Article

Automated real-world data integration improves cancer outcome prediction

https://doi.org/10.1038/s41586-024-08167-5

Received: 9 January 2024

Accepted: 8 October 2024

Published online: 6 November 2024

Open access

Check for updates

Justin Jee[1,5], Christopher Fong[1,5], Karl Pichotta[1,5], Thinh Ngoc Tran[1,5], Anisha Luthra[1,5], Michele Waters[1], Chenlian Fu[1], Mirella Altoe[1], Si-Yang Liu[1], Steven B. Maron[1,2], Mehnaj Ahmed[1], Susie Kim[1], Mono Pirun[1], Walid K. Chatila[1], Ino de Bruijn[1], Arfath Pasha[1], Ritika Kundra[1], Benjamin Gross[1], Brooke Mastrogiacomo[1], Tyler J. Aprati[2], David Liu[2], JianJiong Gao[3], Marzia Capelletti[3], Kelly Pekala[1], Lisa Loudon[1], Maria Perry[1], Chaitanya Bandlamudi[1], Mark Donoghue[1], Baby Anusha Satravada[1], Axel Martin[1], Ronglai Shen[1], Yuan Chen[1], A. Rose Brannon[1], Jason Chang[1], Lior Braunstein[1,2], Anyi Li[1], Anton Safonov[1], Aaron Stonestrom[1], Pablo Sanchez-Vela[1], Clare Wilhelm[1], Mark Robson[1,2], Howard Scher[1,2], Marc Ladanyi[1], Jorge S. Reis-Filho[1], David B. Solit[1], David R. Jones[1], Daniel Gomez[1], Helena Yu[1], Debyani Chakravarty[1], Rona Yaeger[1,4], Wassim Abida[1,4], Wungki Park[1,4], Eileen M. O'Reilly[1,4], Julio Garcia-Aguilar[1,4], Nicholas Socci[1], Francisco Sanchez-Vega[1], Jian Carrot-Zhang[1], Peter D. Stetson[1], Ross Levine[1,4], Charles M. Rudin[1,4], Michael F. Berger[1], Sohrab P. Shah[1], Deborah Schrag[1,4], Pedram Razavi[1,4], Kenneth L. Kehl[2], Bob T. Li[1,4], Gregory J. Riely[1,4], Nikolaus Schultz[1]✉ & MSK Cancer Data Science Initiative Group*

The digitization of health records and growing availability of tumour DNA sequencing provide an opportunity to study the determinants of cancer outcomes with unprecedented richness. Patient data are often stored in unstructured text and siloed datasets. Here we combine natural language processing annotations[1,2] with structured medication, patient-reported demographic, tumour registry and tumour genomic data from 24,950 patients at Memorial Sloan Kettering Cancer Center to generate a clinicogenomic, harmonized oncologic real-world dataset (MSK-CHORD). MSK-CHORD includes data for non-small-cell lung (n = 7,809), breast (n = 5,368), colorectal (n = 5,543), prostate (n = 3,211) and pancreatic (n = 3,109) cancers and enables discovery of clinicogenomic relationships not apparent in smaller datasets. Leveraging MSK-CHORD to train machine learning models to predict overall survival, we find that models including features derived from natural language processing, such as sites of disease, outperform those based on genomic data or stage alone as tested by cross-validation and an external, multi-institution dataset. By annotating 705,241 radiology reports, MSK-CHORD also uncovers predictors of metastasis to specific organ sites, including a relationship between *SETD2* mutation and lower metastatic potential in immunotherapy-treated lung adenocarcinoma corroborated in independent datasets. We demonstrate the feasibility of automated annotation from unstructured notes and its utility in predicting patient outcomes. The resulting data are provided as a public resource for real-world oncologic research.

The ubiquity of electronic health records offers a largely untapped data substrate for translational medicine. Although abstraction of key elements from free-text patient visit, radiology, histopathology and procedural notes has traditionally limited analysis, natural language processing (NLP) now allows for automatic annotation of such features[1,2]. Massive, context-aware transformer architectures[3], including those pretrained on health records[4,5], have reshaped the NLP landscape and have shown promise at a number of medical tasks including predicting hospital readmission[4] and providing medical advice[6]. In oncology, immunohistochemistry[7] and clinical tumour sequencing[8,9] are standard of care for many patients because of their potential to guide therapy. Combining real-world data (RWD) has enormous potential to aid in prediction of tumour trajectories.

The separation of hospital, academic and commercial entities responsible for genomic sequencing, radiology, histopathology and electronic health record data is a hurdle to integrative analysis[10]. Several studies have begun to overcome these silos (for example, through the integration of tumour sequencing with treatment data to uncover genomic modifiers of response[11], or the integration of billing codes to uncover mutations associated with specific organ sites of metastasis[12]).

[1]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Dana Farber Cancer Institute, Boston, MA, USA. [3]Caris Life Sciences, Irving, TX, USA. [4]Weill Cornell Medicine, Cornell University, New York, NY, USA. [5]These authors contributed equally: Justin Jee, Christopher Fong, Karl Pichotta, Thinh Ngoc Tran, Anisha Luthra. *A list of authors appears at the end of the paper. ✉e-mail: schultzn@mskcc.org

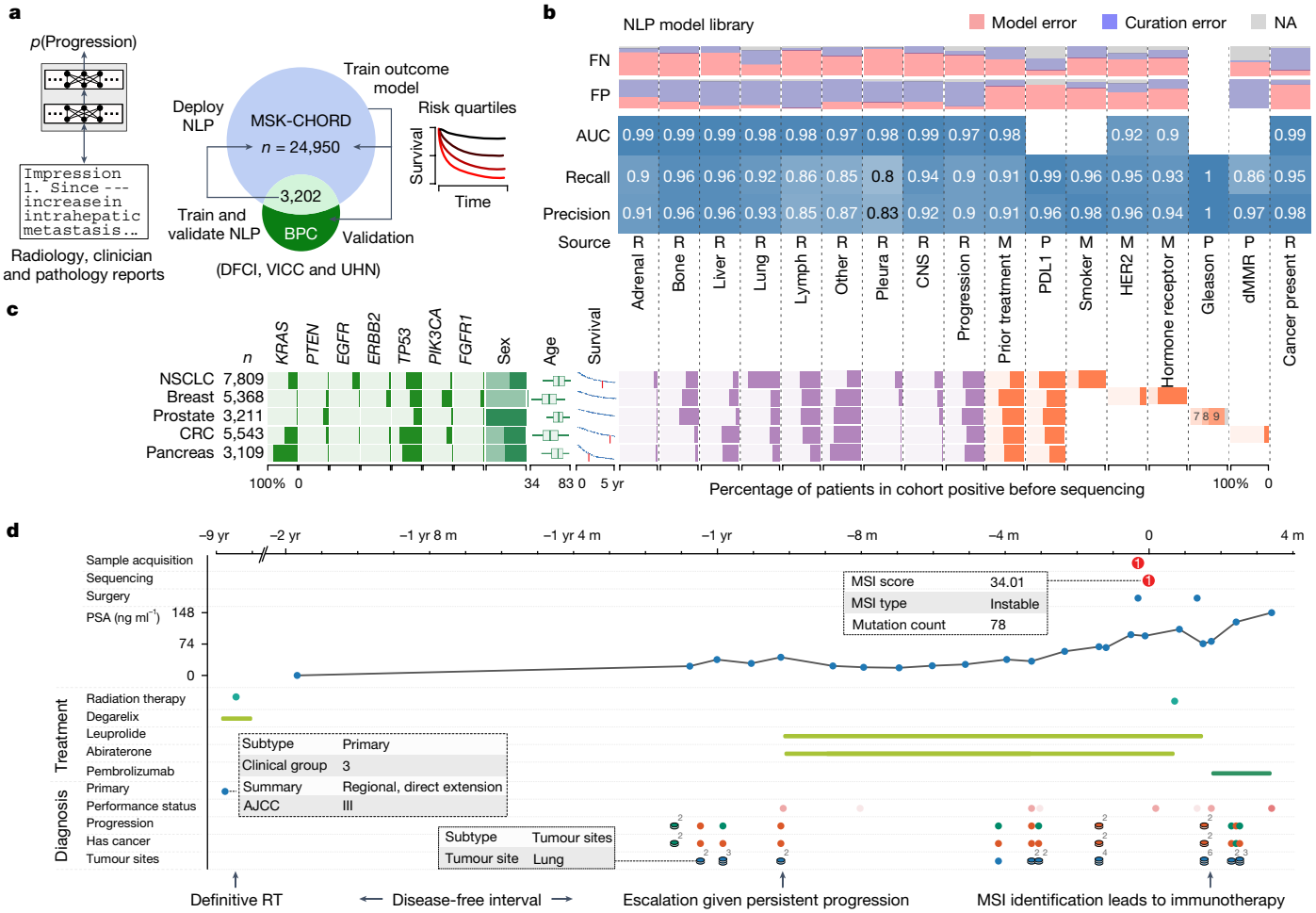


Fig. 1 | Study overview. a, Creating MSK-CHORD. *p*, probability; DFCI, Dana Farber Cancer Institute; UHN, University Health Network; VICC, Vanderbilt-Ingram Cancer Center. **b**, NLP model library performance assessed by either cross-validation or held-out validation in the MSK-BPC cohort (Methods). Source text includes radiology impressions (R), medical oncology notes (M) or histopathology reports (P). Randomly selected false positive (FP) and false negative (FN) cases were independently reviewed to audit reasons for model failure; in several cases (purple), the original curation labels were incorrect. Raw numbers are given in Supplementary Table 1. NA, not applicable; that is, an independent curator determined that the source document did not actually contain sufficient information to determine the status of the variable in question. **c**, MSK-CHORD characteristics overview. Age box plots show median, quartiles and ±95th percentile. Bar charts show proportion of patients with a given feature. Genomic alterations include only those annotated as oncogenic by OncoKB and were derived from tumour biopsy sequencing by MSK-IMPACT. Age, sex (male reference) and survival outcomes were derived from structured data. Kaplan–Meier survival curves for the individual cohorts are shown with median survival denoted by a red hash mark. Bar charts represent the percentage of patients with a given characteristic at time of cohort entry. Additional characteristics in MSK-CHORD such as tumour stage, specific institutional treatments and tumour markers are not shown. **d**, Visualizing patient-level data in cBioPortal, in this case a patient (P-0050196) with prostate adenocarcinoma who was treated with definitive radiation for stage III disease, and then developed metastatic recurrence in the lung and received treatment with multiple lines of therapy including pembrolizumab for MSI found on MSK-IMPACT. m, months; PSA, prostate-specific antigen; AJCC, American Joint Committee on Cancer; RT, radiation therapy.

Models incorporating more detailed clinical, genomic, radiomic and histopathologic data[13] have shown promise in better risk stratification (for example, following immunotherapy[14,15]), although these efforts frequently rely on and are limited by manual extraction of key data elements and are studied in cohorts of modest size.

In this study, we used a large, integrated dataset to develop improved models of cancer outcome. Specifically, we sought to overcome bottlenecks of manual extraction for RWD by developing methods to automatically annotate free-text clinician notes as well as radiology and histopathology reports, and then to combine these annotations with structured treatment, survival, tumour registry, demographic and tumour genomic data to create MSK-CHORD; test whether MSK-CHORD can uncover clinicogenomic associations not apparent in smaller datasets; study whether integrated, multimodal models would outperform traditional single-modality models, including American Joint Committee on Cancer stage, at predicting overall survival (OS); and identify genomic features associated with metastasis to specific organs.

Automatic annotation of free-text notes

To develop algorithms that automatically annotate free-text reports, we leveraged the Project GENIE Biopharma Collaborative (BPC) dataset of the American Association for Cancer Research[16], a structured curation of electronic health records including those for patients with non-small-cell lung (NSCLC), breast, colorectal, prostate and pancreatic cancer at four cancer centres using the PRISSMM method[17]. We trained and validated NLP transformer models using BPC-curated annotations derived from specific radiology, histopathology or clinical notes with corresponding records at Memorial Sloan Kettering Cancer Center (MSK), an academic cancer centre in New York, NY (MSK-BPC, *n* = 3,202 patients with 38,719 corresponding radiology

reports; Fig. 1a), to annotate features requiring nuanced interpretation of language such as negation or context: cancer progression, sites of tumours and the presence of any cancer from the impression section of radiology reports; prior outside treatment from clinician initial visit notes; and hormone receptor and HER2 receptor status from clinician initial visit or follow-up notes. We created additional rule-based models to annotate features stored in a more structured format (that is, smoking status from clinician notes, as well as Gleason score, PDL1 (also known as CD274) status and mismatch repair (MMR) deficiency from histopathology reports).

Transformers were validated using fivefold cross-validation; rule-based models were created on the basis of annotations from previously published cohorts[14,18,19] and validated with MSK-BPC annotations (Methods section NLP models). All NLP models had an area under the curve (AUC) of >0.9 and precision and recall of >0.78 when treating manually curated labels as ground truth, with several models achieving precision and recall of >0.95 (Fig. 1b and Supplementary Table 1). A random sample of instances in which model predictions and curation labels were discrepant were retrospectively reviewed by clinicians, who found that many of the original curation labels were incorrect, with the NLP annotations inferred correctly (Fig. 1b and Supplementary Table 1). More 'confident' transformer probability scores (that is, closer to 0 or 1) were associated with a greater likelihood of curator error across radiographic annotation tasks (Supplementary Fig. 1). In annotating HER2 and hormone receptor status, we observed multiple instances among discrepant cases that could be explained by complex clinical situations, highlighting challenges for both human and NLP curation methods for certain tasks (Supplementary Discussion).

We tested the extent to which NLP model choice affected annotation quality, evaluating several models. Transformer architectures consistently outperformed logistic regression and feed-forward neural network approaches (Supplementary Fig. 2 and Supplementary Table 3). Model performance was dependent on training sample size and the number of positive samples per class (Supplementary Figs. 3 and 4 and Supplementary Discussion); tumour site models for reproductive organs, for example, had worse performance as a result of fewer positive examples in training data. Tumour site annotation was also modestly improved by using a single joint classifier rather than separate, individual classifiers (Supplementary Fig. 5). We also compared the accuracy of NLP-derived annotations for metastatic sites to those of billing codes for those sites. In a patient-wise analysis, NLP-derived annotations had better accuracy for metastatic site involvement than billing codes, with precision and recall improvements ranging from 0.03 to 0.32 (Supplementary Table 3).

We assessed heterogeneity in NLP model performance in specific individual cancer types. In general, models performed comparably well across cancer types, except for identification of prior treatment in NSCLC, for which the precision was 0.78 although the AUC was 0.98 and the recall was 0.92 (Supplementary Table 4).

To test the extent to which our NLP models generalize to cancer types absent from training data, we performed hold-one-cancer-out experiments in which NLP models were trained on four out of the five cancer type cohorts in the MSK-BPC dataset and validated in the held-out cancer type. In these experiments, models had similar precision and recall to those in fivefold cross-validation (Supplementary Table 5), suggesting potential generalizability to out-of-distribution datasets. In summary, NLP can annotate free-text oncologic notes with an accuracy approaching that of manual curation across cancer types.

