scientific reports

Check for updates

OPEN

A pipeline of machine learning-driven multi-modal data fusion methods for prognostic risk analysis in bevacizumab-treated metastatic colorectal cancer

Valentina Thomas[1,24], Gift Nyamundanda[2,24], Adrian Lärkeryd[3], P. S. Hari[2,4], Ian S. Miller[1], Tom Venken[5,6], Dominiek Smeets[5,6], Bram Boeckx[5,6], Johannes Betge[7,8,9], Matthias P. A. Ebert[7,9], Timo Gaiser[10], Verena Murphy[11], Elaine Kay[12], Henk M. Verheul[13], Alice C. O'Farrell[1], Chiara Cremolini[14,15], Federica Marmorino[14,15], William M. Gallagher[16], Ana Barat[17], Rut Klinger[18], Bozena Fender[16], Bauke Ylstra[19], Nicole van Grieken[19], Deborah A. McNamara[12], Bryan T. Hennessy[12,20], Sudipto Das[21], Bruce Moran[18], Darran P. O'Connor[21], Rodrigo Dienstmann[22], Diether Lambrechts[5,6], Jochen H. M. Prehn[17], Anguraj Sadanandam[2,4,23,25]✉ & Annette T. Byrne[1,25]✉

We introduce a comprehensive multi-step machine-learning driven pipeline which fuses multi-modal omics datasets and clinical outcomes with survival and treatment response to predict patient outcome following anti-angiogenic therapy in the metastatic colorectal cancer (mCRC) setting. The approach encompasses the following steps: a) employing a sparse Bayesian factor analysis method (PhenMap) to select significant mapping-variables (MVs) and associated biomarkers (features) through joint multivariable modelling of copy number aberrations (CNA) and clinical covariates, including mutations and clinical demographics/outcomes; b) utilizing Cox-proportional hazard analysis on the MVs to select those associated with progression-free survival (PFS); and c) employing elastic net Cox regression analysis on the prognostic features selected by PhenMap to delineate risk groups associated with bevacizumab (BVZ) response. Through this approach, we have identified three putative features (CNA—15q21.1 and 1p36.31 deletions and *BRAF* mutation) from two prognostically-significant MVs to stratify N = 117 mCRC patients, who received BVZ combination therapy, into 3 (low, medium, high) prognostic risk groups that were significantly associated with PFS and treatment response. Mortality risk was significantly greater in the high-risk group with 100% (n = 12) of patients showing no response to BVZ, compared to the low-risk group where 10 out of 12 patients (88%) showed a response to BVZ. The risk groups were independently negative predictors of survival in BVZ-treated mCRC patients. Overall, Wwe have established a machine learning pipeline integrating multi-modal omics data with response, and have implicated a putative combined CNA/mutation candidate biomarker with associated risk scores that could, in the future, help stratify mCRC patients unlikely to benefit from BVZ combination therapy. Our novel precision medicine approach applies disruptive advancements in artificial intelligence and bioinformatics methodologies to tumour biology datasets.

Keywords Machine learning, Multi-modal data fusion, Prognostic risk analysis, Metastatic colorectal cancer, Personalised medicine, Biomarkers, PhenMap

[1]RCSI Precision Cancer Medicine Group, Department of Physiology & Medical Physics, Royal College of Surgeons in Ireland, University of Medicine and Health Sciences, Dublin D02 YN77, Ireland. [2]Division of Cancer Biology, The Institute of Cancer Research, London SW3 6JB, UK. [3]Centre for Translational Immunotherapy, Division of Radiology and Imaging, The Institute of Cancer Research, London SW3 6JB, UK. [4]Centre for Global Oncology and Research, Division of Cancer Biology, Institute of Cancer Research, London SW3 6JB, UK. [5]VIB Center for Cancer Biology, VIB, Herestraat 49, 3000 Leuven, Belgium. [6]Department of Human Genetics, University of Leuven (KU Leuven), Herestraat 49, 3000 Leuven, Belgium. [7]Department of Medicine II, University Hospital Mannheim, Heidelberg

University, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany. [8]Junior Clinical Cooperation Unit Translational Gastrointestinal Oncology and Preclinical Models, German Cancer Research Center (DKFZ), Heidelberg, Germany. [9]DKFZ-Hector Cancer Institute at the University Medical Center Mannheim, Mannheim, Germany. [10]Institute of Pathology, University Hospital Mannheim, Heidelberg University, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany. [11]Cancer Trials Ireland, RCSI House, 121 St Stephen's, Green, Dublin D2, Ireland. [12]Department of Surgery, Beaumont Hospital, Beaumont Rd, Beaumont, Dublin D9, Ireland. [13]Erasmus University, Rotterdam, the Netherlands. [14]Unit of Medical Oncology 2, University Hospital of Pisa, Pisa, Italy. [15]Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy. [16]OncoMark Limited, NovaUCD, Belfield Innovation Park, Dublin D4, Ireland. [17]Centre for Systems Biology, Department of Physiology & Medical Physics, Royal College of Surgeons in Ireland, University of Medicine and Health Sciences, Dublin D02YN77, Ireland. [18]UCD School of Biomolecular and Biomedical Science, UCD Conway Institute, University College Dublin, Dublin D4, Ireland. [19]Department of Pathology, Cancer Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, De Boelelaan 1117, 1081 HV Amsterdam, the Netherlands. [20]Department of Medicine, Royal College of Surgeons in Ireland, University of Medicine and Health Sciences, Dublin D02YN77, Ireland. [21]School of Pharmacy and Biological Sciences, Royal College of Surgeons in Ireland, University of Medicine and Health Sciences, Dublin D02 YN77, Ireland. [22]Vall d'Hebron Institute of Oncology (VHIO), Universitat Autònoma de Barcelona, Barcelona, Spain. [23]Botton-Champalimaud Pancreatic Cancer Center, Lisbon, Portugal. [24]Valentina Thomas and Gift Nyamundanda, contributed equally to the research work. [25]Anguraj Sadanandam and Annette T. Byrne are co-senior authors who contributed equally to the work. ✉email: Anguraj.Sadanandam@icr.ac.uk; annettebyrne@rcsi.ie

