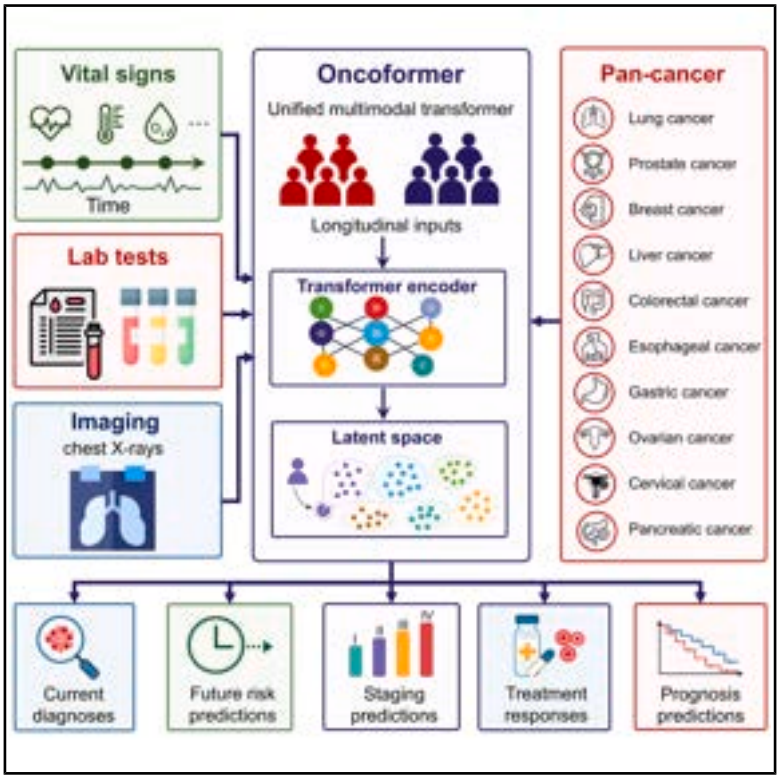
Cell

Resource

Advancing cancer detection and treatment using longitudinal routine clinical data

Graphical abstract



Highlights

- Oncoformer models the whole cancer continuum from routine lab tests and chest X-rays
- It detects existing cancer and predicts onset up to 1 year before clinical diagnosis
- It predicts tumor stage from routine clinical data
- It forecasts treatment response and recurrence-free survival across ten cancer types

Authors

Fei Liu, Kai Wang, Hui Xu, ..., Eric Oermann, Kang Zhang, International Consortium of Digital Twins in Healthcare and Medicine

Correspondence

kang.zhang@gmail.com

In brief

Liu et al. develop Oncoformer, a multimodal transformer that reads routine laboratory tests and chest X-rays already collected in everyday care. Across more than 3.6 million individuals, it detects cancer, infers tumor stage, and stratifies treatment response and recurrence risk, pointing toward risk-adapted cancer care built on data already in hand.

Liu et al., 2026, Cell *189*, 5995–6011
September 17, 2026 © 2026 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2026.07.009

CellPress

Cell

CellPress
OPEN ACCESS

Resource

Advancing cancer detection and treatment using longitudinal routine clinical data

Fei Liu,[1,2,3,4,29] Kai Wang,[3,5,29] Hui Xu,[1,5,29] Cheng Tang,[1,5,29] Xian Shen,[1,6,29] Meihao Wang,[7,8,29] Lei Yang,[9,29] Li Yang,[10,29] Li Liu,[11,29] Changxi Hu,[1,5] Gen Li,[1,5] Wei Wu,[12] Zixing Zou,[3,4] Bingzhou Li,[3,4] Sian Liu,[1,5] Jin Kang,[3] Jungho Kong,[13] Ting Li,[3] Io Nam Wong,[3] Xiaoying Huang,[14] Gang Chen,[1,6] Wenyang Lu,[1,5] Ian Ziyar,[3]

(Author list continued on next page)

