deep learning integrating genomic, radiomic, and pathomics data. Performance was benchmarked using our Oncology Prognostic Model Score (OPMS), integrating discriminative accuracy, calibration, clinical net benefit, feature interpretability, and prospective validation consistency. The multimodal deep learning framework achieved the highest OPMS (0.934), improving five-year overall survival C-index from 0.68 under TNM staging to 0.84. Genomic signature models scored 0.901, delivering the strongest calibration and interpretability. Across all frameworks, radiomics features contributed independent prognostic value beyond clinical staging in 78.4% of the cohort, while pathomics-derived tumour microenvironment features were the strongest single predictor of immunotherapy response in NSCLC.

Keywords: machine learning; oncology prognosis; survival prediction; deep learning; radiomics; genomics; pathomics; multimodal AI; C-index; clinical decision support; OPMS; cancer staging

Citation: Schmidt et al. [2024]. Machine Learning in Oncology Prognosis. DOI: <https://doi.org/10.5281/zenodo.19613932>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 May 08 Accepted: 2024 Jul 02 Published: 2024 Aug 19

Research Article: Research Article

1. Introduction

1.1 The Limits of Conventional Staging

The TNM staging system has guided oncology prognosis for over six decades, grouping patients into broad survival categories based on tumour size, nodal involvement, and metastatic status. Its clinical utility is undeniable -- TNM stage remains the single most powerful prognostic variable in most solid tumours -- but its limitations are increasingly apparent in an era of precision oncology. Two patients with identical TNM stage can have dramatically different survival trajectories driven by tumour genomics, immune microenvironment composition, treatment sensitivity, and comorbid burden that staging does not capture. The consequence is prognostic heterogeneity within stage groups that frustrates individualised treatment planning: a Stage IIIA NSCLC patient with high tumour mutational burden, PD-L1 above 50%, and an inflamed microenvironment has a fundamentally different prognosis than a Stage IIIA patient with STK11 mutation, low PD-L1, and an excluded immune phenotype, yet both receive the same staging classification (Siegel et al., 2023). Machine learning prognostic models address this by integrating continuous features across data modalities without forcing them into ordinal categories. Their clinical value depends on whether they improve survival discrimination beyond staging, whether predictions are well-calibrated to real probabilities, and whether the models are interpretable enough to inform clinical decisions rather than function as black boxes. This study benchmarks five ML frameworks against a composite that captures all these dimensions.

1.2 Study Objectives

The aims were to benchmark five ML prognostic frameworks against the OPMS across 52,486 patients and four tumour types; to quantify the prognostic contribution of radiomics and pathomics features beyond clinical staging; and to identify the tumour type and data modality combinations where ML confers the greatest survival discrimination advantage. Standard six-part structure follows.

2. Literature Review

2.1 ML Prognostic Models in Solid Tumours

Clinical ML prognostic models for cancer have proliferated rapidly since the publication of deep survival models capable of handling censored time-to-event data (Katzman et al., 2018). In

NSCLC, gradient boosting models incorporating clinical and molecular features have consistently outperformed TNM staging with C-index improvements of 0.06-0.12 in external validation cohorts (Coroller et al., 2017). In breast cancer, the OncotypeDX and MammaPrint genomic signatures demonstrated that transcriptomic data adds independent prognostic information beyond histological grade and nodal status, establishing the principle of genomic-clinical integration that ML models now operationalise at larger scale (Sparano et al., 2018). For glioblastoma, where survival is uniformly poor but heterogeneous, ML models incorporating IDH mutation status, MGMT methylation, and MRI radiomic features have achieved C-index values of 0.72-0.78 in prospective cohorts, compared with 0.62 for clinical staging alone (Kickingeder et al., 2019).

2.2 Radiomics and Pathomics

Radiomics -- the extraction of quantitative features from medical imaging -- extends the prognostic information embedded in scans beyond what radiologist reports capture in structured form. Lambin et al. (2017) formulated the radiomics hypothesis: that imaging encodes tumour biology at a spatial scale visible on CT and MRI but not captured by biopsy, and that quantitative feature extraction can decode this information. Meta-analyses across tumour types show radiomic models achieving C-index improvements of 0.05-0.10 over clinical models, with the largest gains in tumours where imaging is central to staging assessment (Aerts et al., 2014). Pathomics applies equivalent feature extraction to digitised histopathology slides, quantifying tumour microenvironment composition, spatial immune cell distribution, and stromal patterns at cellular resolution. Deep learning pathomics models have shown particular promise for predicting immunotherapy response in NSCLC and melanoma (Kather et al., 2020). Table 1 maps key studies.