The advent of precision medicine has been significantly accelerated by the latest advancements in bioinformatics and genomics technologies, which mandate the integration of multi-omics and clinical (multi-modal) data using cutting-edge machine learning methods for translational oncological research. This transformative approach focuses on leveraging machine learning techniques within the clinical setting to uncover multimodal biomarkers, with the ultimate objective of prolonging patients' lives and enhancing their quality of life. Recently, several disruptive studies have demonstrated the potential of machine learning and artificial intelligence (AI) in colorectal cancer (CRC). For example, Liu et al. (2025)[1] have employed a machine learning-driven multi-targeted drug discovery approach using biomarker signatures to identify precision therapeutic targets in colon cancer, exemplifying how machine learning can reveal actionable molecular patterns. Moreover, Bahrambanan et al. (2025)[2] developed a deep-learning–based classification system to predict response to radiochemotherapy in CRC, showing that advanced AI architectures outperform conventional machine learning in modeling treatment sensitivity. Similarly, Li et al. (2025)[3] applied quantitative systems modeling of the cancer-immunity cycle to predict disease progression in advanced metastatic (m)CRC, demonstrating how mechanistic and machine learning-based frameworks can capture tumour-immune dynamics relevant to clinical outcomes. Here, we have investigated the application of our machine learning tool, Phenotype Mapping (PhenMap)[4], to identify candidate multi-modal biomarkers in CRC patients, followed by machine learning-based prognostic risk scoring to associate the identified biomarkers with resistance to bevacizumab (BVZ) combination therapy.

CRC ranks as the third most frequently diagnosed malignancy in both men and women and is linked to significant mortality and morbidity[5,6]. Nearly 50% of patients diagnosed will develop metastatic disease. The current treatment approach for *RAS* mutant mCRC involves 5-fluoruracil-based standard-of-care chemotherapy (SOC) combined with the angiogenesis inhibitor BVZ. Results from phase III clinical trials have shown that the addition of BVZ to chemotherapy enhances response rates and extends the survival of patients with mCRC[7–9]. However, response is only observed in a subset of patients, and the overall clinical benefit of BVZ is limited, leading to eventual disease progression[8]. BVZ therapy is further associated with off target toxicities and high treatment costs. Nevertheless, in the absence of other treatments, BVZ is administered to *RAS* mutant mCRC patients as first-line treatment[9,10]. Critically, while we and others have previously proposed various genomic entities as potential predictors of BVZ response[11–17], no potential biomarkers are available to identify mCRC patients that will *not* respond to BVZ. To address this outstanding clinical issue, here we have developed a pipeline of feature selection and risk-scoring machine learning methodologies to detect both BVZ responders and non-responders. Specifically, we have leveraged our previously established unsupervised machine learning tool, PhenMap[4], which supports the simultaneous integration of cancer multi-omics and phenome data. This tool can further facilitate the identification of clinically relevant and functional subtypes associated with specific endpoints. For example, in the context of breast cancer, PhenMap has previously identified subtypes for response prediction to a cyclin-dependent kinase (CDK) 4/6 inhibitor[4].

Here, PhenMap was applied to the ANGIOPREDICT[18] clinical cohort, a pooled retrospective tumour (and associated clinical metadata) collection from patients which primarily underwent combination BVZ therapy, at centres in Ireland, The Netherlands, and Germany[19]. In the present study, we exploited previously generated copy number alteration (CNA) and mutational profile datasets for these patients[16,17,19] which were integrated with clinical data using the PhenMap machine learning tool. PhenMap output was further employed to develop prognostic risk groups that identified biomarkers associated with patients who were resistant to BVZ combination treatment. A comprehensive flow diagram summarizing the workflow pipeline is shown in Fig. 1**.** Overall, we describe the development of a machine learning framework that fuses multi-modal data with anti-angiogenic patient prognosis and treatment responses to predict outcome in mCRC. Our novel approach aligns with the new era of cancer precision medicine, which has been reinforced by recent advancements in the application of AI and bioinformatics methodologies to tumour biology datasets.

Methods

Datasets

Data generated from the ANGIOPREDICT retrospective cohort[19] (locally irresectable or mCRC tumours from patients treated with BVZ and chemotherapy) was employed in the current study. Briefly, samples had previously been subjected to low-coverage whole genome sequencing, and Genomic Identification of Significant Targets in Cancer (GISTIC) 2.0[20] scores for the CNA data generated[17], where 107 regions were shown to be significantly deleted/amplified. Further, gene mutational status had previously been derived from whole exome sequencing (paired germline and tumour DNA available)[17]. Finally, clinical data associated with each patient were also collated, including clinicopathological variables such as gender, age, stage, tumour sidedness, chromosomal instability (CIN), number of break points and number of mutations (see previous study[17]). These datasets (i.e. copy number, mutation, and clinical) were filtered to specifically retain patients with matching clinical, mutational, and copy number information. In total, data from 117 patients with matching features across the three data types (107 CNA regions, 5 gene mutational data, and clinicopathological data) were used as data input for PhenMap.

PhenMap

PhenMap[4] is a factor analytic model that models high-dimensional observed data as a linear function of corresponding low-dimensional mapping-variables (MVs). The model attempts to explain this variation captured by MVs using additional patient phenotypic information or covariates. Consider measurements, $y_i = (y_{i1} \cdots y_{ip})^T$, which may represent expression of p genes from tumour i, and characterized with l phenotypes, $x_i = (x_{i1} \cdots x_{il+1})^T$ (here, 1 represents intercept), including clinical and/or genotype parameters, where T represents transpose of a matrix. PhenMap can model tumour i (i = 1… n tumours) correlated measurements, y_i, using low-dimensional MVs, $z_i = (z_{i1} \cdots z_{iq})^T$, where q is the number of MVs, by assuming the following Gaussian distributions;