Assembling MSK-CHORD

To allow for integration of data at scale, we sought to create a single cohort containing clinical, radiographic, histopathologic, laboratory and tumour genomic sequencing data. MSK-CHORD combines NLP-derived features with institutional demographic, treatment and tumour registry data, along with tumour genomic profiling using MSK-IMPACT, a Food and Drug Administration-authorized, targeted sequencing assay[20] with matched blood sequencing to filter germline and clonal haematopoiesis variants. MSK-CHORD is at least six times larger than the underlying BPC training data across NSCLC, breast, colorectal, prostate and pancreatic cancer while containing its core clinical data elements (Fig. 1c and Supplementary Table 6). NLP-derived patient characteristics, such as metastatic site incidence, were similar among BPC and MSK-CHORD, suggesting the validity of our NLP approach. However, as a more modern cohort, MSK-CHORD had more modern diagnostic and therapeutic characteristics, such as higher rates of PDL1 testing, than BPC (Supplementary Table 7). MSK-CHORD is available through cBioPortal, allowing for additional visualization and cohort selection[21] (Fig. 1d).

Discovery of associations in MSK-CHORD

The modest size of many manually curated cohorts often leads to insufficiently powered analyses, impeding discovery of meaningful associations. For example, PDL1 expression is a known biomarker of response to immunotherapy in NSCLC; however, of patients with NSCLC in MSK-BPC treated with immunotherapy and PDL1 testing ($n = 29$), there was equivocal evidence that PDL1 (≥1% 'positive' versus <1% 'negative') was associated with longer OS (hazard ratio 0.58, 95% confidence interval (CI) 0.11–1.1, $P = 0.07$). MSK-CHORD showed a similar magnitude of benefit, but with 754 patients with NSCLC receiving immunotherapy at time of cohort entry with PDL1 testing, statistical power was greater (hazard ratio 0.64, 95% CI 0.54–0.77, $P < 0.001$; Fig. 2a).

Genomic alterations may be associated with prior treatment, but the size of the MSK-BPC cohort precluded discovery of enrichment of several known post-treatment alterations (Fig. 2b). At the same time, many patients receive treatment at multiple centres, making analysis based on prior treatment challenging. Using MSK-CHORD, we found that, as expected, *ESR1*, *CCND1* and *NF1* mutations in breast cancer[22], *EGFR*^{T790M} and *MET* amplifications in *EGFR*-mutant NSCLC[23], *AR* and *TP53* mutations in prostate cancer[24], and clonal haematopoiesis *CHEK2*, *PPM1D* and *TP53* mutations[25] were enriched in patients exposed to prior systemic therapy as annotated by NLP (Fig. 2b). As expected, patients with known, institutionally administered treatments before sample acquisition also had enrichment in those alterations (Fig. 2b). Thus, MSK-CHORD's size enables adequately powered identification of post-treatment mutations across multiple cancers, and NLP-derived prior treatment is an important complement to institutional treatment records in such analyses.

Similarly, small studies have suggested a higher incidence of *TP53* and *PTEN* loss and homologous recombination deficiency in patients with prostate cancer of high Gleason grade[26]. After multiple-hypothesis correction, the MSK-BPC was underpowered to discover significant associations between tumour genomics and Gleason score (Fig. 2c). In MSK-CHORD, we observed a dose-dependent relationship between NLP-annotated highest Gleason grade and several gene-level alterations including *TP53*, *PTEN* and *BRCA2* (Fig. 2c). Thus, our cohort allows for validation of proposed genomic–histopathologic associations.

MSK-CHORD's size also enables analyses of patients with less common combinations of features. For example, among patients with stage IV colorectal cancer (CRC), microsatellite instability (MSI) on genomic sequencing or MMR deficiency (dMMR) on immunohistochemistry are two highly concordant biomarkers of response to immunotherapy[27]. However, some patients have a rare combination of these factors (that is, either MSI on genomic sequencing and proficient MMR (pMMR) on immunohistochemistry, a possible result of MMR gene mutations[28], or dMMR and microsatellite stability (MSS) on genomic sequencing). Leveraging MSK-CHORD's size, after excluding patients with equivocal MSI status, we identified ten patients with such discrepancies between dMMR and MSI status treated with

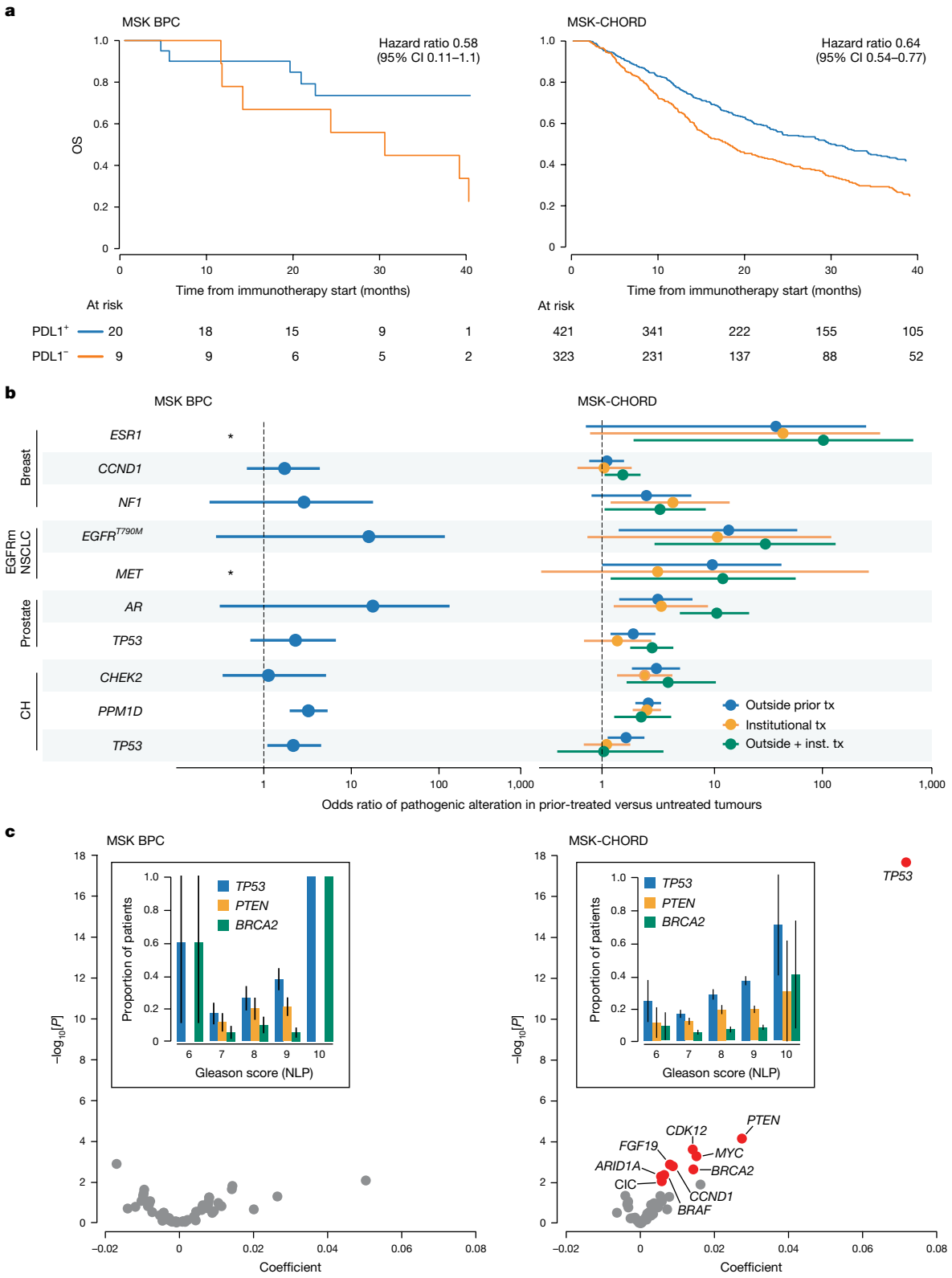


Fig. 2 | Using MSK-CHORD for adequately powered clinicogenomic analysis. **a**, Kaplan–Meier curves depicting OS and hazard ratios for patients with NSCLC treated with immune checkpoint blockade at time of cohort entry to time of death, stratified by PDL1 status in the MSK-BPC cohort and MSK-CHORD cohort. **b**, Left: odds ratios ± 95% CI for known post-treatment alterations in the smaller, manually curated MSK-BPC cohort. Right: odds ratios ± 95% CI for known post-treatment alterations in the MSK-CHORD cohort, stratified by NLP-identified or institutionally given prior treatment (tx) or both. inst., institutional. Clonal haematopoiesis (CH) analyses are performed using a subset of MSK-CHORD with previously published clonal haematopoiesis calls[25]. *0/34 patients with breast cancer without prior treatment in MSK-BPC had *ESR1* alterations, and 0/30 patients with EGFR-mutant (EGFRm) NSCLC without prior treatment had MET alterations; hence, the odds ratio is infinity for these groups. **c**, Proportion of patients with prostate cancer with the listed gene alterations (oncogenic by OncoKB) as a function of Gleason score (NLP-derived) in the MSK-BPC cohort ($n = 561$) and MSK-CHORD cohort ($n = 3,211$). Volcano plots show slope coefficients and two-sided P values from linear regression, with dots in red showing relationships with multiple-hypothesis-corrected q values < false discovery rate 0.05 by Benjamini–Hochberg method, and insets show proportions of the total cohort for selected genes ± binomial 95% CI.

immunotherapy. Manual review of these cases suggests that these discrepancies could not be explained by NLP errors or sequencing artefacts such as low tumour purity (Supplementary Table 8). As expected, patients with dMMR and MSI had better survival than those with pMMR and MSS. Patients with discordant MMR and MSI status had a longer time on immunotherapy than those with pMMR and MSS but a shorter time on immunotherapy than those with dMMR and MSI (Extended Data Fig. 1). Together, these results indicate that patients with discrepant genomic and immunohistochemical results do benefit from and should be offered immunotherapy, but the absence of either positive biomarker may have prognostic importance.

Similarly, MSK-CHORD included patients with lung cancer who were current or former smokers[16] but whose tumours lacked smoking mutational signatures[29]. In these cases, *EGFR* and *KRAS* drivers, classically associated with non-smokers and smokers, respectively[30], were seen in similar proportions, suggesting that neither genomic nor clinical data are sufficient on their own to predict tumour biology in these cases (Extended Data Fig. 2 and Supplementary Discussion).

MSK-CHORD also has important differences compared with previous large-scale tumour genomic profiling efforts such as The Cancer Genome Atlas including size, modernity and clinical annotations[31] (Supplementary Table 9). These differences allow for discovery of relationships between, for example, tumour genomic alterations and OS in lung adenocarcinoma (LUAD) not apparent in The Cancer Genome Atlas (Extended Data Fig. 3). These results show that NLP-derived features in MSK-CHORD have meaningful biologic correlations, although caution should be taken to ensure that confounders are considered. For example, in MSK-CHORD, *EGFR* is associated with better OS in LUAD, but in multivariate analysis, it was receipt of targeted therapy that is associated with better survival, not *EGFR* mutation status itself (see Supplementary Discussion on OS modelling). MSK-CHORD enables adequately powered discovery of clinicogenomic associations including those among patients with less common characteristics.

Integrated multimodal models for OS

To study whether models combining the orthogonal data elements in MSK-CHORD improve prediction of cancer outcomes, we constructed random survival forest (RSF) models to predict OS from time of cohort entry (that is, sequencing report date; see the Methods section OS modelling). We systematically tested the performance of models trained on related subsets of variables (that is, stage, demographics, genomic drivers, pathology, tumour markers, treatment and tumour sites) and compared them with models with access to all variables to assess the benefit of multimodal integration. We tested these models using both fivefold internal cross-validation and external validation on the non-MSK portion of BPC.

As expected, among patients with stage I–III disease and no progression event before cohort entry, most RSFs trained on tumour stage-related variables (that is, stage at diagnosis and time since diagnosis) had prognostic value. The fivefold cross-validation *c* indices for tumour stage-related variables ranged from 0.53 (95% CI 0.51–0.55) for pancreatic cancer to 0.78 (95% CI 0.76–0.80) for breast cancer. However, in all cancer types, combined multimodal models that used all variables outperformed those based on tumour stage and prior progression alone (Fig. 3a and Supplementary Table 10).

Among patients with stage IV disease, models trained only on tumour site data, an NLP-derived feature, had greater prognostic value than models based on genomic drivers alone across all cancer types. In prostate, colorectal and pancreatic cancer, tumour markers were the single modality with the greatest prognostic value; however, in all cancer types, a full multimodal model had greater prognostic power than models trained on tumour marker or site data alone (Fig. 3a). Different specific features were key to multimodal prognostication for different cancer types, although because of frequent correlation of variables across classes, no single class of variables emerged as the most necessary for OS prediction across cancer types (Supplementary Discussion and Supplementary Figs. 6 and 7).

The performance of full multimodal models varied by cancer type, ranging from a *c* index of 0.58 for stage IV pancreatic cancer to 0.83 for stage I–III breast cancer (Fig. 3a). Model performance was generally consistent in both fivefold internal and external validation, apart from tumour markers as a single-modality category in colorectal, pancreatic and breast cancer (Fig. 3b). In these cancer types, because of sparser tumour marker data among BPC patients relative to those in MSK-CHORD, fivefold cross-validation showed substantial prognostic value in tumour marker data, but these models achieved lower performance in the non-MSK BPC validation cohort (Fig. 3b and Supplementary Table 10).

Model architecture and start time selected for OS analysis may influence model results. We performed sensitivity analyses examining OS using a variety of different time-to-event machine learning architectures with a start time of diagnosis left-truncated at time of cohort entry. In these analyses, model performance per cancer type and the importance of multimodal variables for predicting OS were observed despite differences in start time and architecture (Supplementary Discussion and Supplementary Table 11).

Thus, models incorporating multiple data streams including NLP-derived variables had superior discriminative power for predicting OS. MSK-CHORD can serve as a core dataset to which numerous other variables might be feasibly added for OS analysis (Extended Data Fig. 4 and Supplementary Discussion). Our results indicate that multimodal biomarkers are superior to disease stage for prognostication. For example, in both NSCLC and pancreatic cancer, a high-risk subset of patients with stage I–III disease is predicted to have a higher risk of mortality than a low-risk subset of patients with stage IV disease (Fig. 3c and Extended Data Fig. 5). Among patients with stage IV NSCLC in those cancer types, there is a difference in survival of several years captured in different risk quartiles (Fig. 3c).