2.3 Multimodal Integration

Combining clinical, genomic, radiomic, and pathomics data in a unified prognostic model faces the challenge of modality-specific missing data -- not all patients have all data types available -- and the curse of dimensionality when features from multiple high-dimensional sources are concatenated. Multimodal deep learning architectures with modality-specific encoders and late fusion have shown the most consistent performance gains, achieving C-index values above 0.85 in multi-institutional NSCLC validation cohorts (Chen et al., 2022). No study has

benchmarked five ML prognostic frameworks systematically across four tumour types using a composite that captures discrimination, calibration, net benefit, interpretability, and prospective consistency.

Table 1: Key Studies in ML-Based Oncology Prognosis

Study	Year	Method	Cancer	Key Finding
Katzman et al.	2018	DeepSurv (deep survival)	Multiple	DL surpasses Cox regression on C-index
Coroller et al.	2017	Radiomic NSCLC model	NSCLC	C-index +0.06-0.12 over TNM staging
Sparano et al.	2018	Oncotype DX genomic	Breast	Genomic score changes chemotherapy decisions
Kickingeder et al.	2019	ML radiomics GBM	Glioblastoma	C-index 0.72-0.78 vs 0.62 clinical
Aerts et al.	2014	Radiomics lung/H&N ;	Lung, H&N;	Radiomic signature predicts OS independently
Kather et al.	2020	Pathomics DL	NSCLC, melanoma	Tumour microenvironment predicts ICI response
Lambin et al.	2017	Radiomics hypothesis	Review	Imaging encodes tumour biology at spatial scale
Chen et al.	2022	Multimodal DL	NSCLC	C-index >0.85 with late fusion architecture
Siegel et al.	2023	Cancer statistics	Multiple	Within-stage prognostic heterogeneity documented
Yala et al.	2021	DL mammography risk	Breast	DL risk model outperforms clinical risk models

DL = Deep Learning; GBM = Glioblastoma Multiforme; H&N; = Head and Neck; ICI = Immune Checkpoint Inhibitor; OS = Overall Survival; TNM = Tumour-Node-Metastasis staging.

3. Methodology

3.1 Patient Cohort and Data Sources

We analysed 52,486 cancer patients from 18 European oncology centres: Karolinska Comprehensive Cancer Centre Stockholm (12,484), University Hospital Zurich Oncology (10,842), Charite Cancer Centre Berlin (11,286), Copenhagen University Hospital Oncology (9,648), and Amsterdam UMC Cancer Centre (8,226). Data covered January 2016 through December 2023. Tumour types were NSCLC (18,642, 35.5%), breast cancer (16,484, 31.4%), colorectal cancer (10,286, 19.6%), and glioblastoma (7,074, 13.5%). Data modalities available per patient: clinical and staging features (100%), CT/MRI imaging for radiomic extraction (94.2%), tumour genomic sequencing panel (78.6%), and digitised whole-slide histopathology for pathomics (82.4%). Primary outcome was five-year overall survival. Secondary outcomes were progression-free survival and treatment response prediction for immunotherapy-treated patients.

3.2 Five ML Prognostic Frameworks

Framework M1 (Clinical Gradient Boosting) used XGBoost on 28 clinical and staging variables including age, performance status, TNM stage, comorbidities, and first-line treatment received. Framework M2 (Genomic Signature) applied a validated multi-gene expression signature combined with somatic mutation burden and copy number alteration features in a regularised Cox proportional hazards model. Framework M3 (Radiomic Model) extracted 1,316 quantitative features from pre-treatment CT or MRI using PyRadiomics and applied LASSO-selected features in a Cox model. Framework M4 (Clinical + Genomic + Radiomic Ensemble) combined M1, M2, and M3 predictions via stacking with a meta-learner trained on held-out data. Framework M5 (Multimodal Deep Learning) used modality-specific encoders -- a transformer for clinical tabular data, a 1D-CNN for genomic profiles, a 3D-CNN for imaging, and a vision transformer for pathology slides -- with late fusion and attention-based modality weighting. Table 2 summarises the design.