$$P(y_i) = MVN_p[Wz_i, \Sigma] \quad (1)$$

$$P(z_i) = MVN_q[\beta x_i, \Phi] \quad (2)$$

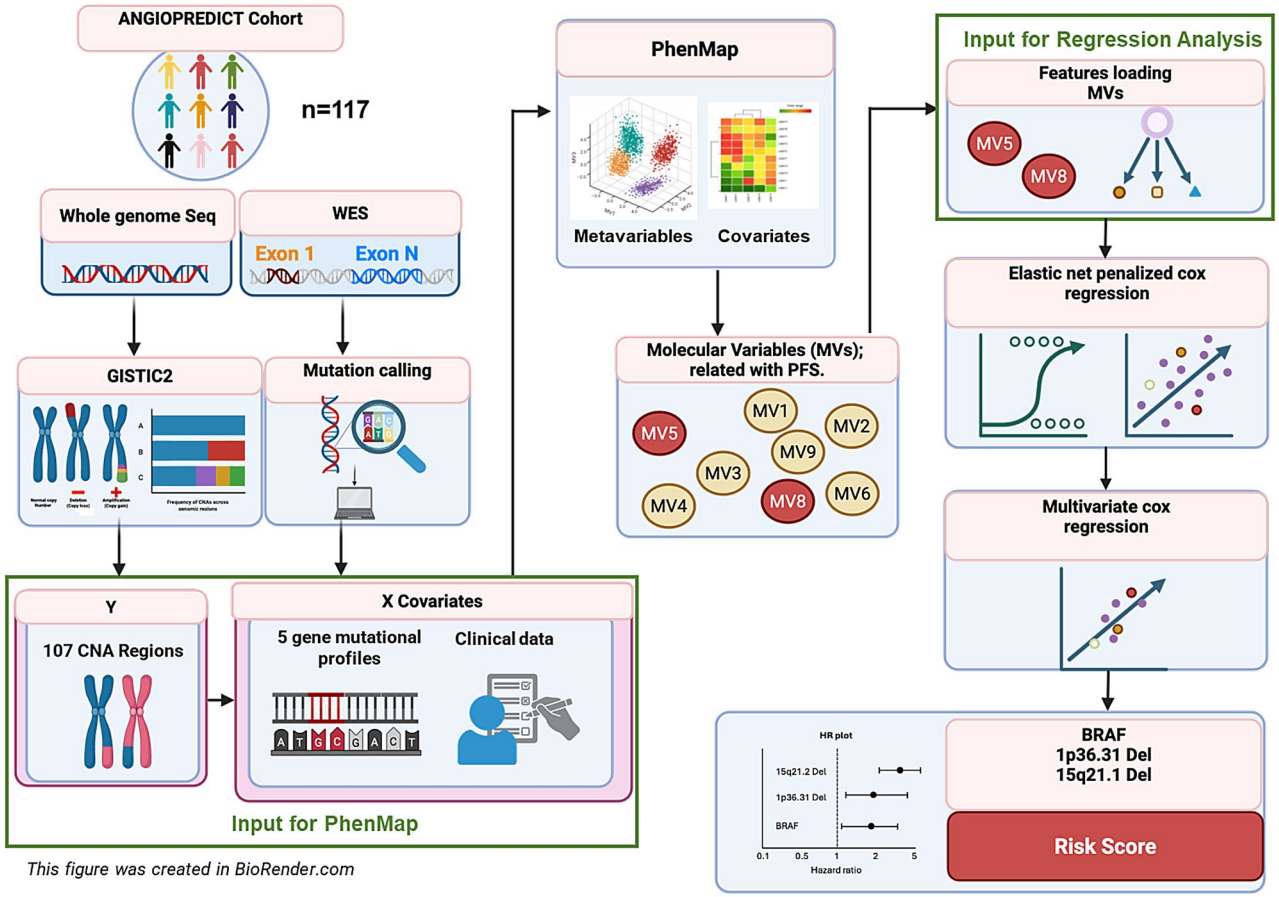

Fig. 1. Flow chart summarizing workflow pipeline.

where W is a p x q loadings matrix relating each feature to q MVs and $\sum = diag\left(\sigma_1^2 \cdots \sigma_p^2\right)$ is the diagonal residual covariance matrix. Regression coefficients, β, quantify the effect of phenotypes (covariates) on each MV and $\Phi = diag\left(\phi_1^2 \cdots \phi_q^2\right)$ is the residual covariance MVs. This model is estimated using a Bayesian framework to address the small-sample-size problem common in genomics. Please see Nyamundanda et al[4]. for more details on the estimation of this model.

To model the correlation between features we introduce the following automatic relevance determination (ARD) sparsity prior on W which enforces element-wise sparsity for feature selection,

$$P(\lambda_{jk}|a, b) = G(\lambda_{jk}|a, b) \tag{3}$$

where G is a gamma distribution, $j = 1 \ldots p_t$, and $k = 1 \ldots q$. The ARD prior on precision hyper-parameter λ_{jk} controls the contribution, w_{jk}, of feature j on the k^{th} MV. Increasing λ_{jk} shrinks w_{jk} the corresponding loading towards zero inducing sparsity in W.

To allow for phenotype (covariates) selection g-prior was set on regression coefficients to regularize β and also to eliminate sensitivity of parameter estimation to changes in the scales of phenotypes. The following g-prior was considered, $p(\beta k|\phi_k^2, g) = MVN_{l+1}[\beta k|0, g(X^T X)^{-1}\phi_k^2]$, where g controls the prior covariance.

Application of PhenMap

The clinical, mutational, and copy number data (detailed above) were integrated using PhenMap, a previously described sparse Bayesian factor analysis machine learning tool that concurrently models omics and phenome data for functional subtyping[4]. Copy number GISTIC scores were treated as output features whilst mutations and patient clinical variables were treated as matched covariates for the sample. The prognostic information associated between MVs from PhenMap and progression-free survival (PFS) for patients was modeled using standard Cox-proportional hazard model.

Significant feature selection and risk scoring

Firstly, to identify the most robust genomic markers that predict PFS in BVZ patients, CNA regions and mutations associated with the prognostic MVs were selected and processed through penalized (elastic net-based) Cox regression[21]. This was performed by repeatedly splitting the data into training/testing data and fitting penalized Cox regression 100 times to remove sampling bias. Features (CNA regions and mutations) selected by penalized (elastic net-based) Cox regression with frequency of 100% were chosen as potential prognostic markers. To identify genomic attributes that are highly predictive of prognosis, we fitted a standard multivariate Cox regression model with the selected markers and selected those significant at 5% significance level (after iteratively removing non-significant features until all features were significant). Finally, the risk score (risk of mortality), Ci, for the ith patient who received BVZ was calculated as follows;

$$Ci = \sum_{j=1}^{g} \log\left(\exp\left[\beta_j\right]\right) * X_j$$

where g is the number of selected genomic markers, β_j is the regression coefficient (natural-logarithm of hazard ratio [HR]) for X_j values for marker j. Hence, risk scores are a product of the selected genomic markers with their corresponding HR (representing the risk associated with each marker).

To assign patients into risk groups, the risk scores were split into quantiles (0.1 and 0.9) to separate the lowest and highest 10% of scores, respectively, into the categories low-risk and high-risk. Often, the extremes or tails of these distributions contain outlier samples that behave differently from others, like low- and high-risk patients. The intermediate scoring patients were assigned to the moderate risk group. The association of the risk groups with clinical outcome (PFS) was evaluated using multivariate Cox proportional hazards models, together with age, tumour location, stage, CIN, and including our previously published ANGIOPREDICT subtypes (derived from CNAs only)[17].