Direct application of transformers to radiology report text may improve prognostication when compared to more interpretable models trained on a small number of annotated variables such as tumour sites and disease stage. We fine-tuned a transformer pretrained on clinical text (radLongformer) to predict mortality within 6 months from radiology reports of computed tomography scans of the chest, abdomen and pelvis (Methods section radLongformer). In all five cancer types, this model had prognostic power for OS. In stage IV CRC, it had superior prognostic power to metastatic sites for predicting OS; however, in no other instances did radLongformer have superior prognostic power to tumour sites alone in predicting OS (Extended Data Fig. 4b). Adding risk scores from radLongformer as a variable to our RSFs did not improve prognostic power. Our results indicate that for predicting cancer mortality, an interpretable model with sufficient variables can perform comparably to a 'black box' neural network model trained directly on free text.

Multimodal models may improve prediction of OS compared to traditional models based on a single modality such as stage. By using an interpretable, late-fusion[13] framework, we identify specific classes of variables, such as metastatic sites, that are an important source of prognostic information.

Genomic predictors of metastatic sites

Metastasis is the leading cause of cancer mortality, and metastatic colonization has clinical implications, such as the frequency of surveillance imaging with magnetic resonance imaging for detection of brain metastases[32]. However, genomic predictors of metastatic tropism are poorly understood[12,33]. We used NLP to annotate the presence or absence of metastatic sites of disease in all 705,241 radiology reports longitudinally for patients in MSK-CHORD (Methods section

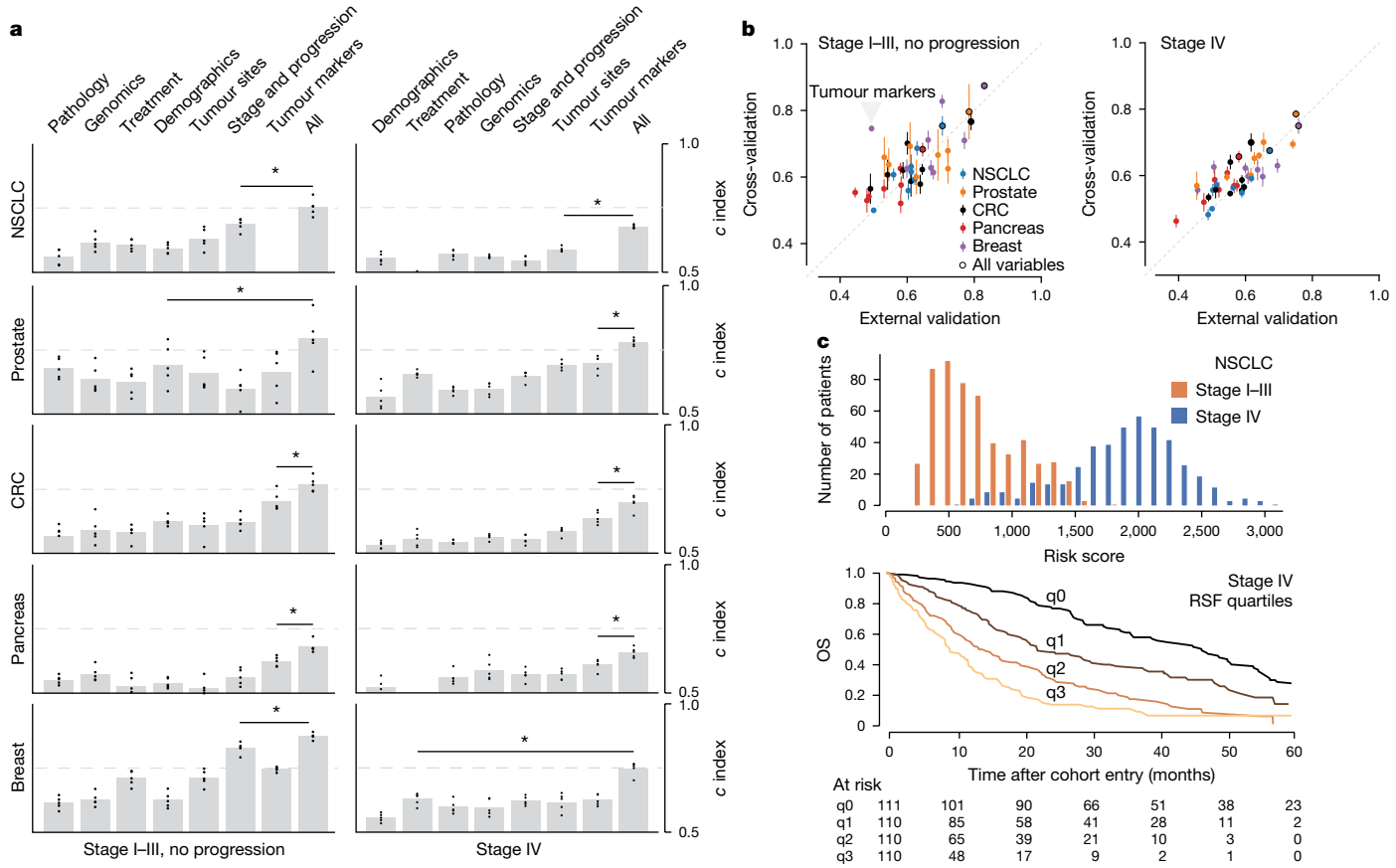


Fig. 3 | Integrated multimodal models predict OS. a, *c* indices from RSFs by cancer type, stage and data modality (*x* axis) validated in fivefold cross-validation. 'All' denotes incorporation of all listed modalities into the model. *$P < 0.05$, unadjusted for multiple hypotheses, by one-sided *t*-test compared to next best-performing model. For stage IV NSCLC, prostate, CRC, pancreas and breast cancer, *P* values are 2×10^{-7}, 0.0003, 0.005, 0.003 and 3×10^{-5}, respectively. For stage I–III disease, *P* values are 0.003, 0.049, 0.008, 0.001 and 0.002, respectively. **b**, Scatter plot comparing mean *c* indices from fivefold cross-validation within MSK-CHORD (error bars represent ±95% CI) versus *c* indices from the same models trained on the entire MSK-CHORD cohort of a given cancer type and tested on an external validation dataset (the corresponding non-MSK-BPC cohorts). Colour legend is the same in both plots. For **a**,**b**, total number of patients in each cohort is given in Supplementary Table 6. **c**, Risk score distribution for the non-MSK-BPC cohorts and survival curves based on computed risk quartiles for patients with NSCLC.

NLP models). Colonization of specific organs may thus be used as an endpoint for time-to-event analyses.

We studied time-to-specific organ metastasis from time of tumour sampling for four specific organ sites of disease with clinical relevance and accurate annotations using NLP: central nervous system (CNS), bone, liver and lung, adjusting for disease stage, prior treatment and histologic subtype. The rate at which patients developed radiographic evidence of disease at those sites was similar in the manually curated BPC versus NLP-derived cohorts (Fig. 4a). We sought to use MSK-CHORD to study whether oncogenic alterations[34] in specific genes were associated with rates of metastasis to those organ sites. We found several associations between genomic alterations and development of future organ metastases (Fig. 4b) despite controlling for histologic subtype, which itself is associated with metastasis to specific organs (Supplementary Fig. 8). Some gene alterations, such as *TP53* and *CDKN2A* in LUAD, were associated with metastases to all sites. Conversely, *RB1* alteration was associated with CNS and liver, but not bone and lung, metastases in LUAD, hormone-receptor-positive breast and prostate cancer. If tumours with *RB1* alterations are more likely to metastasize to the brain and liver, tumours in those sites may be more likely to harbour such alterations. Examination of *RB1* alteration prevalence based on site of disease sequenced revealed enrichment of oncogenic *RB1* alterations in brain and liver metastases (Extended Data Fig. 6).

Aggregating alterations at the pathway[35] level uncovered further specific associations with propensity to metastasize to specific organ sites among cancer types. *TP53* pathway alterations were associated with higher rates of liver but lower rates of CNS metastasis in pancreatic cancer. RTK–RAS pathway alterations in prostate cancer were associated with higher rates of bone but lower rates of liver metastasis (Extended Data Fig. 7a). We also investigated the extent to which chromosome arm-level amplifications or deletions were associated with metastasis to future organ sites. In this analysis, arm-level amplifications and deletions were generally associated with multiple sites of metastasis across cancer types with some notable exceptions (Extended Data Fig. 7b). CRC with MSS was generally unaffected by these changes except for 1p and 1q amplifications and 3p, 11p, 11q and 17p deletions, which were all prognostic for brain metastases. Prostate cancer arm-level changes mostly predisposed to brain and liver metastases. Pancreatic cancer arm-level changes seem to mostly predispose to liver metastases. In breast cancer, 16q and 16p deletions were associated with lower rates of CNS and lung metastases.

Overall, our analysis confirms several genomic–metastasis site associations observed in smaller[33,36] or non-temporal[12] cohorts but also identifies new potential genomic changes of prognostic importance that can be prospectively validated.

SETD2 and immunotherapy in LUAD

Of 5,957 patients with LUAD, 204 (3%) had *SETD2* driver mutations, and these emerged as predictors of longer OS (Extended Data Fig. 3)

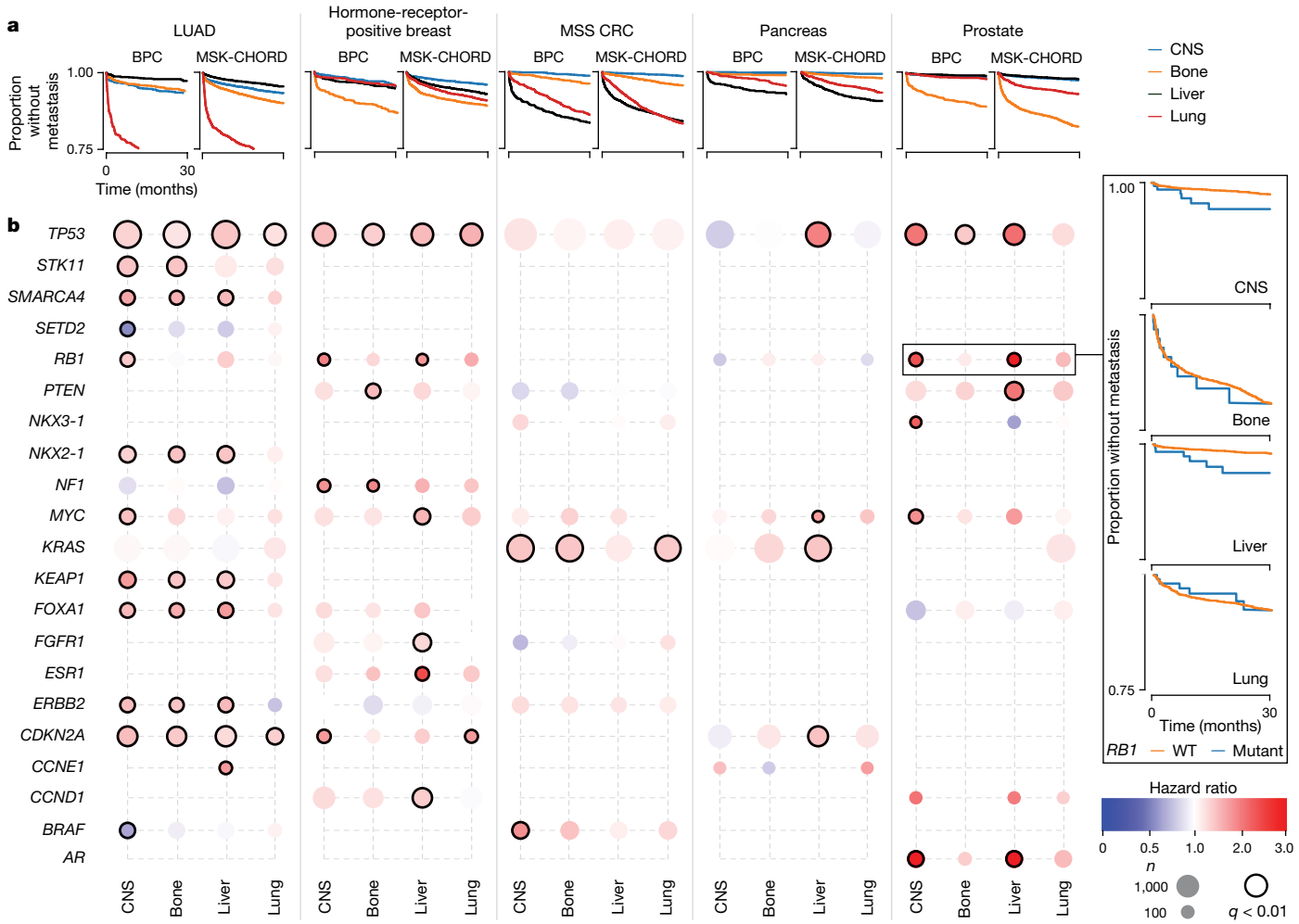


Fig. 4 | Analysis of time to metastatic site colonization. a, Time to metastatic site colonization among patients with LUAD, hormone-receptor-positive breast cancer, CRC with MSS, pancreatic adenocarcinoma (pancreas) and prostate cancer. Cohorts included the manually curated BPC and NLP-derived MSK-CHORD cohorts. **b**, Hazard ratios (colour), number of patients with alteration before site colonization (size) and statistical significance (Benjamini–Hochberg false discovery rate of 0.01, black outline) within MSK-CHORD. Analyses are adjusted for prior treatment, stage and histologic subtype. Only genes with at least one significant association in at least one cancer type (Benjamini–Hochberg $q < 0.01$) are shown. The inset depicts Kaplan–Meier curves of the cancer type and metastatic site highlighted in the grey rectangle stratified by *RB1* status. WT, wild type.

and lower rates of CNS metastasis (Fig. 4b). We sought to: corroborate these findings in independent cohorts; and further study why *SETD2* alterations might affect LUAD outcomes using MSK-CHORD. We identified two non-overlapping cohorts of patients with tumour genomic sequencing and longitudinal outcomes, one of which also included annotated metastatic sites of disease[1] (Methods). In both datasets, *SETD2* alteration was associated with better OS, and in the dataset in which CNS metastasis annotations were available, lower rates of CNS metastasis (Extended Data Fig. 8).

We studied whether *SETD2* driver alterations were associated with specific genomic alterations, histologic subtypes or other features. *SETD2* alterations were positively associated with *BRAF* and *ARID1A* alterations and negatively associated with *EGFR* and *MDM2* alterations and mucinous subtype but not otherwise associated with histologic subtype, PDL1 or smoking status (Fig. 5a,b and Extended Data Fig. 9). *SETD2* mutation was associated with a small but statistically significant difference in tumour mutational burden (TMB; Fig. 5c), consistent with previous observations[37].

We further examined whether *SETD2* mutation was associated with response to specific antineoplastics. *SETD2* mutation was associated with a longer time to next treatment or death following treatment with immune checkpoint blockade but not cytotoxic chemotherapy or molecularly targeted therapy (Fig. 5d). The association between *SETD2* mutation and longer immunotherapy response held among only patients with low TMB (<10 mutations per megabase) and in both validation cohorts (Fig. 5e). In summary, leveraging MSK-CHORD's size and rich annotations, we identified *SETD2* as an uncommon but promising biomarker of immunotherapy response in LUAD not explainable by other histopathologic, clinical or genomic features. We corroborated these findings in independent datasets.