[1]Clinical Data Science Institute, State Key Laboratory of Macromolecular Drugs and Large-scale Manufacturing, and The Zhejiang Key Laboratory of Intelligent Cancer Biomarker Discovery and Translation, the First Affiliated Hospital, Wenzhou Medical University, Wenzhou, China
[2]Department of Urology, State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[3]Artificial Intelligence Cross Disciplinary Research Institute, Faculty of Medicine, Macau University of Science and Technology, Macau, China
[4]Guangzhou National Laboratory, Guangzhou, China
[5]State Key Laboratory of Eye Health, Eye Hospital, Wenzhou Medical University, Wenzhou, China
[6]Department of General Surgery, Department of Hepatobiliary Surgery, Zhejiang-Germany Interdisciplinary Joint Laboratory of Hepatobiliary-Pancreatic Tumor and Bioengineering, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[7]Department of Radiology, The Second Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[8]Key Laboratory of Intelligent Medical Imaging of Wenzhou, Healthcare Center, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[9]Department of Orthopedics, Key Laboratory of Hepatosplenic Surgery of Ministry of Education, the First Affiliated Hospital, Harbin Medical University, Harbin, China
[10]Department of Respiratory and Critical Care Medicine, Center for Big Data and Intelligent Medicine, the First Affiliated Hospital of Chongqing Medical University, Key Laboratory of Digital Health and Intelligent Medicine, Chongqing Municipal Health Commission, Chongqing Translational Medicine Center, Chongqing, China
[11]Department of Health Management, Department of Infectious Diseases, Nanfang Hospital, Southern Medical University, Guangzhou, China
[12]Department of Big Data and Biomedical AI, College of Future Technology, Peking University, Peking-Tsinghua Center for Life Sciences, Beijing, China
[13]Departments of Medicine and Bioengineering, Moores Cancer Center, School of Medicine, UC San Diego, CA, USA
[14]Division of Pulmonary Medicine, the First Affiliated Hospital, Wenzhou Medical University, Wenzhou Key Laboratory of Interdisciplinary and Translational Medicine, Wenzhou Key Laboratory of Heart and Lung, Wenzhou, Zhejiang, China

(Affiliations continued on next page)

SUMMARY

Cancer management remains fragmented across its continuum, from late-stage diagnosis and salvage therapies to non-personalized surveillance. Here, we present Oncoformer, a unified multimodal transformer model trained on the China Oncology Multimodal Prediction and Surveillance Study (COMPASS) cohort (3.67 million individuals, 17.7 million clinical visits) and validated on independent external cohorts, including the UK Biobank. Oncoformer integrates longitudinal electronic health records with chest X-ray imaging to address multiple clinical tasks: pan-cancer diagnosis (area under the receiver operating characteristic curve [AUROC] = 0.956), future cancer prediction up to 1 year before diagnosis (AUROC = 0.869), tumor stage inference (mean AUROC > 0.90), patient-specific treatment-response forecasting, and recurrence-free survival stratification across ten cancer types (all $p < 0.01$). Staging predictions were independently validated against postoperative pathological endpoints and shown to converge on core cancer genomic pathways. By translating routine clinical data into a dynamic view of cancer evolution, Oncoformer provides a framework for risk-informed cancer prediction and treatment stratification using routine clinical data.

INTRODUCTION

Cancer remains a major global health challenge,[1] largely defined by a reactive clinical paradigm.[2] Diagnoses are predominantly triggered by the onset of symptoms, a point at which malignancies have often progressed to advanced stages, severely limiting therapeutic options and compromising patient outcomes.[3] The stark contrast in survival rates between

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Charlotte L. Zhang,[3] Yiwen Sun,[15] Weihong Lin,[8] Caiwen Ou,[16] Manson Fok,[3] Taiwa Hou,[17] Winston Wang,[18] Kanmin Xue,[19] Yun Yin,[1,20] Hao Zhu,[21,22] Jonathan Gootenberg,[21,22] Omar O. Abudayyeh,[21,22] Michael Karin,[13] Alexandre Loupy,[23,24] John E.J. Rasko,[3,25] Trey Ideker,[13,26] Huiyan Luo,[27] Eric Oermann,[28] and Kang Zhang[1,3,4,5,30,*]
International Consortium of Digital Twins in Healthcare and Medicine