For survival analyses, the Kaplan–Meier method was used to plot survival curves. The endpoint studied was PFS. PFS was the time between the study entry and either the date of the first recurrence or the date that the last follow-up took place. All statistical analyses were performed in R Version 4.3.0.

Results

PhenMap selects integrated multi-modal features in BVZ-treated mCRC samples

PhenMap-based joint integration of CNAs (as outcomes), and mutations and clinical demographics/outcomes (as covariates) (original data derived from the ANGIOPREDICT study[17]), revealed the top ten MVs (MV1 to MV10) (refer to methods section for MV explanation) (Fig. 2a) that represented distinct contexts of molecular heterogeneity in the ANGIOPREDICT cohort between CNAs and mutational data.

Interestingly, MV5 and MV8, were associated with PFS (prognostic; based on Cox-Proportional Hazard analysis of PFS) (Fig. 2a). In order to efficiently derive important clinical interpretations, PhenMap identified integrated features, CNAs and covariates (phenotypes), responsible for driving the heterogeneity in MV5 and MV8 (Fig. 2b and c). The main CNA regions significantly associated with MV5 were 15q26.1, 2q33.1, 15q22.33, 15q11.2, and 15q21.1 (at a significance level of 0.25%). Furthermore, *BRAF* was found to be the only significant covariate driving MV5 heterogeneity (at a significance level of 5%). Similarly, the main CNA regions significantly associated with MV8 (at significance level of 0.25%) were 1p36.31, 4q13.3, 1p36.11, 4p16.2, 1p33, 1p31.1, 1p34.2, and 1p21.1. No mutation or clinical co-variates were found to be associated with this MV (Fig. 2b and c). Hence,

PhenMap selected significant prognostic features associated with MV5 and MV8 that were employed for the next set of downstream analyses.

Signature of CNAs/mutations identifies risk groups in BVZ treated mCRC

To derive a joint effect of these relationships (between CNAs and mutations) on prognosis, integrated CNAs and mutations were reduced to a risk score, i.e., weighing the risk of mortality for each patient based on PFS after treatment with combination BVZ therapy. This reduction of features was performed by first fitting penalized Cox regression (with elastic net machine learning approach[22]) on all PhenMap-selected CNAs/mutations associated with MV5 and MV8 in 117 patients who received BVZ. Six CNAs/mutations were all associated (100% frequency) with PFS in the elastic net regression model (Fig. 2d). Multivariate standard Cox regression analysis on the selected six CNAs/mutations reduced the number of relevant features to two CNA regions (15q21.1 and 1p36.31 deletions) and *BRAF* mutational status. The HRs of the final three features is shown in Fig. 2e.

A risk score for each patient was generated by multiplying the normalised values of GISTIC scores for the two CNAs and *BRAF* mutation by their corresponding HRs. The relationship between the risk scores and their originating MVs (MV5 and MV8) is shown as a scatter plot with correlation co-efficients and p values in Fig. 3a & b, whilst Fig. 3c visualises these associations in a combined format, illustrating the derivation of the signature from these two prognostic MVs. Subsequently, all patients were divided into three risk groups (low, moderate, and high; designated CRC-BVZ-RiskAssigner groups) based on their risk scores by quantiles assigning the lowest 10% of risk scores to the low risk group and the highest 10% of risk scores to the high risk group, with remaining patients constituting the moderate risk group (Fig. 4a). The risk of mortality for the high-risk group (n = 12) was higher than that of the low-risk group (n = 12) (p = 1.29e-07 HR: 10.78) (Fig. 4b). Additionally, there was a significant prognostic difference between low-risk and mid-risk groups (p = 0.034 HR: 1.99). Overall, the performance of CRC-BVZ-RiskAssigner scoring model (resulting from Receiver Operating Characteristics (ROC) analysis) was shown to have area under curve (AUC) of 0.685 (95% CI 0.672–0.698) (Fig. 4c).

In a multivariate analysis incorporating age, tumour location, stage, CIN and including ANGIOPREDICT subtypes[17], the risk groups proved to be an independent predictor of PFS in patients treated with BVZ (Fig. 5a). The concordance index (C-index) of the CRC-BVZ-RiskAssigner group is 0.60 (95% CI: 0.56, 0.65) whereas groups based only on the ANGIOPREDICT subtypes has a C-index of 0.58 (Fig. 5b). The confidence intervals