Discussion

RWD may help scientists and clinicians better understand diseases such as cancer. Storage of RWD in free-text notes and siloed datasets has previously posed limitations to analysis. The Project GENIE BPC of the American Association for Cancer Research represents one large-scale effort to mine RWD by means of manual curation using an investment from a consortium of biopharmaceutical companies. In its initial years, the BPC has produced cancer cohorts with data from 2,004 patients with NSCLC[16] and 1,551 patients with CRC published until now. We have leveraged these along with recent advances in NLP, tissue genomic sequencing and health record digitization to create a richly annotated dataset including RWD from patients with multiple cancer types that

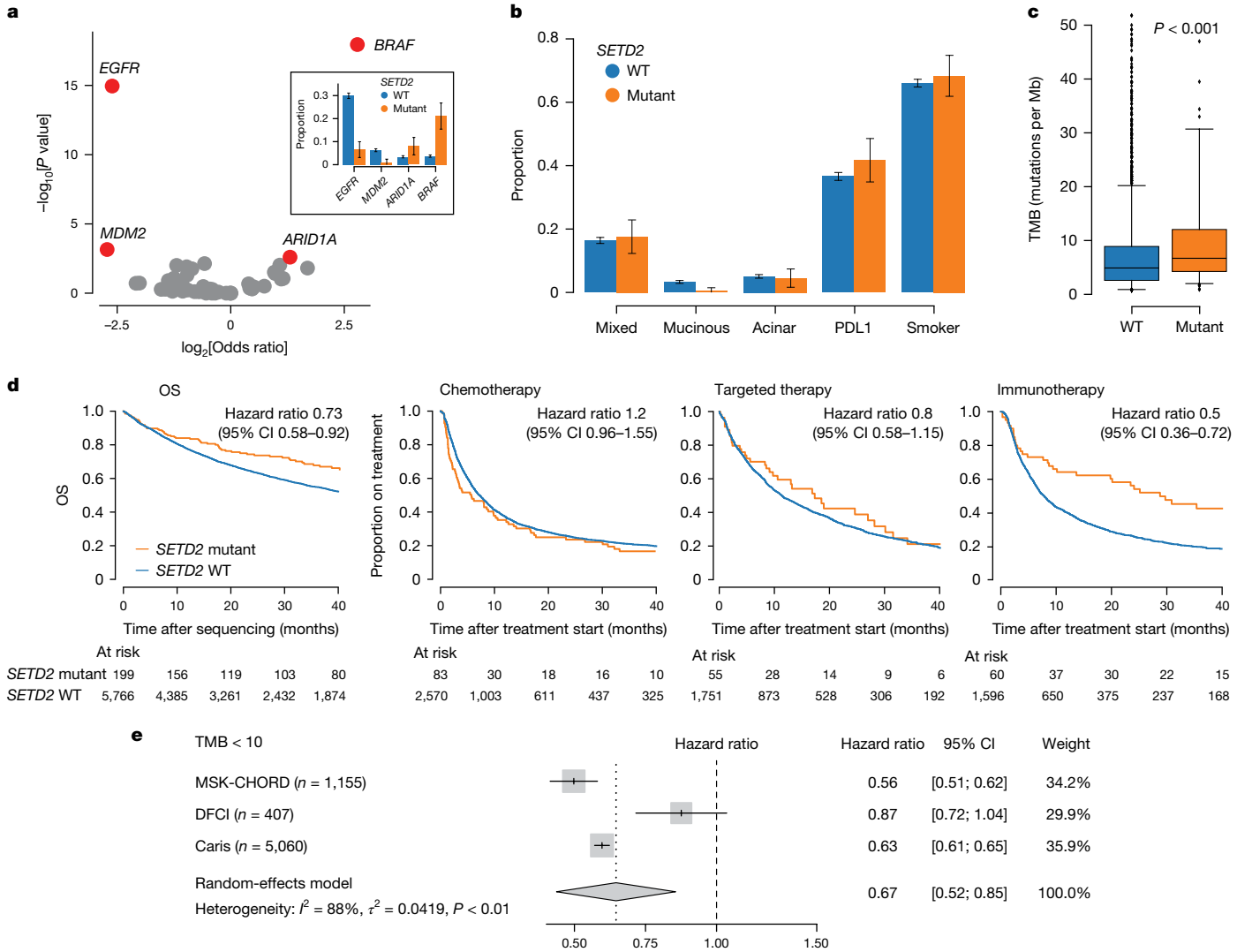


Fig. 5 | *SETD2* in LUAD. a, $\log_2$[Odds ratio] and *P* value (from two-sided Fisher's exact test) for associations of *SETD2* oncogenic alterations with other oncogenic gene alterations. Red indicates $q < 0.05$ by Benjamini–Hochberg. Inset: frequencies of associated genes ± binomial 95% CI. **b**, Proportion with features ± binomial 95% CI. *P* values by two-sided Fisher's exact test for mixed adenocarcinoma, mucinous, acinar, PDL1 and smoker variables were 0.63, 0.02, 0.87, 0.16 and 0.54, respectively. **c**, TMB with *P* value from two-sided Mann–Whitney *U* ($P = 2 \times 10^{-9}$). Box plots show medians and inner quartile ranges with ±95th percentile whiskers. For **a**–**c**, $n = 199$ *SETD2* mutant cases and $n = 5{,}766$ wild-type cases. **d**, OS from time of tumour sequencing and time to next treatment or death by treatment. Groups compared with Cox proportional hazards. **e**, Hazard ratios (mean ± 95% CI) for time to next treatment or death for patients with TMB < 10 mutations per megabase treated with immunotherapy based on *SETD2* status. Left dashed line, hazard ratio for all cohorts in meta-analysis. Right dashed line, hazard ratio of 1.0.

is many times larger than the original BPC, which enabled the findings shown here. We present this cohort as a community resource to aid in discovery of clinicogenomic relationships.

The NLP tools to extract data can be updated in real time, with minimal cost relative to that of manual curation. Validation experiments of NLP models in held-out cancer types suggest that our NLP tools generalize across solid tumours.

The prognostic models demonstrate the importance of rich annotation for predicting OS in most situations. For patients with stage I–III NSCLC, CRC and pancreatic cancer, models trained on single classes of data, including American Joint Committee on Cancer stage, which dictates many adjuvant therapy decisions[38], had worse performance than those trained on all classes of data. NLP-derived features, particularly tumour sites, emerged as important for predicting outcomes. Our results also highlight the challenges in predicting OS in diseases such as pancreatic cancer. Future studies will explore whether other modalities, such as liquid biopsy and laboratory data[39], further improve outcome prediction.

We explored the utility of neural networks applied to notes to predict outcomes, but neural networks may also be applied directly to images[13]. Further work comparing interpretable 'late fusion' models such as those explored here to 'early fusion' models trained jointly on higher-dimensional data such as images would elucidate the extent to which the raw data contain prognostic information not encapsulated by the features present here.

Our study has limitations. Although we attempted to circumvent immortality bias using left truncation and controlling for progression and tumour sites at the start date of cohort entry, cohort entry (that is, genomic sequencing) is not random and disproportionately represents patients with advanced disease or recent progression, which may affect the generalizability of our prognostic models[40]. As with any real-world dataset, there are potential confounders not directly reported here. Comorbidities and symptoms are two patient features often crucial to clinical decision-making that have an impact on outcomes but are not captured in either the BPC or MSK-CHORD. Future iterations of MSK-CHORD will include these and other data elements. Our cohort

consists predominantly of patients from a catchment based on New York and New Jersey. Although the size of our patient base with tumour sequencing has previously enabled findings in populations of diverse backgrounds[41,42], careful future work involving multiple centres and a more diverse patient population is required to disentangle socioeconomic, demographic and geographic effects on outcome models presented here. Whole-genome and RNA sequencing, as well as single-cell studies incorporating the tumour microenvironment[43], are necessary to derive mechanistic insights into the process of metastatic colonization.

NLP combined with results from tissue genomic sequencing, tumour registry and other siloed data sources can empower RWD analysis. Our results highlight the importance of multiple data streams in predicting outcomes. It is our hope that MSK-CHORD will fuel further research into real-world genotype–phenotype relationships in cancer.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-08167-5.

1. Kehl, K. L. et al. Artificial intelligence-aided clinical annotation of a large multi-cancer genomic dataset. *Nat. Commun.* **12**, 7304 (2021).
2. Fries, J. A. et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nat. Commun.* **12**, 2017 (2021).
3. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 6000–6010 (2017)
4. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* https://doi.org/10.1038/s41586-023-06160-y (2023).
5. Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. A comparative study of pretrained language models for long clinical text. *J. Am. Med. Inform. Assoc.* **30**, 340–347 (2023).
6. Haupt, C. E. & Marks, M. AI-generated medical advice—GPT and beyond. *JAMA* **329**, 1349–1350 (2023).
7. Molina, M. A. et al. Trastuzumab (herceptin), a humanized anti-Her2 receptor monoclonal antibody, inhibits basal and activated Her2 ectodomain cleavage in breast cancer cells. *Cancer Res.* **61**, 4744–4749 (2001).
8. Kris, M. G. et al. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. *JAMA* https://doi.org/10.1001/jama.2014.3741 (2014).
9. Singal, G. et al. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non–small cell lung cancer using a clinicogenomic database. *JAMA* **321**, 1391–1399 (2019).
10. Mahon, P., Hall, G., Dekker, A., Vehreschild, J. & Tonon, G. Harnessing oncology real-world data with AI. *Nat. Cancer* https://doi.org/10.1038/s43018-023-00689-7 (2023).
11. Liu, R. et al. Systematic pan-cancer analysis of mutation–treatment interactions using large real-world clinicogenomics data. *Nat. Med.* **28**, 1656–1661 (2022).
12. Nguyen, B. et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell* **185**, 563–575 (2022).
13. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126 (2022).
14. Vanguri, R. S. et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* https://doi.org/10.1038/s43018-022-00416-8 (2022).
15. Chowell, D. et al. Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nat. Biotechnol.* **40**, 499–506 (2022).
16. Choudhury, N. J. et al. The GENIE BPC NSCLC cohort: a real-world repository integrating standardized clinical and genomic data for 1,846 patients with non-small cell lung cancer. *Clin. Cancer Res.* https://doi.org/10.1158/1078-0432.CCR-23-0580 (2023).
17. Lavery, J. A. et al. A scalable quality assurance process for curating oncology electronic health records: the Project GENIE Biopharma Collaborative approach. *JCO Clin. Cancer Inform.* https://doi.org/10.1200/CCI.21.00105 (2022).
18. Keegan, N. M. et al. Clinical annotations for prostate cancer research: defining data elements, creating a reproducible analytical pipeline, and assessing data quality. *Prostate* **82**, 1107–1116 (2022).
19. Chatila, W. K. et al. Genomic and transcriptomic determinants of response to neoadjuvant therapy in rectal cancer. *Nat. Med.* **28**, 1646–1655 (2022).
20. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* https://doi.org/10.1038/nm.4333 (2017).
21. de Bruijn, I. et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).
22. Razavi, P. et al. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* **34**, 427–438 (2018).
23. Piper-Vallillo, A. J., Sequist, L. V. & Piotrowska, Z. Emerging treatment paradigms for EGFR-mutant lung cancers progressing on osimertinib: a review. *J. Clin. Oncol.* https://doi.org/10.1200/JCO.19.03123 (2020).
24. Abida, W. et al. Genomic correlates of clinical outcome in advanced prostate cancer. *Proc. Natl Acad. Sci. USA* **116**, 11428–11436 (2019).
25. Bolton, K. L. et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nat. Genet.* https://doi.org/10.1038/s41588-020-00710-0 (2020).
26. Isaacsson Velho, P. et al. Molecular characterization and clinical outcomes of primary Gleason pattern 5 prostate cancer after radical prostatectomy. *JCO Precis. Oncol.* https://doi.org/10.1200/PO.19.00081 (2019).
27. André, T. et al. Pembrolizumab in microsatellite-instability–high advanced colorectal cancer. *N. Engl. J. Med.* **383**, 2207–2218 (2020).
28. Hechtman, J. F. et al. Retained mismatch repair protein expression occurs in approximately 6% of microsatellite instability-high cancers and is associated with missense mutations in mismatch repair genes. *Mod. Pathol.* **33**, 871–879 (2020).
29. Selenica, P. et al. APOBEC mutagenesis, kataegis, chromothripsis in EGFR-mutant osimertinib-resistant lung adenocarcinomas. *Ann. Oncol.* **33**, 1284–1295 (2022).
30. Dogan, S. et al. Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin. Cancer Res.* **18**, 6169–6177 (2012).
31. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018).
32. Wu, J. et al. Predictive model to guide brain magnetic resonance imaging surveillance in patients with metastatic lung cancer: impact on real-world outcomes. *JCO Precis. Oncol.* https://doi.org/10.1200/PO.22.00220 (2022).
33. Lengel, H. B. et al. Genomic mapping of metastatic organotropism in lung adenocarcinoma. *Cancer Cell* **41**, 970–985 (2023).
34. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* https://doi.org/10.1200/po.17.00011 (2017).
35. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
36. Skakodub, A. et al. Genomic analysis and clinical correlations of non-small cell lung cancer brain metastasis. *Nat. Commun.* **14**, 4980 (2023).
37. Lu, M. et al. Pan-cancer analysis of *SETD2* mutation and its association with the efficacy of immunotherapy. *npj Precis. Oncol.* **5**, 51 (2021).
38. Pignon, J.-P. et al. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE Collaborative Group. *J. Clin. Oncol.* **26**, 3552–3559 (2008).
39. Jee, J. et al. Overall survival with circulating tumor DNA-guided therapy in advanced non-small-cell lung cancer. *Nat. Med.* **28**, 2353–2363 (2022).
40. Kehl, K. L. et al. Clinical inflection point detection on the basis of EHR data to identify clinical trial–ready patients with cancer. *JCO Clin. Cancer Inform.* https://doi.org/10.1200/CCI.20.00184 (2021).
41. Jiagge, E. et al. Tumor sequencing of African ancestry reveals differences in clinically relevant alterations across common cancers. *Cancer Cell* https://doi.org/10.1016/j.ccell.2023.10.003 (2023).
42. Arora, K. et al. Genetic ancestry correlates with somatic differences in a real-world clinical cancer sequencing cohort. *Cancer Discov.* **12**, 2552–2565 (2022).
43. Vázquez-García, I. et al. Ovarian cancer mutational processes drive site-specific immune evasion. *Nature* **612**, 778–786 (2022).