[15]Nepean Hospital, Sydney, Australia
[16]Dongguan Hospital, Southern Medical University, Guangzhou, China
[17]Conde S. Januario Hospital, Macau, China
[18]Department of Internal Medicine, Mayo Clinic, Arizona, USA
[19]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
[20]Faculty of Health and Wellness, Faculty of Business, City University of Macau, Taipa, Macau SAR, China
[21]Division of Engineering in Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[22]Gene and Cell Therapy Institute, Mass General Brigham, Cambridge, MA, USA
[23]Université Paris Cité, INSERM U970 PARCC, Paris Institute for Transplantation and Organ Regeneration, F-75015 Paris, France
[24]Department of Kidney Transplantation, Assistance Publique – Hôpitaux de Paris, Paris, France
[25]Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, Peking University, Beijing, China
[26]Big Data Institute, Oxford University, Oxford, UK
[27]Department of Medical Oncology, Sun Yat-Sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Guangzhou, China
[28]Departments of Neurosurgery, Radiology, and Data Science, Neuroscience Institute, NYU Langone Medical Center, New York University, New York, NY, USA
[29]These authors contributed equally
[30]Lead contact
*Correspondence: kang.zhang@gmail.com
https://doi.org/10.1016/j.cell.2026.07.009

early- and late-stage disease underscores the need for approaches that can support earlier, risk-informed detection and intervention.[4,5] Achieving this requires tools that can prospectively identify individuals at high risk, enabling early intervention when treatments are most effective.

Significant progress has been made toward proactive cancer detection.[6–9] Organ-specific imaging, such as low-dose computed tomography (CT) for lung cancer and mammography for breast cancer, has proven effective but is limited by costs, radiation exposure, and an inability to provide a pan-cancer risk assessment. Concurrently, the advent of liquid biopsies, particularly multi-cancer early detection (MCED) tests based on circulating tumor DNA (ctDNA), holds good promise.[10–13] However, their widespread implementation is currently hampered by high costs and variable sensitivity for early-stage cancer detection, with their clinical utility in population-scale screening still under prospective evaluation.[14] A fundamental limitation shared by these approaches is that they often provide a static snapshot of risk at a single point in time, failing to capture cancer as a dynamic, longitudinal disease process.[10,15] A practical approach would support longitudinal, low-cost risk assessment using data already collected in routine care.

Routine clinical data generated daily in healthcare systems provide an opportunity to study longitudinal cancer-associated signals.[16] This longitudinal data, encompassing electronic health records (EHRs), laboratory tests, and opportunistic medical images, constitutes a detailed digital record of a patient's health trajectory over years or even decades.[17–20] In theory, these multi-faceted records contain the subtle, early footprints of nascent malignancy, long before a tumor becomes clinically apparent. Recent population-scale plasma proteomic studies in the UK Biobank, for instance, have detected systemic signals predictive of cancer years before clinical diagnosis,[21] albeit using specialized assays rather than the routine, low-cost data leveraged here. The routinely collected and relatively low-cost nature of these data makes them a practical substrate for building scalable models for longitudinal cancer risk assessment.[22–25] However, learning faint disease signals from such heterogeneous and noisy real-world data requires a purpose-built computational framework capable of modeling complex temporal and multimodal patterns.

Generative large language models (LLMs) are particularly suited for disease progression modeling, given their ability to utilize past events and their mutually dependent connections to predict the future sequence of medical outcomes.[26] Foundational models like Med-BERT (a BERT-based model pretrained on structured electronic health records) have adapted transformer architectures to learn powerful representations from large-scale structured EHRs,[19,27–29] enabling pan-cancer risk models[30–33] and cancer-specific prediction models.[34–46] Paralleling to these efforts, a trend toward multimodal integration has emerged, enhancing prognostic accuracy by combining clinical data with specialized inputs like histopathology or omics data.[24,47–57] Despite these advances, a critical gap persists. Existing frameworks are often task-specific—addressing either risk prediction, diagnosis, or prognosis in isolation—and frequently rely on specialized data not ubiquitously available for inexpensive, population-wide, proactive surveillance.[45,58–62] An effective approach should integrate widely accessible data modalities and support multiple clinically relevant prediction tasks within a shared framework.