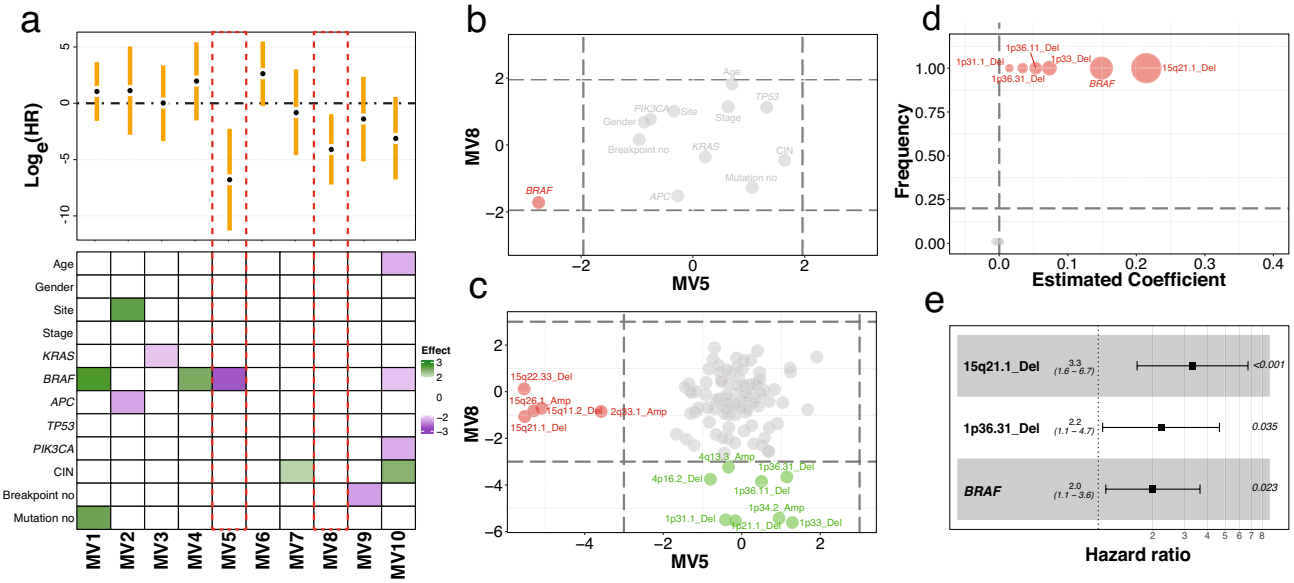

Fig. 2. Identification and phenotypic interpretation of PhenMap meta-variables (MVs) and associated genomics features in mCRC. (**a**) Upper part Forest plot assessing the prognostic value (PFS) of the ten MVs selected by PhenMap. Multivariate Cox regression estimates and corresponding 95% confidence intervals (CI) for the ten MVs selected by PhenMap applied to the ANGIOPREDICT cohort. MVs (MV5 and MV8) highlighted in red mark significant associations with PFS. Lower part Heatmap of PhenMap-scaled regression coefficients for the effect of covariates on each MVs. This phenotypic heatmap shows the correlation between MV scores and clinical variables, mutations and genomic features, to identify the phenotypic variables driving heterogeneity in each MV. "no" represents number. (**b**) Plot showing the scaled regression coefficients (dashed lines represent the 5% significance level) for MV5 and MV8. The coloured dots represent variables with significant effects on the MVs. The grey dots represent variables with non-significant effects on the MVs. 'Amp' = amplification, 'Del' = deletion. (**c**) The scaled loading coefficients showing significant CNAs associated with MV5 and MV8, coloured in the same way as (b). (**d**) Elastic net regression selects features (CNAs) most frequently (with 100 iterations) associated with PFS. (**e**) Estimated hazard ratios (HR) from the cox proportional hazard regression model, based on the three finally selected features. These features are then used to calculate risk scores. Del represents deletion/loss, and Amp represents amplification/gain.

of CIN, stage, and location was not greater than 0.5 c-index (not significant). Critically, we have shown that the high-risk group only contains patients that do not respond to BVZ when patient response was associated with risk groups. The low-risk group contained the majority of responders (10 out of 12 patients; 83.33%). The moderate-risk group was comprised of both responders and non-responders (Fig. 5c).

Discussion

Here we introduce a comprehensive multi-step machine-learning driven pipeline which fuses multi-modal omics and clinical outcome datasets, along with prognostic information and treatment response to predict patient outcome following anti-angiogenic combination therapy in the mCRC setting. Specifically, by combining a previously published method (PhenMap[4]) of multi-modal and phenomic data integration, with risk prediction machine-learning approaches, we identified prognostic risk groups from a cohort of mCRC patients who received combination BVZ therapy[17,19]. Risk of mortality was significantly greater for patients in the high-risk group, with 100% (n = 12) of patients showing no response to BVZ, compared to the low-risk group (16.67%, 2 out of 12). Overall, assignment to the high-risk group was an independent negative predictor of survival in BVZ-treated mCRC patients.

The simultaneous integration of tumour multi-modal and phenomic data, through the application of PhenMap, identified ten MVs, which associated (regression) features from different omics data types and clinical/mutational covariates in different contexts. Two of these MVs, namely MV5 and MV8, were shown to be significantly associated with PFS and were therefore selected for downstream analysis. Interestingly, we observed that MV5 is negatively associated with *BRAF* status which has been associated with poor prognosis in mCRC[23]. Indeed, all patients in the high-risk group bore the *BRAF* tumour mutation (along with the other two CNAs) and were non-responders to BVZ therapy. Elsewhere, prior data from the TRIBE clinical study (NCT00719797) indicated that patients with *BRAF* mutated tumours had longer median overall survival (mOS) when treated with FOLFOXIRI + BVZ (N = 16 patients; 19.0 months mOS) compared to FOLFIRI + BVZ (N = 12 patients; 10.7 months mOS). Nevertheless, overall these patients had shorter mOS (with either regimen) compared to *RAS* and *BRAF* wild-type patients (N = 45; 33.5 months mOS [FOLFIRI + BVZ]), N = 48; 41.7 months mOS [FOLFOXIRI + BVZ])[24]. In this context, and particularly when combined with adverse genomic alterations such as 15q21.1 or 1p36.31 deletions, the *BRAF* mutation remains strongly associated with poor outcome and appears enriched in our high-risk non-responder group.

For each MV, PhenMap provided associations of phenotypic variables with CNA patterns. Here, alterations in chromosome 15q were shown to have a significant impact on the association with MV5, while alterations in chromosome 1p were associated with MV8. Interestingly, deletions in the long arms of chromosome 15 (15q) have previously been shown to be indicators of poor prognosis and disease progression in mCRC patients[25,26], suggesting the harboring of important tumour suppressor genes. Specifically, 15q21.1 deletion is known to be associated with *B2M* gene loss, which codes for major histocompatibility complex (MHC-I)[27]. Moreover, Sheffer et al. propose a role for 1q deletion in alterations of the oxidative phosphorylation pathways leading to worse prognosis[25]. Further investigation into the role of 15q and 1q loss is warranted.