MSK Cancer Data Science Initiative Group

Aaron Lisman[1], Benjamin Gross[1], Brooke Mastrogiacomo[1], Gaofei Zhao[1], Ino de Bruijn[1], Ritika Kundra[1], Thinh Ngoc Tran[1,5], Walid K. Chatila[1], Xiang Li[1], Aarman Kohli[1], Darin Moore[1], Raymond Lim[1], Tom Pollard[1], Arfath Pasha[1], Robert Sheridan[1], Avery Wang[1], Calla Chennault[1], Manda Wilson[1], Hongxin Zhang[1], Robert Pimienta[1], Surya Rangavajhala[1], Guru Subramanian[1], Jowel Garcia[1], Naveen Rachuri[1], Kevin Boehm[1], Mitchell Parker[1], Henry Walch[1], Subhiksha Nandakumar[1], Jordan Eichholz[1], Ayush Kris[1], Paolo Manca[1], Xuechun Bai[1], Tejiri Agbamu[1], Justin U[1] & Xinran Bi[1]

Methods

Patients

This study primarily analysed data for patients with tumour genomic sequencing and completed tumour registry entries from two partially overlapping sources: patients with MSK-IMPACT sequencing (forming the basis of MSK-CHORD) and the Project GENIE BPC cohort of the American Association for Cancer Research, which includes patients with tumour genomic profiling and clinical annotation from four institutions including MSK. Details regarding the BPC have been published previously[7]. Here we included patients in the BPC with single-primary NSCLC, breast, colorectal, prostate or pancreatic cancer. The MSK-CHORD cohort comprises patients at MSK, an academic cancer hospital with tumour genomic sequencing using MSK-IMPACT, a Food and Drug Administration-authorized tumour genomic profiling assay, which uses matched white blood cell sequencing to filter clonal haematopoiesis and germline variants. All MSK patients were enrolled as part of a prospective sequencing protocol (NCT01775072) or analysed as part of institutional review board (IRB)-approved retrospective research protocols (MSK IRB protocols 16–1463 and 19–368). The study was independently approved by the IRBs of MSK and DFCI. Patients provided written, informed consent and were enrolled in a continuous, non-random fashion. Data here are from a 9 September 2023 snapshot.

Genomic annotation

For all analyses involving tumour genomic alterations aggregated at the gene level, a Food and Drug Administration-recognized molecular knowledge database (OncoKB[34]) was used to annotate all mutations, copy number changes and structural variations as oncogenic or not; any such oncogenic alteration led to a gene being labelled positive for the purposes of analysis. For OS modelling in which non-MSK BPC patients were used as an external validation cohort, only genes present in all sequencing panels across the BPC were used as variables. For other genomic analyses, the 341 genes included in the first MSK-IMPACT sequencing panel[20] were used as variables. The presence or absence of genomic gains and losses of each chromosome arm were identified from MSK-IMPACT data. Genomic coordinates for the chromosome arms in the GRCh37 (also known as hg19) human genome assembly were considered gained or lost if most of the arm (>50%) is made up of segments with an absolute value log ratio of ≥0.2 (ref. 44).

NLP models

Radiology reports. *Data preprocessing*. Radiology reports for computed tomography (CT), positron emission tomography and magnetic resonance imaging examinations of chest, abdomen, pelvis, head and/or extremities were queried for all patients within the MSK-CHORD cohort. Report sections were segmented using regular expression to separate the 'impression' section from the full report, for cases in which it was available. The pieces of impression text corresponding to the manually curated MSK-BPC labels for presence of cancer, tumour sites and cancer progression were merged to create a direct mapping of label and text that the labels were derived from.

Radiographic progression. We fine-tuned a RoBERTa transformer model[45] on impression sections extracted from radiology reports paired with binarized human-curated progression labels. Labels were binarized by calling the two GENIE BPC label classes 'Progressing/Worsening/Enlarging' and 'Mixed' as positive, and calling other classes ('Improving/Responding', 'Stable/No change' and 'Not stated/Indeterminate') as negative.

Binarized supervision labels were provided at the document level (that is, the model was trained to predict a single binary variable for a given impression section). We used the PyTorch[46] implementation of RoBERTa and pretrained model weights from the HuggingFace library and model hub[47]. Text was tokenized with the default RoBERTa tokenizer and report-level predictions are pooled using the default method of conditioning on the first [CLS] pseudo-token prepended to the sequence comprising the impression section. We used a batch size of 128, fine-tuning with the AdamW optimizer[48] using a learning rate of 2×10^{-6}, and fine-tuning for 20 epochs with a linearly decaying learning-rate scheduler with a 2-epoch warm-up period. Hyperparameter values were selected through a random search using a holdout validation set of 20% of reports across learning rate values $\{1 \times 10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-6}\}$, batch size values {8, 16, 32, 64, 128, 256} and num-epochs {5, 10, 20, 50}. Extrinsic results (that is, main results incorporating model predictions) were presented on models trained on the full MSK-BPC cohort.

Tumour sites. We fine-tuned a ClinicalBERT model[49], which is itself a BioBERT model[50] fine-tuned on reports from the MIMIC-III v1.4 database[51]. We extracted impression sections from radiology reports and paired them with report-level supervision from the GENIE BPC dataset. Labels were transformed into ten binary variables corresponding to a closed inventory of nine common disease sites (adrenal gland, bone, CNS or brain, intra-abdominal, liver, lung, lymph nodes, pleura and reproductive organs), along with one 'other' variable, describing whether the report is labelled as indicating tumour presence in that organ site.

The model was trained in a multi-labelling setup: pooled transformer output was input to a single-layer fully connected feed-forward network of width d with a tanh nonlinearity, whose output is linearly transformed to a ten-dimensional vector giving ten logits, from which binary cross-entropy losses were computed against the gold-standard labels and mean-pooled. In other words, the network computes

$$f(x) = \sigma(V(\tanh(W\varphi(x))))$$

in which x is the tokenized document, $\varphi(x)$ is the pooled transformer output vector, W is a learned affine transformation outputting a d-dimensional vector, tanh is applied element-wise, V is a learned affine transformation mapping d-dimensional vectors to ten-dimensional vectors, and σ is a plain element-wise sigmoid function; $f(x)$ is a ten-dimensional vector of values between 0 and 1. Note that the different per-site predictions are non-mutually-exclusive and are conditionally independent given the post-pool d-dimensional hidden state.

The ClinicalBERT model was implemented in PyTorch[46]; we used the model and pretrained model weights in the HuggingFace library and model hub[47]. We pooled transformer model output using the default method of conditioning on the first [CLS] pseudo-token prepended to the sequence comprising the impression section. We trained using AdamW[48] trained using a batch size of 8, a learning rate of 2×10^{-6}, a dropout probability of 0.2 (applied to the post-pool single-hidden-layer feed-forward network) and a pre-logit hidden dimension of 1,024, training for 15 epochs with a warm-up period of 1.5 epochs. Extrinsic results (that is, main results incorporating model predictions) were presented on models trained on the full MSK-BPC cohort.

Cancer presence. We fine-tuned a BERT[52] base model (uncased)[53] on impression sections extracted from radiology reports paired with binarized human-curated cancer evidence labels. Labels were binarized by calling the MSK-BPC label class 1 as 'yes' for presence of cancer and calling label class 0 as 'no' for absence of cancer. Binarized supervision labels were provided at the document level (that is, the model was trained to predict a single binary variable for a given impression section). BERT models were trained as described for tumour sites. Text was tokenized with the default HuggingFace AutoTokenizer for BERT, and report-level predictions were pooled using the default method of conditioning on the first [CLS] pseudo-token prepended to the sequence comprising the impression section. We used a batch size of 32 and fine-tuned for a maximum of 10 epochs. We trained the models using the AdamW optimizer[48] using a learning rate of 1×10^{-5}, epsilon of

1×10^{-8}, weight decay of 1×10^{-4} and no warm-up period. During training, model weights were optimized to minimize cross-entropy loss.

Clinician notes. *Data preprocessing*. Clinician notes for patients were queried and filtered by initial consult (IC) and follow-up notes created by medical oncologists, radiation oncologists, surgery, inpatient services and others. Notes in the institutional database are segmented into subsections including family history, present illness, comorbidities and so on. Further filtering or combining of note subsections was dependent on the application. For inferring prior outside medications, IC notes were filtered and included sections relevant to external treatments, such as past medical history, history of present illness and chief complaint, while excluding sections mentioning future treatment plans. Patients with no IC notes in the allowable note categories were excluded from the training and validation set. We excluded patients with IC notes dated more than 90 days after their initial visit date. We selected one note per patient to analyse. If a patient had multiple notes, the IC note with the earliest creation time was used. Preprocessing for inference of HER2 and hormone receptor consisted of filtering notes created by the breast medicine division, for which entire IC and follow-up notes were used as the input to the model.

Prior external treatment. The other transformer-based models presented above operate on impression sections that are generally very short and therefore do not see appreciably degraded marginal performance from truncating documents to the maximum model input sequence size of 512 subtokens. This relatively short input limit is necessary for the full-self-attention parameterizations used by these models, which require memory scaling quadratically in input sequence length. However, full IC reports are appreciably longer than impression sections, and any mention of prior anti-neoplastic treatments occurs within a much longer textual context. We therefore use a transformer model engineered to have subquadratic memory requirements; in particular, we fine-tune a Clinical-Longformer model[54], which is itself a Longformer model[55] fine-tuned on the MIMIC-III v1.4 database[51]. This model has a maximum input sequence length of 4,096 subtokens.

The Clinical-Longformer model is implemented in PyTorch[46]; we use the model and pretrained model weights in the HuggingFace library and model hub[47]. We pool transformer model output using the default method of conditioning on the first [CLS] pseudo-token prepended to the sequence comprising the impression section. We train AdamW[48] using a batch size of 64 and a learning rate of 1×10^{-6}, training for 20 epochs with a warm-up period of 2 epochs. We upsample minority-class examples uniformly with replacement to achieve class balance during training. Extrinsic results (that is, main results incorporating model predictions) are presented on models trained on the MSK-BPC cohort.

HER2 and hormone receptor. As HER2 and hormone receptor can be heterogeneous across pathology samples, we sought to create a classifier based on clinician notes to identify the overall receptor subtypes for a patient's cancer used to inform treatment. For training, we used clinician notes from a cohort of 6,053 patients with single-primary breast cancer with manually annotated HER2 and hormone receptor subtypes to train separate HER2 and hormone receptor binary classifiers. We performed training and testing within this cohort with a 90/10 split. Specifically, the clinician note chronologically closest to sequencing was used as features and the expert-annotated subtypes as targets. For final validation, we used a held-out set of 1,489 patients from a previously published breast cancer dataset[22]. As with the prior treatment model, we used Clinical-Longformer models for both HER2 and hormone receptor classifiers using a 2,000-subtoken input, padded as necessary. We used the AdamW optimizer with a batch size of 64 and a learning rate of 1×10^{-6}, training for 30 epochs without a warm-up period.

Smoking status. Smoking status (former or current versus never) was obtained from dedicated smoking or social history sections through regular expression extraction applied to the first available clinician assessment for a given patient. The algorithm was created on the basis of a previously published cohort of 247 patients with NSCLC and previously annotated smoking status[14], withholding data from patients also present in the MSK-BPC NSCLC cohort. The model was validated on the basis of the MSK BPC NSCLC cohort.

Pathology reports. *PDL1*. PDL1 status (positive defined as 1% or higher versus negative) was obtained through regular expression extraction applied to the first available clinician assessment for a given patient. The algorithm was created on the basis of a previously published cohort of 247 patients with NSCLC and previously annotated smoking status[14], withholding data from patients also present in the MSK BPC NSCLC cohort. The model was validated on the basis of the MSK BPC NSCLC cohort.

Gleason score. Gleason score (6–10) was obtained through regular expression extraction applied to pathology reports from either prostatic biopsies or resections. The algorithm was created on the basis of iterative fine-tuning on a previously published cohort of 451 patients with prostate cancer and previously annotated Gleason score[56], withholding data from patients also present in the MSK BPC Prostate cohort. The model was validated on the basis of the MSK BPC Prostate cohort.

MMR. Mismatch status (proficient versus deficient) was obtained through regular expression extraction applied to histopathology reports. The algorithm was created on the basis of a previously published cohort of 224 patients with CRC and previously annotated MMR status[57], withholding data from patients also present in the MSK CRC cohort. The model was validated on the basis of the MSK BPC CRC cohort.

Billing code annotation metrics. We sought to assess the accuracy of structured data elements (that is, billing codes[12]) to recover tumour site information and to compare this accuracy with that of our NLP algorithms. As the timing of billing codes is not necessarily tied to particular radiology reports, we aggregated labels at the patient level, wherein cancer detection in a given tumour site at any point in the patient's history was considered positive overall for that site. Patient-level billing code labels and, separately, NLP labels (from radiology impressions as above) were compared to gold-standard curated BPC labels, all aggregated at the patient level. The patient-level accuracies for these annotations are provided in Supplementary Table 2.

OS modelling. RSFs[58] to predict time to death from time of cohort entry, right-censored at time of last follow-up, were trained using pre-assigned hyperparameters (*n* trees = 1,000, minimum *n* splits = 10, minimum *n* samples per leaf = 15). In exploratory secondary analyses, a random hyperparameter grid search to find 'optimal' hyperparameters using a 20% holdout for evaluation was conducted (*n* tree range 200–2,000, minimum *n* splits range 5–20, minimum *n* samples per leaf range 5–30, *n* search iterations = 100, threefold internal cross-validation for hyperparameter selection); a model trained on optimal hyperparameters did not yield better results (*c*-index 'improvement' of −0.01 using optimal versus pre-assigned hyperparameters). We included all variables in Supplementary Table 6, grouped according to the schema in that table.

To predict time to death while accounting for left truncation and right censoring, we used the OncoCast package (https://github.com/AxelitoMartin/OncoCast) updated from previous work[59,60] with the RF (Random Forest) method. In brief, this method fits an elastic net-regularized Cox proportional hazards model to the data, and then applies a random forest to estimate the Martingale residuals; this correction term is applied when the model is tested on new data. We created an ensemble learning model through cross-validation or by training on the whole MSK-CHORD dataset and validating the model on the non-MSK BPC dataset as for the RSF model. The OncoCast model, configured with 500 trees, 5 terminal nodes and 50 runs, was fitted to

the training set. Predictions of risk for the test set were made across all iterations. Model performance was assessed using the concordance probability index at each iteration.

radLongformer. We fine-tuned a Clinical-Longformer[54] model to take as input the full text of CT chest, abdomen and pelvis (CAP) reports and predict binarized OS within 6 months, a clinically meaningful endpoint and a time frame in which a single radiology report might meaningfully prognosticate. We split all cohorts into training and test sets at the patient level, reserving 20% of the cohort or all patients with a CT CAP within 3 months of cohort entry for testing, whichever was smaller. In the training set, all CT CAP reports from all patients were annotated according to survival status within 6 months; those with insufficient follow-up were excluded. The Clinical-Longformer was fine-tuned in this dataset using a batch size of 64 and a learning rate of 1×10^{-6}, training for 20 epochs with a warm-up period of two epochs.

Time to metastasis. The association of genomic alterations with time to metastasis was analysed using Cox proportional hazards models. Death was treated as a censoring event. Patients with metastasis to a given site of interest before the start time (time of sample acquisition; that is, the earliest time a given alteration could be confirmed for a given tumour) were excluded from analysis. Prior treatment (any versus none) and stage (I–III versus IV) were included as variables in all multivariable analyses. Histologic subtype was included as a variable where indicated.