Here, we address this challenge by developing a unified framework for cancer detection, staging, treatment-response prediction, and prognosis using routine clinical data. We first established the China Oncology Multimodal Prediction and Surveillance Study (COMPASS), one of the largest real-world cohorts of its kind, providing a large-scale basis for model development and validation. At the core of our framework is Oncoformer, a multimodal time-series transformer architecture designed to integrate longitudinal EHR and imaging data at a scale of 3.67 million individuals. We demonstrate that Oncoformer performs consistently in predictions

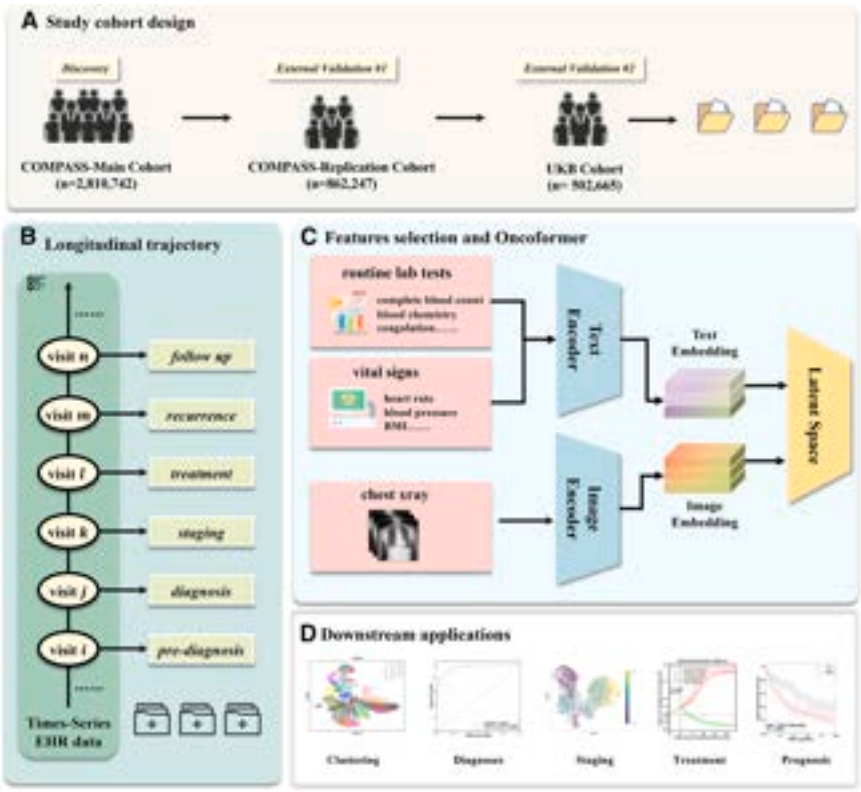


Figure 1. The COMPASS framework and Oncoformer architecture for multi-task cancer modeling from routine clinical data

(A) Overview of the multi-cohort study design. A large-scale discovery cohort (COMPASS-Main, *n* = 2,810,742) was used for model development. The model's robustness was assessed on two independent validation sets: a large internal hold-out cohort (COMPASS-Replication, *n* = 862,247) and an external cohort from the UK Biobank (UKB, *n* = 502,665 total participants, of whom 44,275 had a cancer diagnosis), ensuring generalization across different populations and healthcare systems.

(B) Conceptual illustration of the cancer-related clinical trajectory captured by longitudinal time-series EHR data. Oncoformer is designed to process sequential clinical visits (visit i to visit n), which contain latent signals corresponding to distinct phases of cancer evolution, including the pre-diagnostic window, initial diagnosis, clinical staging, treatment, recurrence, and long-term follow-up.

(C) The Oncoformer architecture. This unified, multimodal framework takes two streams of routine clinical data as input: (i) longitudinal EHR data, including laboratory tests and vital signs, processed by a text encoder, and (ii) opportunistic chest X-rays, processed by an image encoder. Both modalities are converted into high-dimensional embeddings and fused within a shared latent space to learn patient representations.

(D) Downstream applications of the learned patient representations. The Unified Model is trained for multiple cancer-related tasks, including unsupervised patient clustering, cancer diagnosis, pre-diagnostic tumor staging, prediction of treatment response, and patient prognosis stratification.

across multiple cancer-related prediction tasks. It achieves high-performance cancer early diagnosis and prediction up to a year in advance and non-invasively infers tumor stage using only pre-diagnostic data. Furthermore, Oncoformer's built-in treatment-response prediction module forecasts patient-specific responses to targeted therapies, and its derived risk scores provide clinically relevant prognostic stratification. These results support the potential of longitudinal routine-care data to inform multiple cancer-related prediction tasks within a shared modeling framework.