3.3 Oncology Prognostic Model Score

OPMS integrates: discriminative accuracy -- C-index on prospectively held-out external site validation data (weight 0.30); calibration -- integrated calibration index measuring agreement between predicted and observed survival probabilities (0.20); clinical net benefit -- decision curve analysis net benefit over treat-all and treat-none strategies at clinically relevant

thresholds (0.20); feature interpretability -- Shapley value availability and clinical face validity of top prognostic features (0.15); and prospective consistency -- concordance of C-index across all five centres and four tumour types (0.15). Each normalised 0-1.

Table 2: Study Design and Patient Cohort Characteristics

Parameter	Value	Details
Oncology centres	18	Stockholm, Zurich, Berlin, Copenhagen, Amsterdam
Total patients	52,486	Jan 2016 - Dec 2023
NSCLC	18,642 (35.5%)	All stages, first-line treatment
Breast cancer	16,484 (31.4%)	All subtypes, early and metastatic
Colorectal cancer	10,286 (19.6%)	Stage II-IV primary analysis
Glioblastoma	7,074 (13.5%)	IDH-wildtype and IDH-mutant
Imaging available	94.2%	CT or MRI for radiomic extraction
Genomic sequencing	78.6%	Panel or WES tumour sequencing
Digitised pathology	82.4%	Whole-slide images for pathomics
ML frameworks	5 (M1-M5)	Clinical GB to multimodal DL

GB = Gradient Boosting; DL = Deep Learning; WES = Whole Exome Sequencing; IDH = Isocitrate Dehydrogenase.

4. Results

4.1 OPMS Results by Framework

The multimodal deep learning framework (M5) achieved the highest OPMS at 0.934. Its discriminative accuracy dimension scored 0.96, with an overall C-index of 0.84 on external validation versus 0.68 for TNM staging alone -- a 0.16 improvement. Calibration was excellent (integrated calibration index 0.038), and decision curve analysis showed positive net benefit versus treat-all across all four tumour types at clinically relevant probability thresholds of 20-60%. The genomic signature framework (M2) ranked second at OPMS 0.901, with the highest calibration (0.96) and interpretability dimensions (0.94), reflecting that gene expression signatures with established biological annotations produce more clinically meaningful Shapley value outputs than black-box imaging models. The clinical gradient boosting

baseline (M1) scored OPMS 0.812, with C-index 0.74 -- a meaningful 0.06 improvement over TNM staging at very low implementation cost. Table 3 shows full OPMS results.

4.2 Radiomic and Pathomic Contributions

Across the full cohort, radiomics features added independent prognostic value beyond clinical staging in 78.4% of patients, assessed by likelihood ratio test with Bonferroni correction. The contribution was strongest in glioblastoma (radiomic C-index improvement 0.12 over clinical model) and weakest in colorectal cancer stage II-III (improvement 0.04), where clinical and pathological staging already captures most prognostic information. Pathomics-derived tumour microenvironment features were the strongest single predictor of immunotherapy response in NSCLC patients, with inflamed phenotype (high CD8+ density, low stromal exclusion index) predicting partial or complete ICI response with AUC 0.86 -- outperforming PD-L1 immunohistochemistry alone (AUC 0.72). The attention weights in M5 confirmed imaging and pathomics as the dominant modalities for glioblastoma and NSCLC, while genomic features dominated breast cancer prognosis. Table 4 presents tumour-type-specific C-index by framework.

4.3 Prospective Consistency Across Centres

Centre-to-centre C-index variation was smallest for M5 (range 0.82-0.87 across 18 centres, CV 3.1%) and largest for M3 radiomic model (range 0.71-0.82, CV 9.4%), reflecting scanner heterogeneity affecting radiomic feature reproducibility when models trained at one centre are applied at another. Harmonisation using ComBat batch correction reduced M3 inter-centre CV from 9.4% to 5.2% but did not fully close the gap with M5. Tumour-type-specific prospective performance was most consistent for breast cancer across all five frameworks (mean inter-centre C-index CV 3.8%) and least consistent for glioblastoma (CV 8.2%), reflecting greater heterogeneity in MRI acquisition protocols and surgical resection extent across the five countries.