***SETD2* validation cohorts.** We utilized two validation cohorts of patients with LUAD and tumour genomic profiling: patients at DFCI; and patients in a commercial real-world dataset. Details of the DFCI cohort have been published previously[1,61]. In the commercial dataset, formalin-fixed paraffin-embedded samples from patients with NSCLC were submitted to a commercial Clinical Laboratory Improvement Amendments-certified laboratory for molecular profiling (Caris Life Sciences, Phoenix, AZ). Any patient with Caris tumour molecular profiling was eligible for inclusion; patient sources include a variety of community and academic settings, and patients were non-overlapping with those in MSK-CHORD. A total of 29,422 NSCLCs with adenocarcinoma histology were analysed by next-generation sequencing, 592 targeted panel or whole-exome sequencing for genomic features. Before molecular testing, tumour enrichment was achieved by collecting targeted tissue using manual microdissection techniques. For NextSeq-sequenced tumours, a custom-designed SureSelect XT assay was used to enrich 592 whole-gene targets (Agilent Technologies, Santa Clara, CA). For NovaSeq whole-exome-sequenced tumours, a hybrid pull-down panel of baits designed to enrich for more than 700 clinically relevant genes at high coverage and high read depth was used, along with another panel designed to enrich for an additional >20,000 genes at a lower depth. A 500-megabase single-nucleotide polymorphism backbone panel (Agilent Technologies, Santa Clara, CA) was added to assist with gene amplification and deletion measurements and other analyses. All variants were detected with >99% confidence, with an average sequencing depth of coverage of >500 and an analytic sensitivity of 5%. This test has a sensitivity to detect as low as approximately 10% population of cells containing a mutation in all exons from the high-read-depth clinical genes and 99% of all exons in the 20,000 whole-exome regions. Genetic variants identified were interpreted by board-certified molecular geneticists and categorized according to the American College of Medical Genetics and Genomics standards. Real-world OS was obtained from insurance claims data and calculated from time of biopsy to last contact.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

MSK-CHORD is available under the Creative Commons BY-NC-ND 4.0 licence and can be accessed on cBioPortal (https://www.cbioportal.org/study/summary?id=msk_chord_2024). Data from the GENIE BPC can be accessed on cBioPortal (https://genie.cbioportal.org/), with additional data available as previously described[16] and further instructions here: https://www.aacr.org/professionals/research/aacr-project-genie/aacr-project-genie-data/. For questions regarding access to Caris validation data contact jxiu@carisls.com.

Code availability

All code to perform the analyses presented here is available at GitHub (https://github.com/clinical-data-mining).

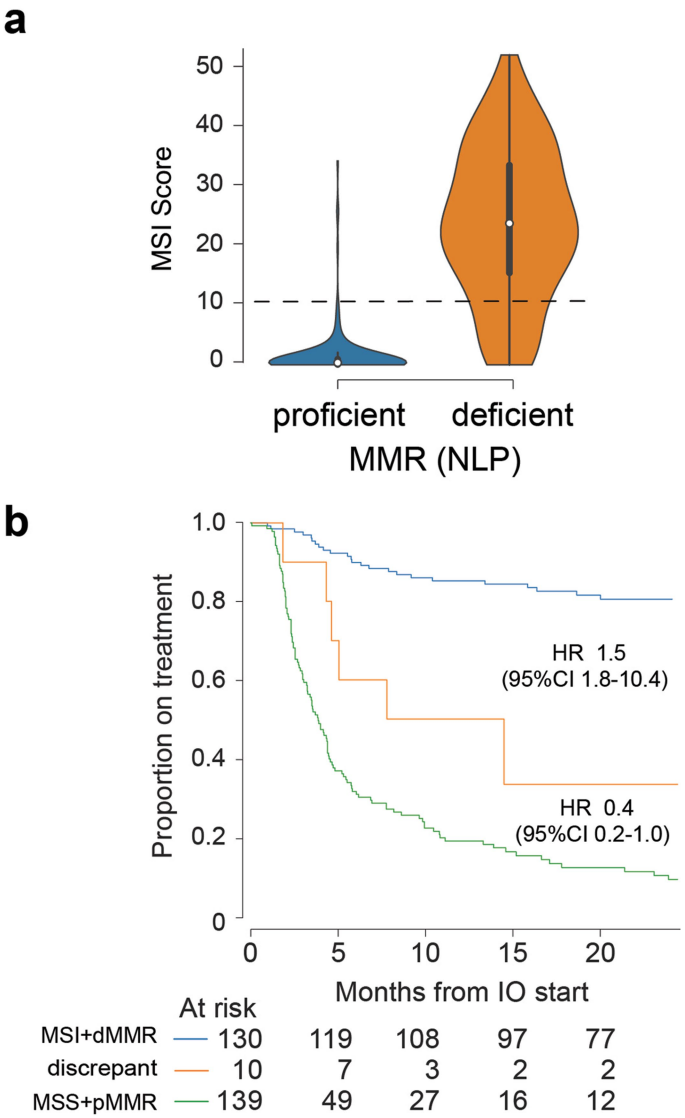
44. Penson, A. et al. Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA Oncol.* **6**, 84–91 (2020).
45. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).
46. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32,** 8024–8035 (2019)
47. Wolf, T. et al. Transformers: State-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, 2020).
48. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. in *7th International Conference on Learning Representations* (ICLR, 2019).
49. Alsentzer, E. et al. Publicly available clinical BERT embeddings. in *Proceedings of the 2nd Clinical Natural Language Processing Workshop* 72–78 (Association for Computational Linguistics, 2019).
50. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
51. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
52. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long and Short Papers) 4171–4186 (Association for Computational Linguistics, 2019).
53. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long and Short Papers) 4171–4186 (Association for Computational Linguistics, 2019).
54. Li, Y, Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. A comparative study of pretrained language models for long clinical text. *J. Am. Med. Inform. Assoc.* **30**, 340–347 (2023).
55. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: the long-document transformer. Preprint at https://arxiv.org/abs/2004.05150 (2020).
56. Abida, W. et al. Prospective genomic profiling of prostate cancer across disease states reveals germline and somatic alterations that may affect clinical decision making. *JCO Precis. Oncol.* **1**, 1–16 (2017).
57. Stadler, Z. K. et al. Reliable detection of mismatch repair deficiency in colorectal cancers using mutational load in next-generation sequencing panels. *J. Clin. Oncol.* **34**, 2141–2147 (2016).
58. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
59. Zauderer, M. G. et al. The use of a next-generation sequencing-derived machine-learning risk-prediction model (OncoCast-MPM) for malignant pleural mesothelioma: a retrospective study. *Lancet Digit. Health* **3**, e565–e576 (2021).
60. Shen, R. et al. Harnessing clinical sequencing data for survival stratification of patients with metastatic lung adenocarcinomas. *JCO Precis. Oncol.* https://doi.org/10.1200/po.18.00307 (2019).
61. Aprati, T. et al. Abstract 2750: Leveraging machine-learning approaches to dissect drivers of clinical metastatic dynamics in lung adenocarcinoma. *Cancer Res.* **84**, 2750 (2024).
62. Middha, S. et al. Reliable pan-cancer microsatellite instability assessment by using targeted next-generation sequencing data. *JCO Precis. Oncol.* **1**, 1–17 (2017).

Acknowledgements This work was supported by the MSK Support Grant/Core Grant (P30 CA008748), the MSK Molecular Diagnostics Service in the Department of Pathology, the Marie-Josee and Henry R. Kravis Center for Molecular Oncology, the Halvorsen Center for Computational Oncology, P50-CA092629, P01-CA228696, R01 CA234361, the Investigational Cancer Therapeutics Training Program (T32-CA009207 to J.J.), the Paul Calabresi Career Development Award for Clinical Oncology (K12 CA184746 to J.J.), the American Society of Clinical Oncology Jill Soffer Young Investigator Award (J.J.), a Lung Cancer Research Foundation-AstraZeneca Grant (J.J.), the National Cancer Institute (R01CA217169 and R01CA240472 to D.R.J. and K08CA286842 to J.J.), the Movember-Prostate Cancer Challenge Award and the Hamilton Family Foundation (to D.R.J.). The authors would like to acknowledge the American Association for Cancer Research (AACR) and its financial and material support in the development of the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing. Interpretations are the responsibility of study authors.

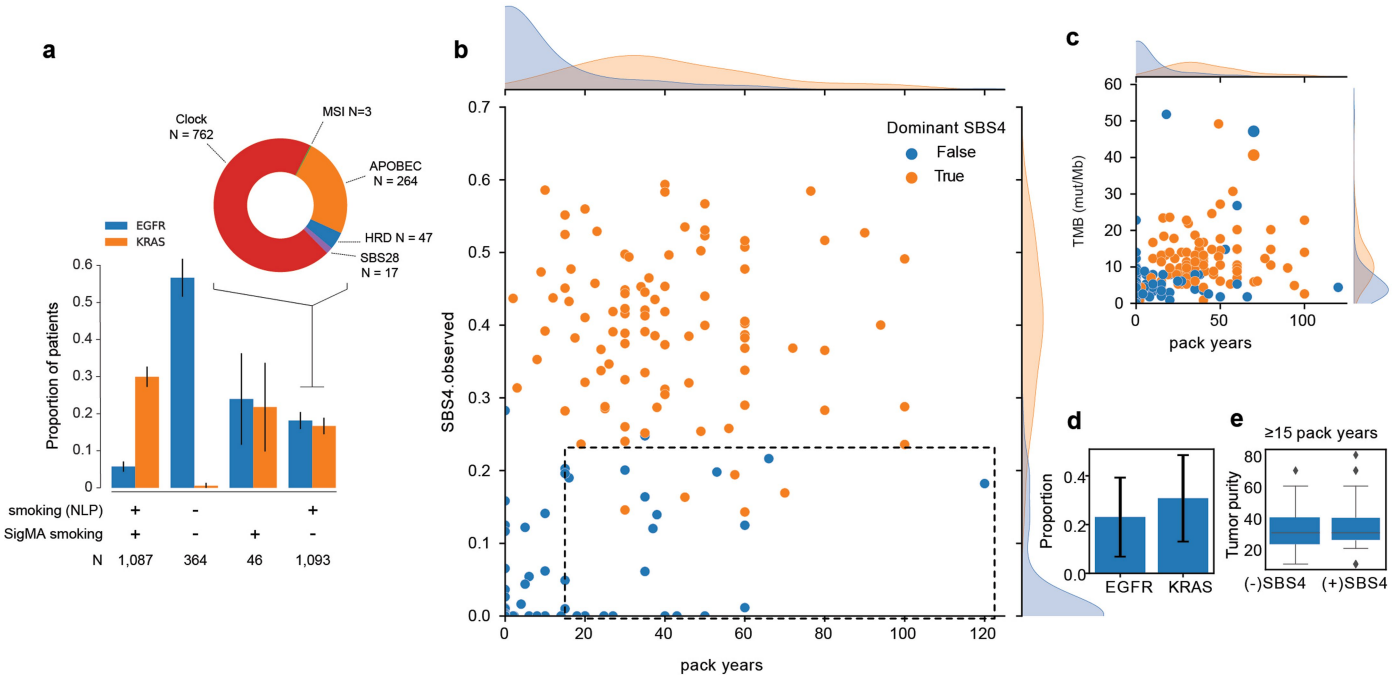
Author contributions Data collection and analysis: J.J., C. Fong, K. Pichotta, T.N.T., A. Luthra, M. Waters, C. Fu, M. Altoe, S.-Y.L., S.B.M., M. Ahmed, S.K., M. Pirun, W.K.C., I.d.B., A.P., R.K., B.G., B.M., T.J.A., D.L., J. Gao, M.C., K. Pekala, L.L., M. Perry, C.B., M.D., B.A.S., A.R.B., J.C., L.B., A. Li, A. Safonov, A. Stonestrom, P.S.-V. and K.L.K. Statistical design: A.M., R. Shen and Y.C. Wrote the first draft of the manuscript: J.J., C.W. and N. Schultz. Supervised the work: M.R., H.S., M.L., J.S.R.-F., D.B.S., D.R.J., D.G., H.Y., D.C., R.Y., W.A., W.P., E.M.O., J.G.-A., N. Socci, F.S.-V., J.C.-Z., P.D.S., R. Levine, C.M.R., M.F.B., S.P.S., D.S., P.R., K.L.K., B.T.L., G.J.R. and N. Schultz. All main authors provided input to the final version of the manuscript. Contributions for the consortium authors are as follows. Data visualization: A. Lisman, B.G., B.M., G.Z., I.d.B., R.K., T.N.T., W.K.C. and X.L. Engineering core support: A. Kohli, D.M., R. Lim, T.P., A.P., R. Sheridan, A.W., C.C., M. Wilson and H.Z. Engineering open systems support: R.P., S.R., G.S., J. Garcia and N.R. Translational data use: K.B., M. Parker, H.W., S.N., J.E., A. Kohli, A. Kris and P.M. Additional data review: X. Bai, T.A., J.U. and X. Bi.