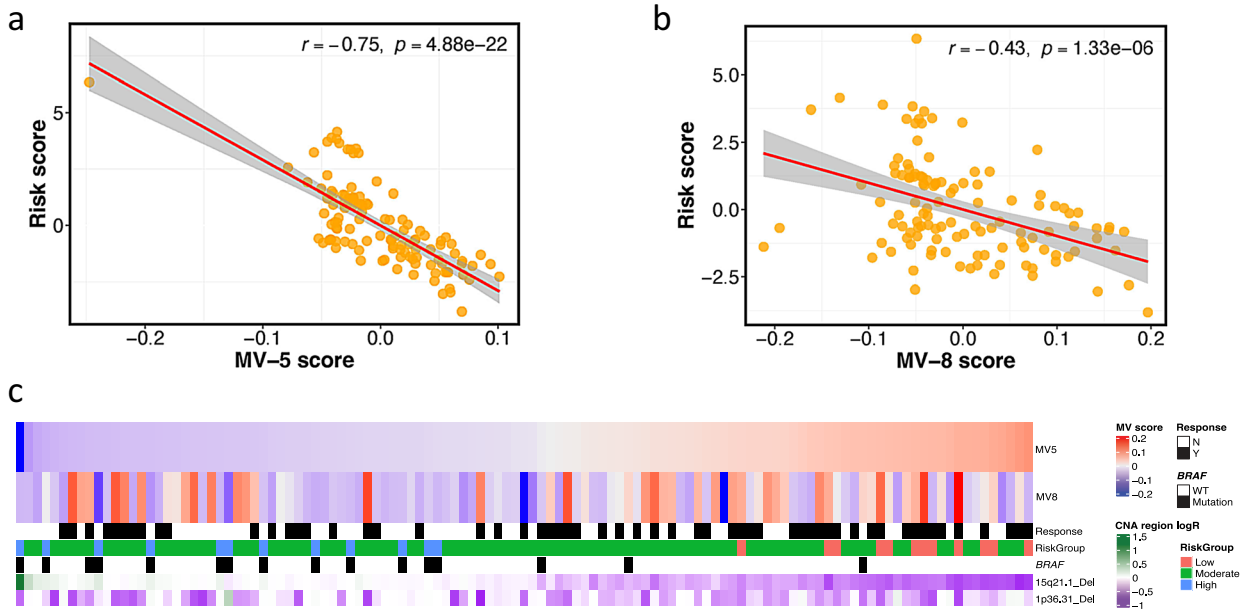

Fig. 3. The relationship between risk scores and their originating MVs. (**a**) Scatter plot (Pearson correlation) showing Risk scores and MV5 scores with a significant negative correlation. (**b**) Scatter plot (Pearson correlation) showing Risk scores and MV8 scores with a significant negative correlation. (**c**) Heatmap of significant MVs, selected features, Risk Group, and Response for all patients in the cohort.

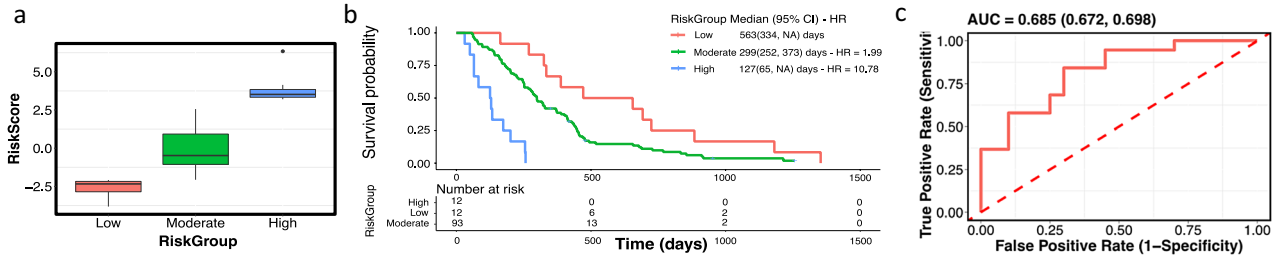

Fig. 4. PhenMap Genomic Signature of CNAs/Mutations Identifies 3 Distinct Risk Groups in BVZ Treated mCRC. (**a**) Boxplot of mCRC patients divided into three distinct risk groups (low, moderate-low, and high) based on their predicted risk scores. (**b**) Kaplan–Meier PFS plot of the three predicted risk groups in mCRC patients. Risk of mortality is significantly greater for the high-risk group, compared to the low-risk group ($p < 0.0001$). (**c**) Receiver operating characteristic curve for the prediction of response based on the Risk Score. 100 iterations of using 2/3 of the dataset for training and 1/3 to test gives the mean AUC and 95% confidence interval.

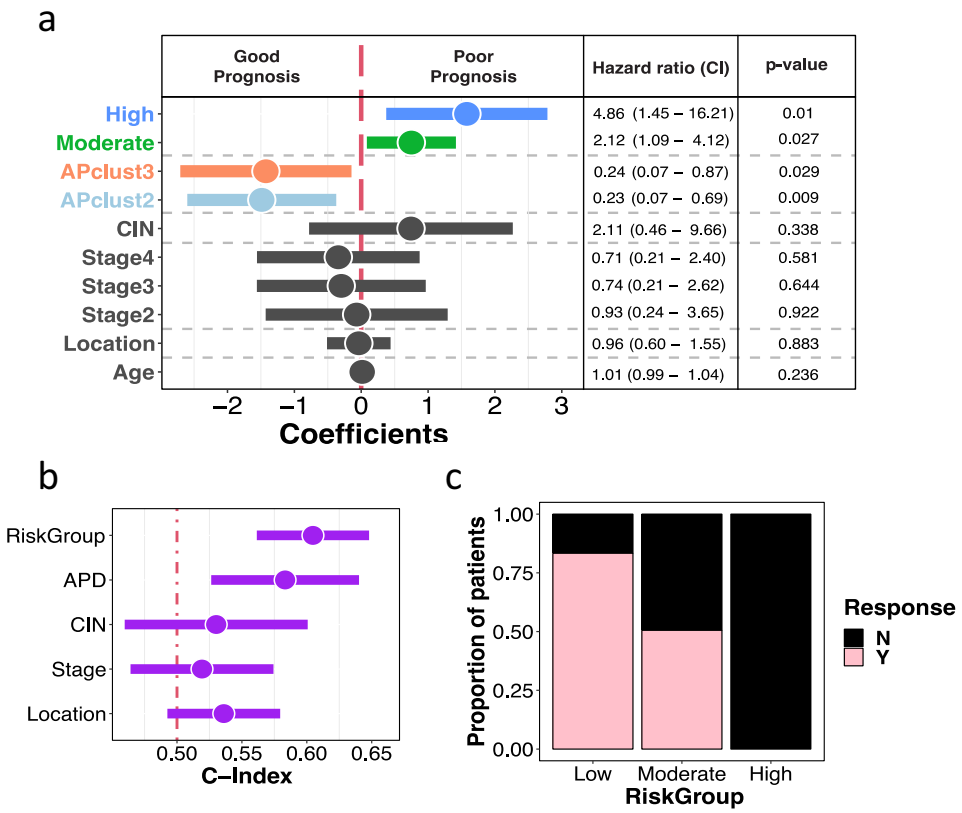

Fig. 5. Comparison of Risk Scores to ANGIOPREDICT subtypes and clinical parameters. (**a**) Multivariate cox proportional hazards model with age, tumour location, stage, CIN and including ANGIOPREDICT subtypes (derived from CNAs only) and risk groups. (**b**) The concordance index (C-index) with 95% confidence intervals of each individual model's ability to predict PFS. (**c**) Bar chart representing the proportion of responders and non-responders to BVZ in each of the three risk groups.