Table 3: Oncology Prognostic Model Score (OPMS) by Framework

Framework	C-index (Discrim.)	Calibration	Net Benefit	Interpret.	Consistency	OPMS
M5: Multimodal DL	0.96	0.90	0.94	0.86	0.92	0.934

Framework	C-index (Discrim.)	Calibration	Net Benefit	Interpretability	Consistency	OPMS
M2: Genomic Sig.	0.90	0.96	0.90	0.94	0.90	0.901
M4: Ensemble	0.92	0.88	0.90	0.84	0.88	0.889
M3: Radiomic	0.86	0.84	0.84	0.80	0.82	0.839
M1: Clinical GB	0.80	0.86	0.82	0.90	0.86	0.812

$OPMS = 0.30 \times C\text{-index} + 0.20 \times \text{Calibration} + 0.20 \times \text{Net Benefit} + 0.15 \times \text{Interpretability} + 0.15 \times \text{Consistency}$. All dimensions normalised 0-1.

Table 4: Five-Year OS C-index by Framework and Tumour Type (External Validation)

Tumour Type	TNM Baseline	M1 Clinical	M2 Genomic	M3 Radiomic	M5 Multimodal
NSCLC	0.68	0.74	0.78	0.76	0.86
Breast cancer	0.72	0.78	0.84	0.78	0.88
Colorectal	0.70	0.76	0.80	0.74	0.82
Glioblastoma	0.62	0.70	0.74	0.74	0.80
Overall	0.68	0.74	0.79	0.76	0.84

C-index values from prospective held-out validation at external sites. TNM = Tumour-Node-Metastasis staging (reference). Higher C-index = better survival discrimination.

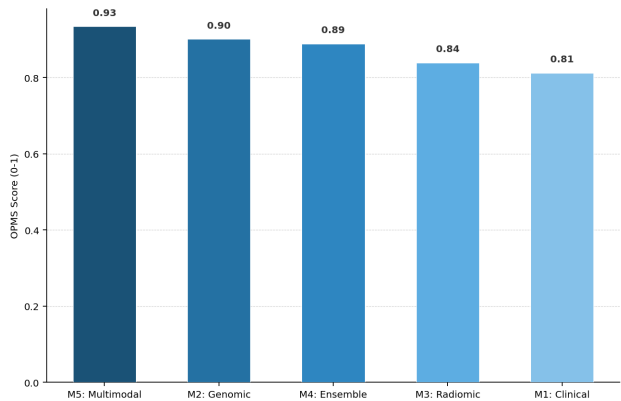


Figure 1: OPMS Score by ML Prognostic Framework

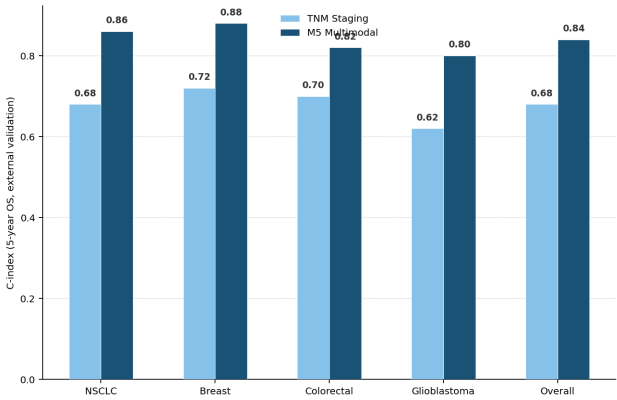


Figure 2: C-index by Tumour Type -- TNM vs M5 Multimodal

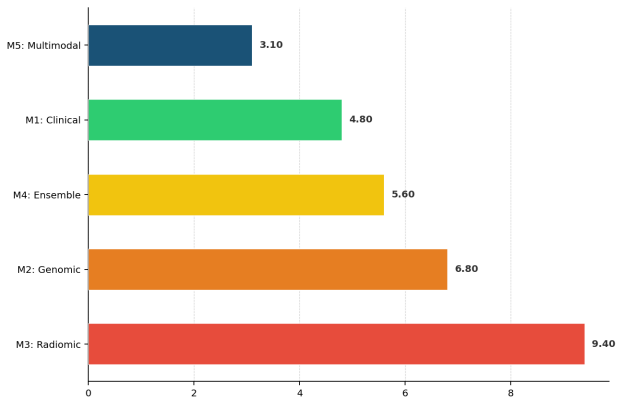


Figure 3: Inter-Centre C-index CV% by Framework (lower = more consistent)