RESULTS

The COMPASS framework and Oncoformer architecture for multi-task cancer modeling

To build a predictive model capable of learning cancer-associated signals from routine clinical data, we developed a unified, multimodal framework, Oncoformer, and validated it on a large-scale, multi-cohort dataset (Figure 1). The core of our study is a large-scale discovery cohort, COMPASS-Main, which includes longitudinal data from 2,810,742 individuals. To ensure the model's robustness and generalizability, we established two independent validation cohorts: a large external #1 hold-out set, COMPASS-Replication (*n* = 862,247), and a geographically and ethnically distinct external cohort from the UK Biobank (UKB, *n* = 502,665) (Figure 1A).

The detailed process for cohort construction, including patient inclusion/exclusion criteria and the final distribution into training, tuning, and validation sets, is outlined in Figure S1. In total, the curated COMPASS dataset encompasses 17,748,334 clinical visits from 3,672,989 unique individuals, including 347,940 patients with cancer and 3,325,049 non-cancer controls (Table S1). The cancer and non-cancer groups were well-matched for age and gender (standardized mean difference [SMD] < 0.1 for both), and the cancer cohort represented a diverse range of common malignancies and was distributed across all clinical stages (I-IV), providing a rich substrate for model training.

Our central hypothesis is that the longitudinal sequence of routine clinical encounters contains latent signals that reflect the entire evolution process of cancer—from a pre-cancer state to various cancer stages, with corresponding predictive tasks in diagnosis, treatment options, recurrence, and survival (Figure 1B). To capture these complex temporal and multimodal signatures, we designed Oncoformer, a unified Transformer-based architecture (Figure 1C). The model takes inputs from two ubiquitous data streams: time-series EHRs, including laboratory tests and vital signs (Table S2), and chest X-rays. These are processed by dedicated text and image encoders, respectively, and their representations are projected onto and fused in a shared latent space.

To overcome the inherent challenges of real-world data, such as missing values and domain shifts across different cohorts, the Oncoformer architecture incorporates an advanced adversarial de-biasing strategy during training (Figure S2). Through self-supervised reconstruction and adversarial learning against both missingness patterns and cohort-specific features,