Having reduced the number of MV specific features through the selection of those most strongly associated with PFS, we used significant features from MV5 and MV8 to calculate a "risk score". Next, we designed a risk score classification system, whereby patients are assigned to three different risk groups (low, moderate, and high) based on their predicted response to BVZ combination therapy. We observed that high- and moderate-risk mCRC patients demonstrated worse PFS compared to low-risk mCRC patients. Moreover, the high-risk group proved to be an independent negative predictor of survival in BVZ-treated mCRC patients. Features associated with high-risk patients can thus be extracted and potentially tested as clinical biomarkers for the high-risk subtype. This would allow clinicians to identify non-responders and spare them from treatment with

BVZ and the associated side-effects, which negatively impact quality of life. Overall, in order to further develop our novel methodological pipeline, the candidate biomarker and risk groups now requires stringent validation in a separate, well-annotated, expanded cohort of mCRC patient samples. Ultimately, the PhenMap tool could, in the future, be combined with a next generation sequencing (NGS) assay such as the FoundationOne companion diagnostic which detects both CNVs and mutational burden in a single patient sample. Output from this FDA approved test could therefore be exploited as input for the PhenMap methodological pipeline. Thus, by integrating copy number, mutational, and clinical response data, robust patient outcome predictions could be accelerated.

While the findings from our integrated machine-learning framework are promising, a number of limitations warrant consideration. First, the ANGIOPREDICT cohort, while uniquely well-characterized, remains relatively modest in sample size for high-dimensional multi-omic modeling; this may limit statistical power and could cause certain associations to be cohort-specific. Second, although PhenMap effectively integrates multi-modal data, the present study primarily focused on CNAs, a limited set of gene mutations, and core clinicopathological variables.To capture additional biology relevant to BVZ resistance the addition of layers such as transcriptomics, epigenomics, proteomics, and tumour microenvironment features (e.g., spatial pathology, immune infiltration metrics) may be necessary. Third, our predictive model was trained and evaluated within a single retrospective dataset and therefore requires validation in independent, prospectively collected cohorts, ideally including patients treated under standardized therapeutic regimens and with harmonized sequencing pipelines. Finally, mechanistic interpretation of the identified CNA/mutation signature, particularly the biological pathways linking 15q21.1 and 1p36.31 loss to anti-angiogenic therapy resistance, remains incomplete. Future studies should therefore prioritize multi-cohort external validation, experimental dissection of the mechanistic consequences of these alterations, and exploration of how these biomarkers may interact with emerging targeted and immunotherapy combinations. Expansion of the pipeline to integrate additional modalities, including digital pathology and liquid biopsy features, may further enhance predictive accuracy and support translation into clinically deployable decision-support tools.

In conclusion, we have established a novel machine learning-driven fusion method to integrate multi-modal omics data with response outcomes. Employing this method, we have identified a putative combined CNA/mutation candidate biomarker with associated risk scores for the detection of mCRC patients who will not respond to BVZ combination therapy. This putative biomarker now requires further validation. Importantly, the presented machine learning tool and disruptive multi-modal data fusion analysis pipeline may now be applied to other multi-modal datasets in the oncology setting.

Data availability

The sequencing datasets analysed during the current study are available in the European Genome-Phenome Archive (https://ega-archive.org/) (EGAS00001002617). Other datasets used/generated during the current study are available from the corresponding authors on reasonable request.

Code availability

The PhenMap script is available through Github: https://github.com/syspremed/PhenMAP

Received: 14 August 2025; Accepted: 3 February 2026
Published online: 13 April 2026

References

1. Liu, T. et al. Machine learning-driven multi-targeted drug discovery in colon cancer using biomarker signatures. *NPJ Precis Oncol* **9**, 297. https://doi.org/10.1038/s41698-025-01058-6 (2025).
2. Bahrambanan, F. et al. The development of an efficient artificial intelligence-based classification approach for colorectal cancer response to radiochemotherapy: deep learning vs machine learning. *Sci. Rep.* **15**, 62. https://doi.org/10.1038/s41598-024-84023-w (2025).
3. Li, C., Wei, Y. & Lei, J. Quantitative cancer-immunity cycle modeling for predicting disease progression in advanced metastatic colorectal cancer. *NPJ Syst Biol Appl* **11**, 33. https://doi.org/10.1038/s41540-025-00513-1 (2025).
4. Nyamundanda, G., Eason, K., Guinney, J., Lord, C. J. & Sadanandam, A. A machine-learning tool concurrently models single omics and phenome data for functional subtyping and personalized cancer medicine. *Cancers (Basel)* https://doi.org/10.3390/cancers12102811 (2020).
5. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **74**, 229–263. https://doi.org/10.3322/caac.21834 (2024).
6. Morgan, E. et al. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut* **72**, 338–344. https://doi.org/10.1136/gutjnl-2022-327736 (2023).
7. Yamazaki, K. et al. Randomized phase III study of bevacizumab plus FOLFIRI and bevacizumab plus mFOLFOX6 as first-line treatment for patients with metastatic colorectal cancer (WJOG4407G). *Ann Oncol* **27**, 1539–1546. https://doi.org/10.1093/annonc/mdw206 (2016).
8. Saltz, L. B. et al. Randomized phase II trial of cetuximab, bevacizumab, and irinotecan compared with cetuximab and bevacizumab alone in irinotecan-refractory colorectal cancer: the BOND-2 study. *J Clin Oncol* **25**, 4557–4561. https://doi.org/10.1200/JCO.2007.12.0949 (2007).
9. Hurwitz, H. et al. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *N Engl J Med* **350**, 2335–2342. https://doi.org/10.1056/NEJMoa032691 (2004).
10. Pataky, R. E. et al. Real-World Cost-Effectiveness of Bevacizumab With First-Line Combination Chemotherapy in Patients With Metastatic Colorectal Cancer: Population-Based Retrospective Cohort Studies in Three Canadian Provinces. *MDM Policy Pract* **6**, 23814683211021060. https://doi.org/10.1177/23814683211021060 (2021).
11. Lambrechts, D. et al. VEGF pathway genetic variants as biomarkers of treatment outcome with bevacizumab: an analysis of data from the AViTA and AVOREN randomised trials. *Lancet Oncol* **13**, 724–733. https://doi.org/10.1016/S1470-2045(12)70231-0 (2012).