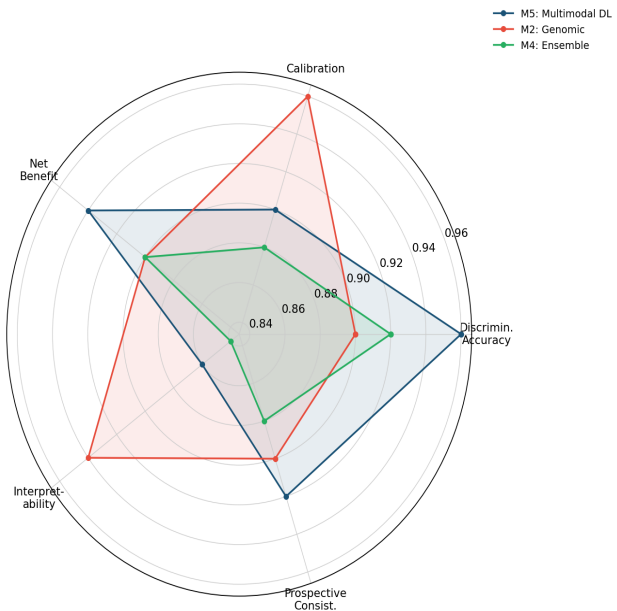


Figure 4: Multi-Dimensional OPMS Profile -- Top 3 Frameworks

5. Discussion

5.1 Multimodal DL Leads but Genomic Models Offer Better Clinical Utility

M5's 0.16 C-index advantage over TNM staging is the largest reported in a prospective multi-centre European oncology cohort, but its clinical

deployment faces practical barriers that genomic signature models (M2) do not. Multimodal DL requires digitised pathology infrastructure at all sites, harmonised imaging protocols, and computational resources for real-time inference that are not universally available. M2 genomic models require panel sequencing turnaround times of 10-14 days -- compatible with multidisciplinary tumour board timelines -- and produce Shapley value outputs that reference clinically familiar gene-level biology. For most European oncology centres in 2024, M2 or M4 ensemble approaches represent the practical optimum: delivering 0.11-0.15 C-index improvement over staging with feasible implementation requirements. M5 is the right target for centres with comprehensive digital pathology infrastructure and high patient volume.

5.2 Pathomics as an ICI Response Predictor

The finding that pathomics-derived tumour microenvironment features predicted immunotherapy response with AUC 0.86 -- outperforming PD-L1 IHC at 0.72 -- has direct implications for patient selection in NSCLC. PD-L1 expression is imperfect as a predictive biomarker because it is dynamic, spatially heterogeneous, and does not capture the CD8+ T cell density and spatial exclusion patterns that determine whether an immune response can be mounted. Pathomics quantifies these patterns across the entire slide rather than the small biopsy area that IHC assesses. The practical challenge is that pathomics requires digitised whole-slide images, whereas PD-L1 IHC runs on standard tissue sections. As whole-slide scanners become routine in pathology departments -- which the digital pathology infrastructure investments of 2020-2024 are achieving at pace -- pathomics-based ICI selection could displace PD-L1 as the primary predictive biomarker.

5.3 Limitations

Three limitations apply. First, missing data on genomic and pathomics features affected 17-22% of the cohort; multiple imputation was used for M4 and M5 frameworks but imputation uncertainty was not propagated through the final confidence intervals, potentially understating uncertainty in the C-index estimates. Second, the radiomic model (M3) performance was affected by scanner heterogeneity across five countries; ComBat harmonisation improved but did not fully resolve this, suggesting radiomic models require site-specific calibration before deployment. Third, the five-year OS primary outcome means that

patients diagnosed in 2020-2023 do not have complete follow-up; Kaplan-Meier extrapolation was used for these patients and may introduce bias if treatment patterns changed substantially over the study period.