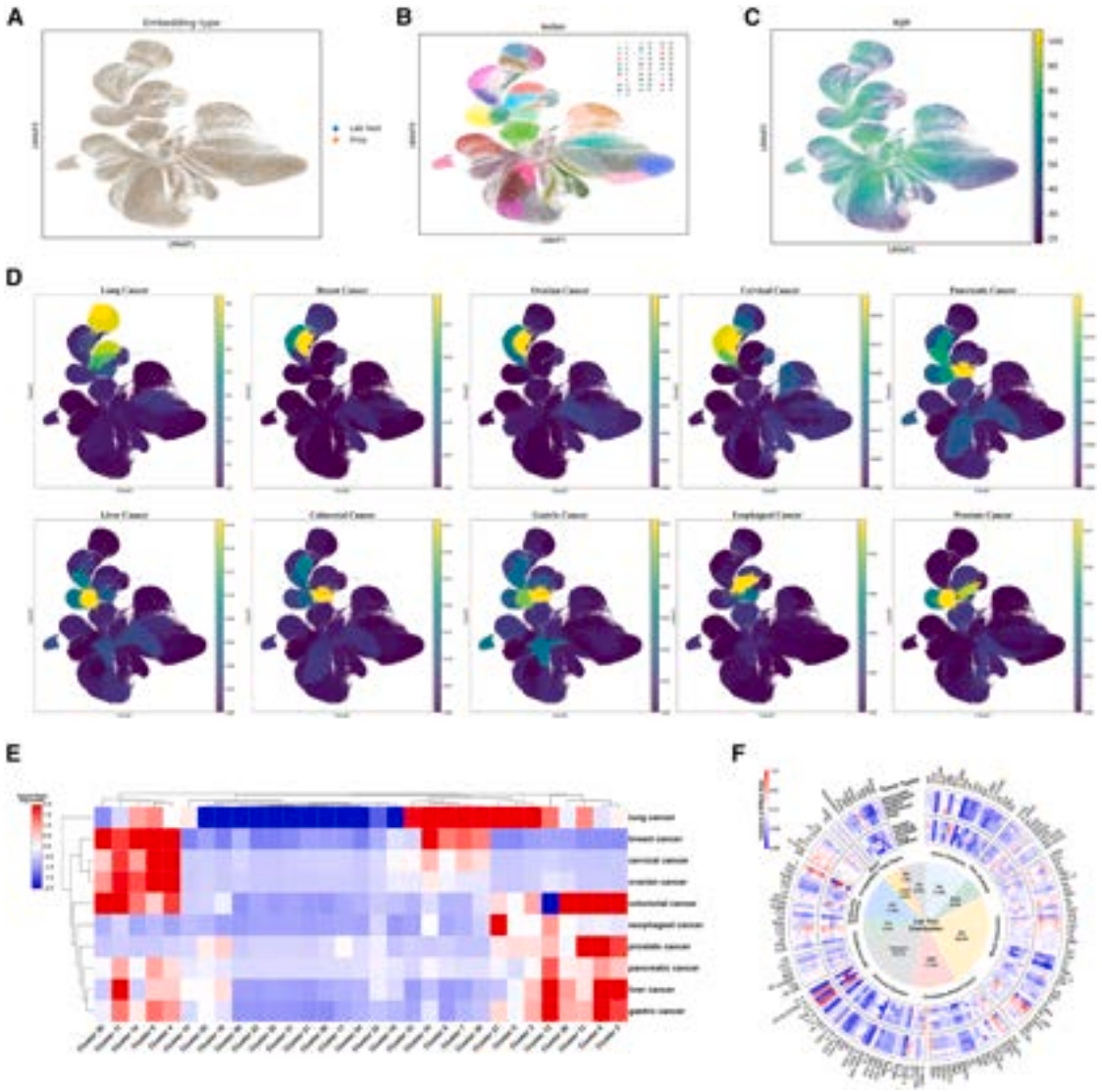


Figure 2. Oncoformer-generated latent space stratifies patients by unsupervised clustering

(A) UMAP visualization demonstrating the successful fusion of multimodal data. Latent representations derived from laboratory tests (lab-only) and from chest X-rays (image-only) for the same patients are projected onto the same latent space. The near-complete overlap of the two modality-specific embeddings, visually manifested as color convergence, validates the effectiveness of the model's cross-modal integration.

(B) Leiden clustering on the unified patient representations revealing a highly structured organization of the patient population.

(C) The latent space is strongly associated with fundamental biological variables. The UMAP embedding colored by patient age demonstrates a clear gradient, indicating that the model captures age-related physiological changes.

(D) UMAP plots illustrating the distribution of the 10 most common cancer types. Each malignancy occupies a distinct, albeit sometimes overlapping, topological region, demonstrating that Oncoformer learns organ-specific disease signatures from systemic data.

(E) Heatmap from a Cox regression analysis showing the HRs associating cluster membership with the risk for 10 major cancer types.

(F) Circular plots illustrating the systemic laboratory profiles for each of the 10 cancer types, with colors indicating the effect size (Cohen's d) of each marker compared with non-cancer controls.

the model learns robust, generalizable representations of a patient's underlying health state. This strong, de-biased representation serves as a versatile foundation for a wide array of downstream applications. The architecture includes dedicated heads for a range of tasks, from standard classification and regression to treatment-response prediction, which is designed to forecast patient-specific therapeutic outcomes[63] (Figures 1D and S2).

Oncoformer-generated latent space stratifies patients by unsupervised clustering

We first investigated the structure of the latent space generated by Oncoformer to assess its capacity for multimodal integration and its ability to reveal clinically relevant patient subgroups through unsupervised clustering (Figure 2). A key architectural feature of Oncoformer is its ability to synergistically integrate