12. Schneider, B. P. et al. Association of vascular endothelial growth factor and vascular endothelial growth factor receptor-2 genetic polymorphisms with outcome in a trial of paclitaxel compared with paclitaxel plus bevacizumab in advanced breast cancer: ECOG 2100. *J Clin Oncol* **26**, 4672–4678. https://doi.org/10.1200/JCO.2008.16.1612 (2008).
13. Van Cutsem, E. et al. Bevacizumab in combination with chemotherapy as first-line therapy in advanced gastric cancer: a biomarker evaluation from the AVAGAST randomized phase III trial. *J Clin Oncol* **30**, 2119–2127. https://doi.org/10.1200/JCO.2011.39.9824 (2012).
14. de Haas, S. et al. Genetic variability of VEGF pathway genes in six randomized phase III trials assessing the addition of bevacizumab to standard therapy. *Angiogenesis* **17**, 909–920. https://doi.org/10.1007/s10456-014-9438-1 (2014).
15. Lambrechts, D., Lenz, H.-J., de Haas, S., Carmeliet, P. & Scherer, S. J. Markers of response for the antiangiogenic agent bevacizumab. *J Clin Oncol* **31**, 1219–1230. https://doi.org/10.1200/JCO.2012.46.2762 (2013).
16. van Dijk, E. Loss of Chromosome 18q11.2-q12.1 Is Predictive for Survival in Patients With Metastatic Colorectal Cancer Treated With Bevacizumab. *J. Clin. Oncol.* **36**, 2052–2060. https://doi.org/10.1200/JCO.2017.77.1782 (2018).
17. Smeets, D. et al. Copy number load predicts outcome of metastatic colorectal cancer patients receiving bevacizumab combination therapy. *Nat Commun* **9**, 4112. https://doi.org/10.1038/s41467-018-06567-6 (2018).
18. ANGIOPREDICT Consortium. https://cordis.europa.eu/project/id/278981 n.d.
19. Betge, J. et al. Outcome of Colorectal Cancer Patients Treated with Combination Bevacizumab Therapy: A Pooled Retrospective Analysis of Three European Cohorts from the Angiopredict Initiative. *Digestion* **94**, 129–137. https://doi.org/10.1159/000449412 (2016).
20. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* https://doi.org/10.1186/gb-2011-12-4-r41 (2011).
21. Cox, D. R. Regression Models and Life-Tables. *J R Stat Soc Series B Stat Methodol* **34**, 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x (1972).
22. Zou, H. & Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J R Stat Soc Series B Stat Methodol* **67**, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x (2005).
23. Tabernero, J., Ros, J. & Élez, E. The Evolving Treatment Landscape in BRAF-V600E-Mutated Metastatic Colorectal Cancer. *Am Soc Clin Oncol Educ Book* **42**, 1–10. https://doi.org/10.1200/EDBK_349561 (2022).
24. Cremolini, C. et al. FOLFOXIRI plus bevacizumab versus FOLFIRI plus bevacizumab as first-line treatment of patients with metastatic colorectal cancer: updated overall survival and molecular subgroup analyses of the open-label, phase 3 TRIBE study. *Lancet Oncol* **16**, 1306–1315. https://doi.org/10.1016/S1470-2045(15)00122-9 (2015).
25. Sheffer, M. et al. Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. *Proc Natl Acad Sci U S A* **106**, 7131–7136. https://doi.org/10.1073/pnas.0902232106 (2009).
26. Poetsch, M. et al. Loss of heterozygosity at 15q21.3 correlates with occurrence of metastases in head and neck cancer. *Mod. Pathol.* **19**, 1462–1469. https://doi.org/10.1038/modpathol.3800666 (2006).
27. Sucker, A. & Paschen, A. Deciphering the genetic evolution of T-cell resistance in melanoma. *Oncoimmunology* **4**, e1005510. https://doi.org/10.1080/2162402X.2015.1005510 (2015).

Acknowledgements

The authors would like to acknowledge the Centre for Translational Immunotherapy for supporting us through A. Lärkeryd. We also thank Colum Connolly (RCSI) for his assistance with figure preparation, particularly the flow diagram. We thank Joan Kehoe for her contribution to the tissue collection that enabled this study.

Author contributions

V.T: Data curation, formal analysis, investigation, writing–original draft. G.N: original method development, development of machine-learning approaches and packages, data curation, investigation, formal analysis, visualization, writing-original draft. A.L: Formal analysis, methodology, writing– review and editing, writing–original draft. H. PS: Data analysis and assisting with packages and visualization. I.S. M: writing–original draft, writing–review, and editing. T. V: formal analysis, investigation. D.S: Resources, investigation. B.B: Resources, investigation. J.B: Resources, investigation. M.P.E: Resources, investigation, supervision. T.G: Resources. V.M: formal analysis, visualization, methodology, resources. E.K: Resources, investigation. H.M.V: Resources. A.C.O: writing–review and editing, project administration. C.C: Resources. F.M: Resources. W.M.G: Resources. A.B: Methodology, Resources. R.K: Methodology. B.F: Methodology. B.Y: Investigation, Resources. N.vG: Resources. D.A.M: Resources. B.T.H: Resources. S.D: Resources. B.M: Data analysis. D.P.O: Resources. R.D: Resources, writing–review, and editing. D.L: Resources, supervision, methodology. J.H.M.P: Resources, Supervision, funding acquisition, writing–review, and editing. A.S: Conceptualization, resources, supervision, funding acquisition, visualization, methodology, writing–original draft, writing–review, and editing, project administration, A.T.B: Conceptualization, resources, supervision, funding acquisition, writing–original draft, project administration, writing–review, and editing. All authors reviewed the manuscript.

Funding

This work was supported by the European Commission Framework Programme Seven (FP7) initiative 'ANGIOPREDICT' [grant number 278981]. This work was further supported by a Health Research Board (HRB) Investigator Led Project (ILP) [grant number ILP-POR-2019–066] awarded to A.T.B and D.L and a Science Foundation Ireland Frontiers for the Future Project [grant number 20/FFP-P/8884] awarded to A.T.B, D.L and A.S. A.T.B, M.E, V.M, D.L, R.D, D.P.O, J.H.M.P and A.S were further supported by the European Union's Horizon 2020 Health Research and Innovation award "COLOSSUS" [grant number 754923]. This research was partly supported by the Institute of Cancer Research and The Ian Harty Charitable Trust to A.S.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Informed consent was obtained from each patient and the study was carried out in accordance with the relevant guidelines and regulations effective for the participating study centers. Specifically, ethical review and approval of health-related research studies which are not clinical trials of medicinal products for human use as defined in SI 190/2004 was undertaken by the Research Ethics Committee of Beaumont Hospital, RCSI Hospital Group in Ireland. The study was approved by the Medical Ethics Commission of the Faculty of Medicine in Mannheim, Heidelberg University in Germany. The collection, storage, and use of archival tissue and patient data were performed in compliance with the Code for Proper Secondary Use of Human Tissue in The Netherlands (2011), developed by the standing committee of the FEDERA (FMWV Foundation, Chamber of Commerce, Rotterdam, 41055219, http://www.fmwv.nl and www.federa.org) and the Commissie Regelgeving in Onderzoek (COREON), in close cooperation with BBMRI-NL and patient organisations.

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

During the preparation of this work, the authors used ChatGPT to query for additional information, improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Additional information

Correspondence and requests for materials should be addressed to A.S. or A.T.B.