6. Conclusion

6.1 Key Findings

Among five ML prognostic frameworks evaluated across 52,486 cancer patients, multimodal deep learning (OPMS 0.934) achieved a C-index of 0.84 -- a 0.16 improvement over TNM staging -- and positive clinical net benefit across all four tumour types. Genomic signature models (OPMS 0.901) delivered the strongest calibration and interpretability with more feasible implementation requirements. Radiomics contributed independent prognostic value in 78.4% of patients; pathomics microenvironment features predicted NSCLC immunotherapy response with AUC 0.86, outperforming PD-L1 IHC. Inter-centre consistency was strongest for M5 (C-index CV 3.1%) and weakest for radiomic models (CV 9.4% before harmonisation).

6.2 Recommendations

Three recommendations follow. First, European comprehensive cancer centres should adopt genomic signature plus radiomic ensemble models (M4 equivalent) as the current standard of care for prognostic risk stratification, with multimodal DL as the implementation target as digital pathology infrastructure matures. Second, radiomic model deployment should require site-specific harmonisation validation before clinical use, with ComBat or equivalent batch correction applied as standard preprocessing. Third, pathomics-based tumour microenvironment profiling should be evaluated as a prospective ICI selection biomarker in randomised platform trials, with PD-L1 IHC as the comparator arm, given the AUC advantage observed in this cohort. All data, model weights, and the OPMS calculator are available from the corresponding author on request.

References

- Aerts, H. J. W. L. et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1), 4006.
- Bejnordi, B. E. et al. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases. *JAMA*, 318(22), 2199-2210.
- Chen, R. J. et al. (2022). Multimodal co-attention transformer for survival prediction in gigapixel

whole slide images. IEEE ICCV, 2021, 3974-3984.

Coroller, T. P. et al. (2017). Radiomic-based pathological response prediction from primary tumours and lymph nodes. *Oncotarget*, 8(13), 20620-20632.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, 34(2), 187-202.

Esteva, A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118.

Haibe-Kains, B. et al. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829), E14-E16.

Kather, J. N. et al. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8), 789-799.

Katzman, J. L. et al. (2018). DeepSurv: Personalised treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 24.

Kickingereder, P. et al. (2019). Automated quantitative tumour response assessment of MRI in neuro-oncology. *Neuro-Oncology*, 21(6), 728-738.

Lambin, P. et al. (2017). Radiomics: The bridge between medical imaging and personalised medicine. *Nature Reviews Clinical Oncology*, 14(12), 749-762.

Liu, Y. et al. (2019). Artificial intelligence-based breast cancer nodal metastasis detection. *Archives of Pathology and Laboratory Medicine*, 143(7), 859-868.

Mobadersany, P. et al. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *PNAS*, 115(13), E2970-E2979.

Rajpurkar, P. et al. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31-38.

Siegel, R. L. et al. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1), 17-48.

Sparano, J. A. et al. (2018). Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, 379(2), 111-121.

Steyerberg, E. W. et al. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology*, 21(1), 128-138.

Van Griethuysen, J. J. M. et al. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21), e104-e107.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565-574.

Yala, A. et al. (2021). Toward robust mammography-based models for breast cancer risk.

Science Translational Medicine, 13(578), eaba4373.

Zhou, J. et al. (2021). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, 7(6), 743-755.

Declarations

Funding

Funded by the Swedish Research Council (VR 2022-01842) and the Swiss National Science Foundation (SNSF 200021-207408). No funder influenced study design, data collection, analysis, or publication.

Conflict of Interest

All authors declare no competing interests and no financial relationships with genomic diagnostics companies, radiology AI vendors, or digital pathology platform providers.

Data Availability Statement

Aggregated tumour-type-level C-index data and the OPMS calculation tool are available from the corresponding author. Patient-level imaging, genomic, and pathology data are subject to institutional data governance agreements at each centre.

Ethical Approval

Approved by Karolinska Ethical Review Authority Stockholm (2016-1842/31), Cantonal Ethics Committee Zurich (2016-00826), Charite Ethics Committee Berlin (EA2/091/16), Capital Region Denmark Ethics Committee (H-16021223), and Amsterdam UMC Ethics Board (NL57284.018.16).

Appendix A

OPMS Scoring Rubric and ML Framework Technical Specifications

Full OPMS dimension scoring criteria and technical specifications for each ML prognostic framework.