Competing interests S.B.M. declares professional services and activities for Amgen, Clinical Care Options, Daiichi Sankyo, Elevation Oncology, MedPage Today, Novartis, Physicians' Education Resource, Pinetree Therapeutics, Purple Biotech and Vindico Medical Education; and equity in McKesson. L.B. declares professional services and activities for the Cancer Prevention & Research Institute of Texas. M.R. declares professional services and activities (uncompensated) for Artios Pharma, AstraZeneca, Foundation Medicine, Pfizer and Tempus Labs; and professional services and activities for Change Healthcare, Clinical Education Alliance, Genome Quebec, MJH Associates and myMedEd. M.L. declares equity in and professional services and activities (uncompensated) for Paige.AI. D.B.S. declares professional services and activities for American Association for Cancer Research, BridgeBio, Fog Pharmaceuticals, Paige.AI, Pfizer, Rain Therapeutics; and equity in and professional services and activities for Elsie Biotechnologies, Fore Biotherapeutics, Function Oncology, Pyramid Biosciences and Scorpion Therapeutics. D.R.J. declares professional services and activities for AstraZeneca, Dava Oncology and MORE Health; and professional services and activities (uncompensated) for Merck & Co. D.G. declares professional services and activities for AstraZeneca, Grail, Johnson & Johnson, Med Learning Group, Medtronic and Varian Medical Systems. H.Y. declares professional services and activities for AbbVie, AstraZeneca, Black Diamond Therapeutics, Blueprint Medicines, C4 Therapeutics, Daiichi Sankyo, Ipsen Pharma, Janssen Pharmaceuticals, Taiho and Takeda Pharmaceuticals. R.Y. declares professional services and activities for Mirati Therapeutics and Zai Lab. W.A. declares professional services and activities for AstraZeneca, Clinical Education Alliance, Janssen Oncology and Touch Independent Medical Education. W.P. declares professional services and activities for Astellas. J.G.-A. declares professional services and activities for Ethicon; and equity in and professional services and activities for Intuitive Surgical. P.D.S. declares professional services and activities for the National Comprehensive Cancer Network and the National Institutes of Health. R. Levine declares equity, a fiduciary role or position and intellectual property rights in and professional services and activities (uncompensated) for Ajax Therapeutics; equity in Anovia Biosciences, Bakx Therapeutics, Epiphanes, Imago Biosciences and Syndax; professional services and activities for AstraZeneca, Genome Quebec, Goldman Sachs, Incyte, Janssen Pharmaceuticals and Jubilant Therapeutics; equity in and professional services and activities (uncompensated) for Auron Therapeutics and the Isoplexis Corporation; equity in and professional services and activities for C4 Therapeutics, Kurome Therapeutics, Mana Therapeutics, Mission Bio, Prelude Therapeutics, Scorpion Therapeutics, Zentalis Pharmaceuticals; intellectual property rights in the Cure Breast Cancer Foundation and Epizyme; professional services and activities (uncompensated) for the ECOG-ACRIN Cancer Research Group; equity and a fiduciary role or position in and professional services and activities (uncompensated) for Qiagen; and a fiduciary role or position in and professional services and activities for The Mark Foundation. C.M.R. declares professional services and activities for Amgen, AstraZeneca, Bridge Medicines, D2G Oncology, Harpoon Therapeutics and Jazz Pharmaceuticals; intellectual property rights in Daiichi Sankyo; and equity in Earli. M.F.B. declares professional services and activities for AstraZeneca and Paige.AI; professional services and activities (uncompensated) for *JCO Precision Oncology* and the *Journal of Molecular Diagnostics*; and intellectual property rights in SOPHiA GENETICS. P.R. declares professional services and activities for Biovica, Inivata, Novartis, Prelude Therapeutics and SAGA Diagnostics; professional services and activities (uncompensated) for Guardant Health, Paige.AI and Tempus Labs; and equity, a fiduciary role or position and intellectual property rights in Odyssey Biosciences. B.T.L. declares professional services and activities (uncompensated) for Amgen, the Asia Society, AstraZeneca, Bolt Biotherapeutics and Daiichi Sankyo; and intellectual property rights in Karger Publishers and Shanghai Jiao Tong University Press. G.J.R. declares professional services and activities (uncompensated) for the American Association for Cancer Research, the American Society of Clinical Oncology, Mirati Therapeutics, Pfizer, Takeda Pharmaceuticals and Verastem; and professional services and activities for Harborside Press, MJH Associates, the National Comprehensive Cancer Network, Phillips Gilmore Oncology Communications, Research to Practice and Triptych Health Partners. H.S. declares professional services and activities for Bayer, Pfizer, Regeneron Pharmaceuticals, Sanofi and WCG Oncology; and intellectual property rights in Elucida Oncology. J.S.R.-F. is an employee of AstraZeneca, has served as a consultant for Goldman Sachs, Paige.AI and REPARE Therapeutics; and has served as an adviser for Roche, Genentech, Roche Tissue Diagnostics, Ventana, Novartis, InVicro, GRAIL, Goldman Sachs, Paige.AI and Volition RX. J. Gao and M.C. are employees of Caris.

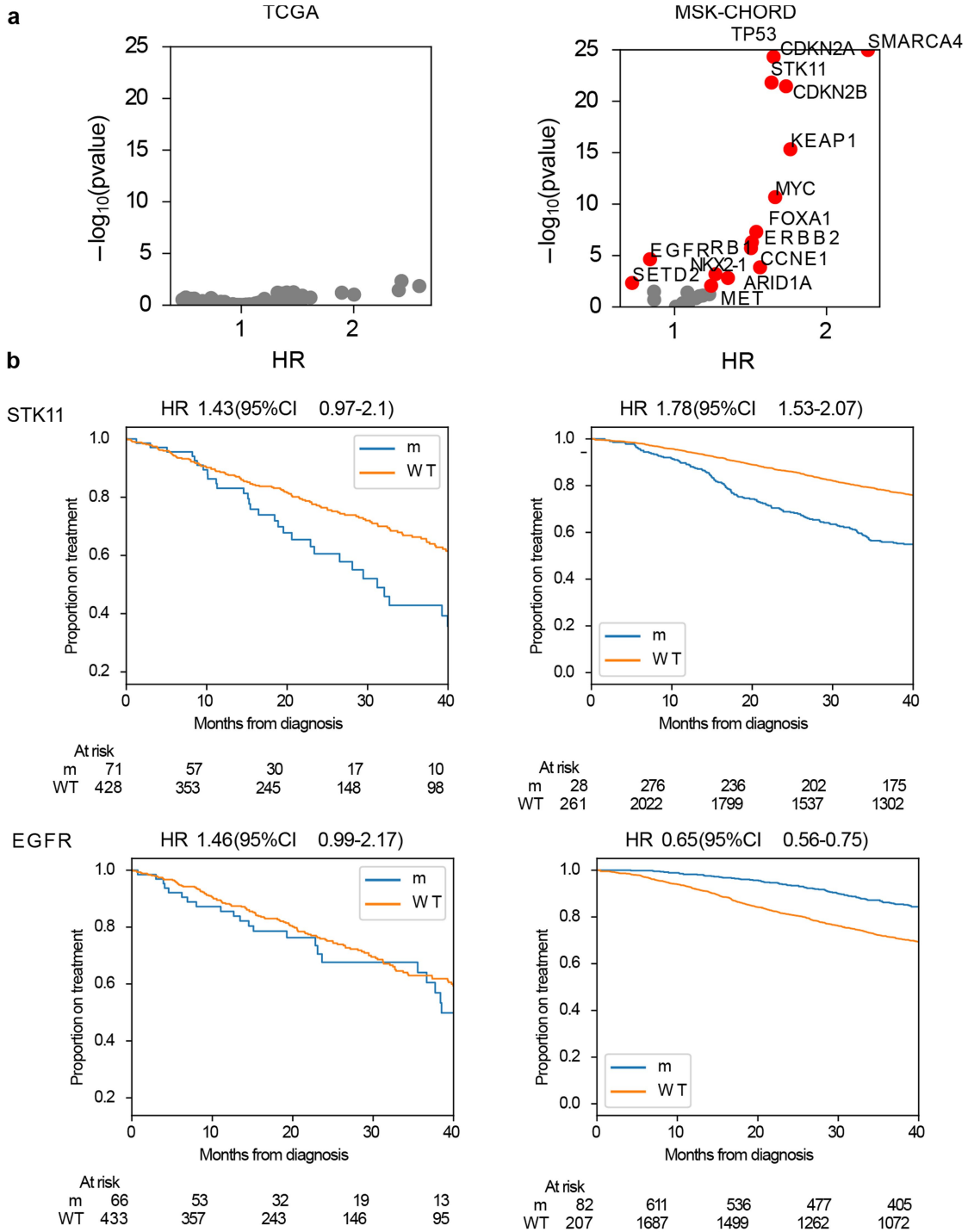
Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08167-5.
Correspondence and requests for materials should be addressed to Nikolaus Schultz.
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.



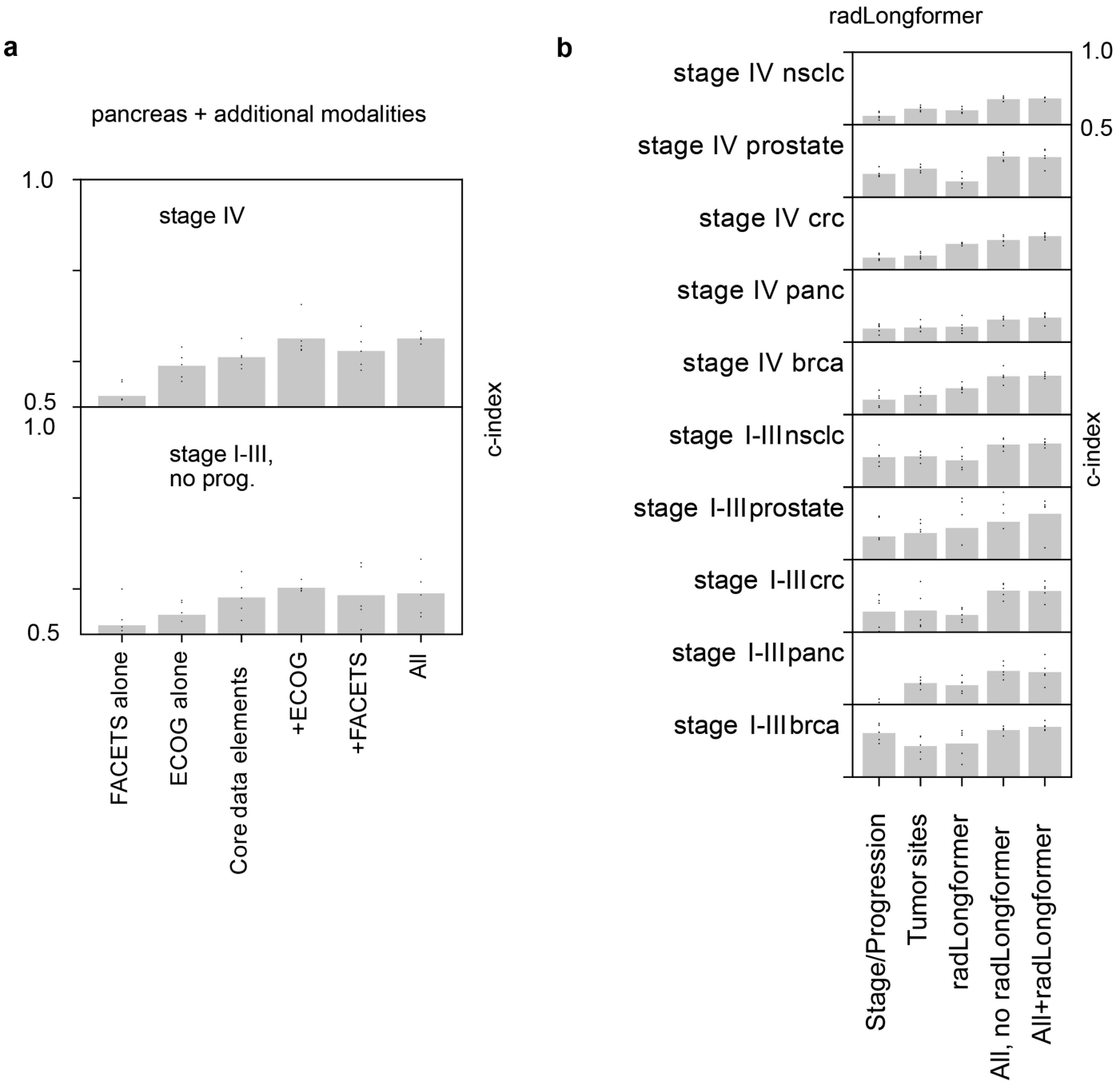
Extended Data Fig. 1 | Mismatch repair in immunohistochemistry and genomics. a. Relationship between mismatch repair (MMR) proficiency (pMMR)/deficiency (dMMR) on immunohistochemistry as annotated by NLP and microsatellite instability (MSI) as determined by MSK-IMPACT (MSISensor cutoff of 10, excluding indeterminate cases[62]). Boxplots depict median and inner quartile ranges (IQRs) with whiskers corresponding to 1.5xIQR. b. Kaplan-Meier curves show time to next treatment with stage IV colorectal cancer treated with immunotherapy (IO) stratified by MMR/MSI type.



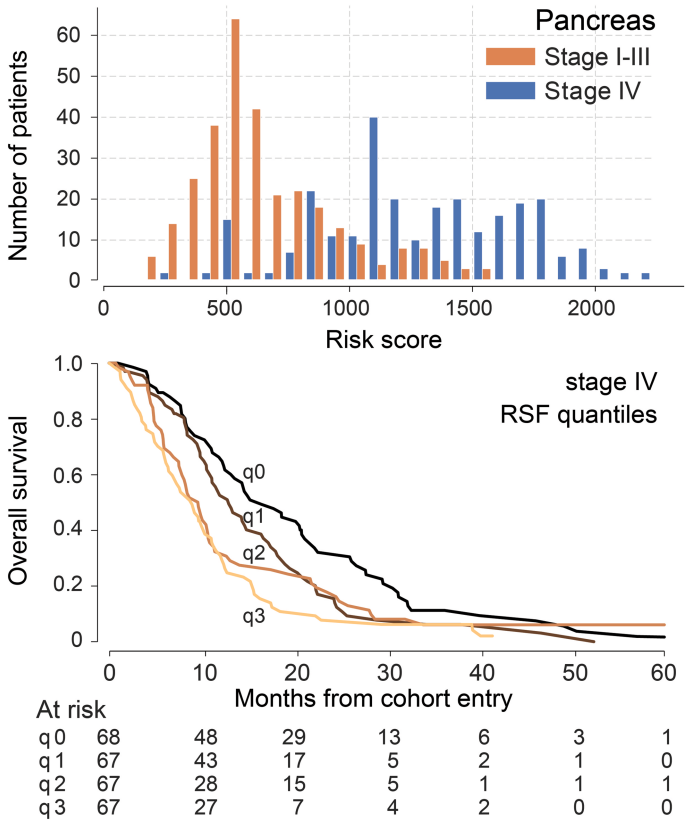
Extended Data Fig. 2 | Clinical and genomic representations of smoking. a. Proportion of patients with NSCLC (of the whole cohort) and oncogenic *EGFR* or *KRAS* alterations by clinical (NLP-derived) smoking status and smoking mutational signature status (+/− binomial 95%CI) in MSK-CHORD. Inset shows the distribution of dominant mutational signatures for the clinical smoking NLP +, SigMA smoking signature − subgroup. b. Scatterplot showing SBS4 observed from whole exome sequencing vs. pack years smoked at time of initial visit based on manual curation. c. Scatterplot showing tumor mutational burden (TMB) vs. pack years smoked in the exome cohort. d. Bar charts showing proportion and binomial 95%CI with a driver *EGFR* or *KRAS* mutation among patients with a significant clinical smoking history (≥ 15 pack years) and a non-dominant smoking signature in the exome cohort. e. Boxplots showing median, Q1-Q3, and 5–95%ile tumor purity among patients with ≥15 pack year smoking history, stratified by SBS4 status in the exome cohort.



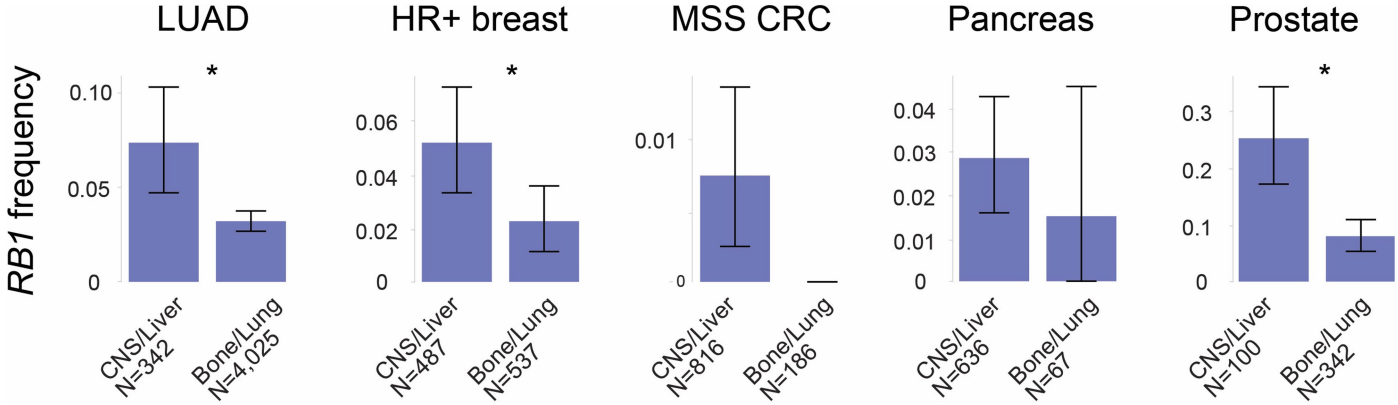
Extended Data Fig. 3 | Comparison of survival analyses between The Cancer Genome Atlas (TCGA) and MSK-CHORD. a. Volcano plots showing Cox proportional hazards models for specific oncogenic (by OncoKB) gene alterations (for all genes altered in at least 2% of the respective cohort) from time of diagnosis to time of death, right censored at last follow-up. For MSK-CHORD data is left truncated at time of sequencing (cohort entry) and only patients with stage I-III disease at diagnosis are shown. b. Selected representative survival curves stratified by oncogenic gene alteration presence. For example, *STK11* mutation is associated with worse survival in both cohorts although requires a sufficiently large cohort to show statistical robustness. *EGFR* mutation is associated with better OS only in MSK-CHORD, as these patients were treated following the advent of EGFR-targeted therapy, which was not standard of care during the timeframe of TCGA.



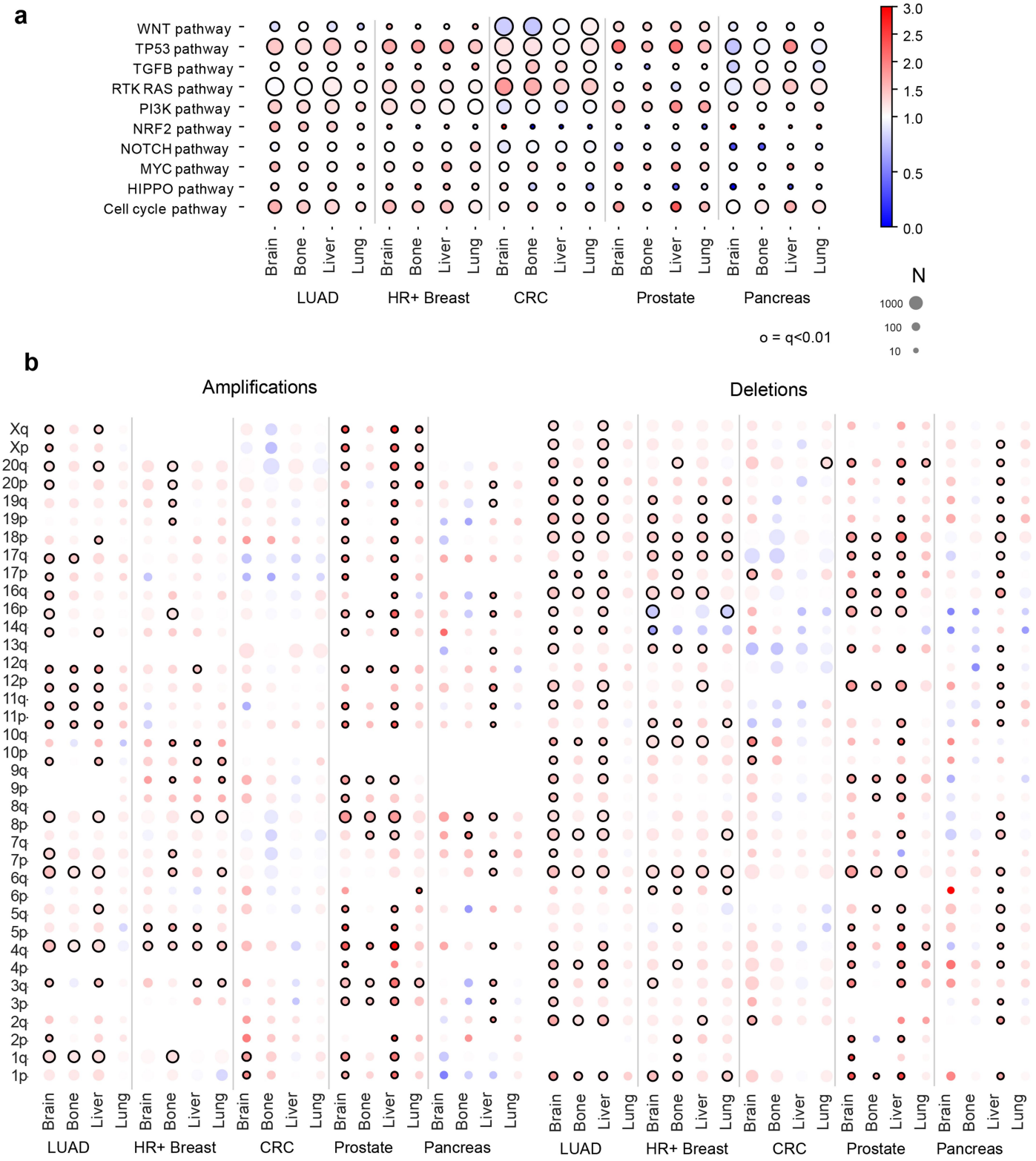
Extended Data Fig. 4 | Augmenting MSK-CHORD for predictive modeling. Mean c-indices from random survival forests by cancer type and stage and data modality (x axis) validated in 5-fold cross-validation using a. Secondary genomic data and performance status within the MSK-CHORD pancreatic cancer cohort and b. radLongformer. Dots correspond to results from individual validation folds.



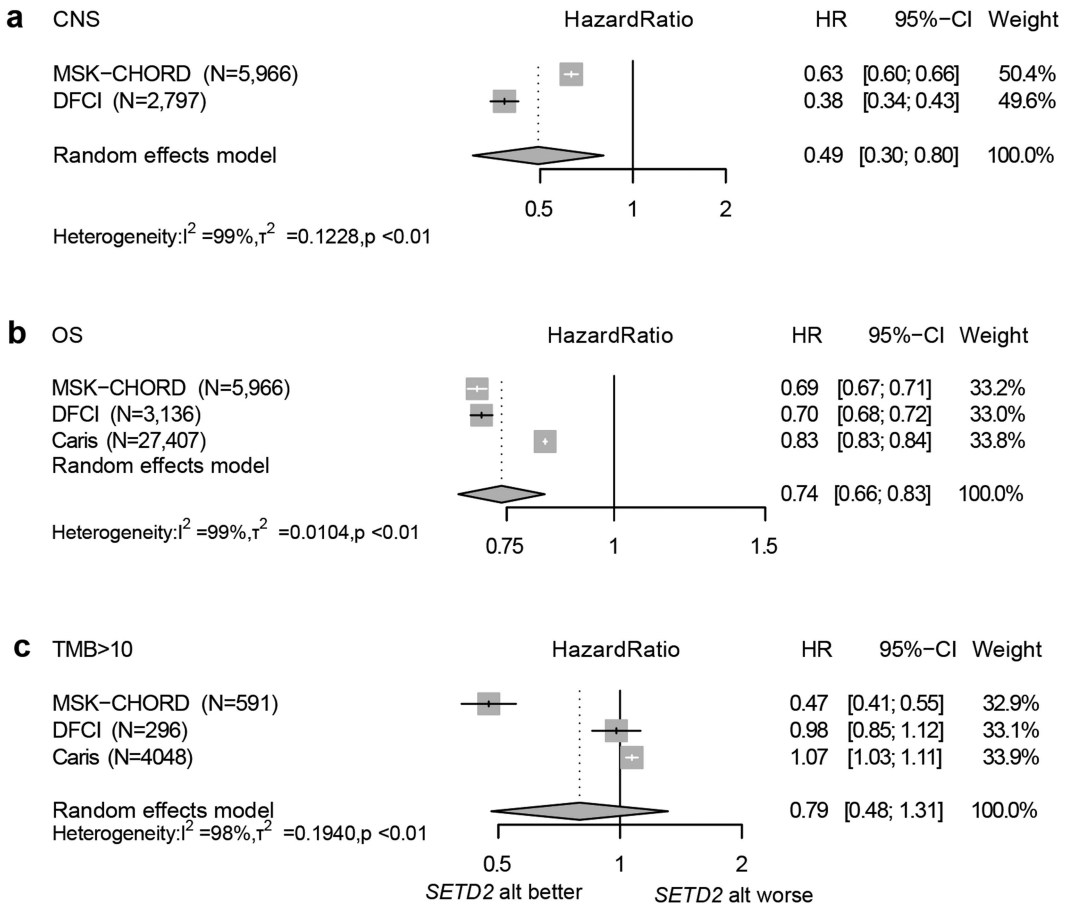
Extended Data Fig. 5 | Risk modeling. Risk score distribution for the non-MSK BPC cohorts and Kaplan-Meier survival curves based on computed risk quartiles for patients with pancreatic cancer.



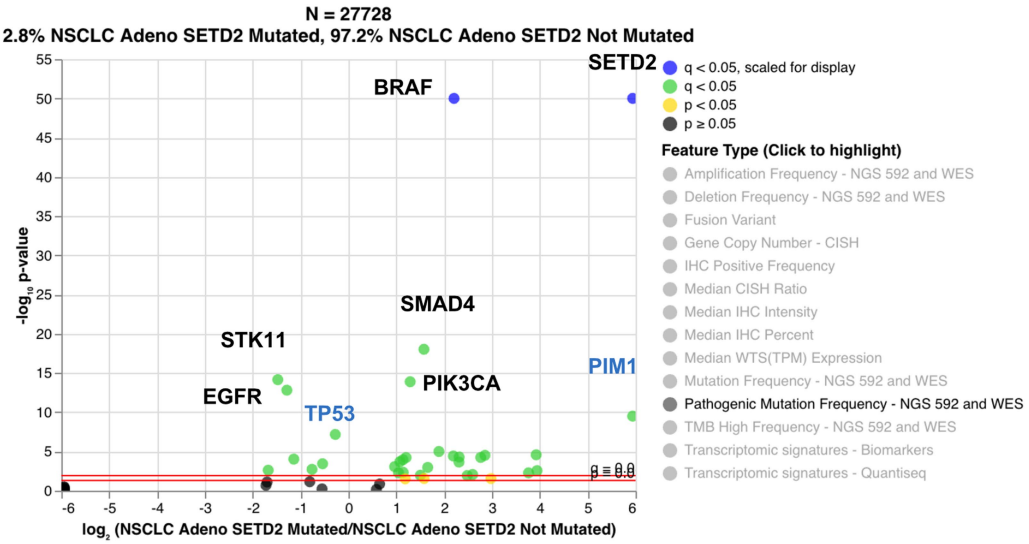
Extended Data Fig. 6 | RB1 alterations in metastatic samples. Frequency (proportion of total cohort) of oncogenic *RB1* alterations (+/− binomial 95%CI) in sequenced samples taken from the listed sites across the five studied cancer types. *=$p < 0.05$ by 2-sided Fisher Exact text.



Extended Data Fig. 7 | Derived genomic features and risk of future metastasis. Bubble plots showing hazard ratios (color), number of patients with alteration prior to site colonization (size) and statistical significance (Benjamini Hochberg FDR 0.01, black outline) for (a) pathway-level oncogenic alterations and (b) chromosome arm-level amplifications or deletions.



Extended Data Fig. 8 | Metastatic potential of *SETD2* mutant lung adenocarcinoma across multiple datasets. Hazard ratios +/−95%CI from Cox proportional hazards models as described in Methods. Combined hazard ratios are from random effects meta-analyses for (a) CNS metastasis, (b) overall survival (OS), and (c) time to next treatment or death from immunotherapy start for patients with lung adenocarcinoma and TMB>10 mut/Mb.



Extended Data Fig. 9 | Further *SETD2* genomic correlations. Volcano plot showing co-alteration or mutual exclusivity with *SETD2* driver mutations in patients with lung adenocarcinoma in a large cohort with exome sequencing (Caris).

nature portfolio

Corresponding author(s): Nikolaus Schultz

Last updated by author(s): Aug 12, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Code to perform the natural language processing used to create MSK-CHORD is restricted due to protected health information. Example code is available at https://github.com/clinical-data-mining/msk-chord-figures-public/tree/main/NLP
Data analysis	All code to perform the analyses presented here are available at https://github.com/clinical-data-mining. OncoCast is available at https://github.com/AxelitoMartin/OncoCast

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

MSK-CHORD will be made available on cBioPortal upon publication (study ID: msk_chord_2024). Data from the GENIE BPC can be accessed as previously described.41 For questions regarding access to Caris validation data contact jxiu@carisls.com.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This is reported in Fig. 1 and Table S6
Reporting on race, ethnicity, or other socially relevant groupings	This is reported in Table S6
Population characteristics	This is reported in Table S6
Recruitment	This is described in the Methods: Patients section of the manuscript. In short all MSK patients were enrolled as part of a prospective sequencing protocol (NCT01775072) or analyzed as part of a IRB-approved retrospective research protocols (MSK IRB Protocols 16-1463 and 19-368). Patients provided written, informed consent and were enrolled in a continuous, nonrandom fashion. Patients were enrolled at their provider's discretion. Selection bias may be present in that only patients from an academic cancer center in New York/New Jersey with providers likely to think patients would benefit from molecular profiling were enrolled. We attempted to study this bias by also validating results in two orthogonal datasets (Dana Farber Cancer Institute and Caris).
Ethics oversight	Institutional review board of Memorial Sloan Kettering and Dana Farber Cancer Institute

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was not predetermined and defined by the number of patients accrued up to the date of data cutoff.
Data exclusions	Exclusion criteria are described in the Patients: Methods section of the manuscript.
Replication	NA--each patient in the cohort acts as a biological replicate for any analysis in which they are included
Randomization	The study was nonrandomized. Controls for each specific analysis were assessed using observational data
Blinding	The study was not blinded. This was an observational study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	NCT01775072
Study protocol	Protocol available from MSK (internal IRB number 12-245)
Data collection	Data here is from patients at Memorial Sloan Kettering, an academic cancer center with sites in New York and New Jersey. Enrollment began in January 2014. Data here is from a September 9, 2023, snapshot
Outcomes	Primary outcomes included overall survival and metastasis to CNS, liver, bone and lung sites and were obtained as described in the manuscript. Secondary outcomes include response to immunotherapy, chemotherapy, and targeted therapy and were not prespecified. These were analyzed in an exploratory manner.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*