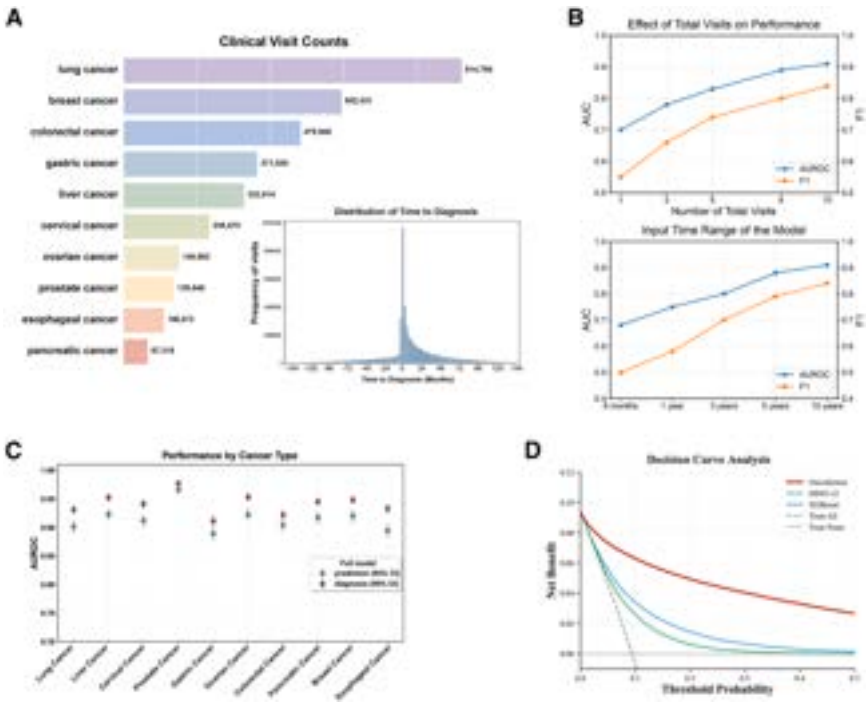


Figure 3. Oncoformer achieves high performance in current cancer diagnoses and future predictions

(A) Characterization of the longitudinal data available for the representative cancer types. The bar chart shows the total number of clinical visits captured for each malignancy. The inset histogram displays the temporal distribution of these visits relative to the date of diagnosis (time = 0), with the *y* axis showing the frequency of visits.

(B) Predictive performance (AUC and F1-score) as a function of data richness and temporal depth. The top panel shows performance stratified by the number of total visits supplied to the model. The bottom panel shows performance stratified by the total time span of patients' available health records.

(C) Comparison of diagnostic (red) and predictive (blue) performance (area under the receiver operating characteristic curve [AUROC]) across ten major cancer types on the COMPASS-Replication cohort. Points and error bars represent the AUROC and 95% CIs for the full multimodal Oncoformer model.

(D) Decision curve analysis (DCA) comparing the clinical net benefit of Oncoformer against two established baseline models, XGBoost and DINO v2, for pan-cancer prediction on the COMPASS-Replication cohort. The dashed gray line represents the "treat all" strategy, and the dotted black line represents the "treat none" strategy.

different data types. Projecting embeddings derived from either laboratory tests or chest X-rays alone or together demonstrated a consistent alignment within the shared latent space, confirming the model's success in creating a unified, holistic patient representation (Figure 2A). Inspired by a single-cell sequencing approach, we apply unsupervised Leiden clustering to this unified representation, which revealed robust patient clusters, indicating that the model organizes the patient population into a highly structured space with distinct, data-driven phenotypes (Figure 2B). This structure was not random but correlated strongly with fundamental biological variables, such as age, which exhibited a clear gradient across the uniform manifold approximation and projection (UMAP) projection (Figure 2C).

We next examined how different malignancies were organized within this space by visualizing the 10 most prevalent cancer types. This method revealed that the latent space intrinsically captures biological relatedness among tumors. For instance, malignancies of the female reproductive system, such as breast, ovarian, and cervical cancer, were concentrated in proximate geometric regions. Similarly, cancers of the digestive system, including esophageal, gastric, liver, and colorectal cancer, also clustered in close proximity to one another (Figure 2D).

Crucially, we sought to determine if these data-driven clusters corresponded to tangible differences in clinical risk. A Cox regression analysis confirmed this hypothesis, revealing statistically significant associations between cluster membership and the future risk for specific cancers (Figure 2E). The resulting heatmap of hazard ratios (HRs) strongly corroborated the spatial organization observed in the UMAP plot. For example, the clusters enriched for gynecological cancers indeed conferred a significantly higher risk for developing breast and ovarian cancer. Likewise, clusters where digestive system cancers were concentrated showed an elevated risk for colorectal and gastric malignancies. This demonstrates that the model's unsupervised clustering identifies patient subgroups with significantly different, organ-system-specific cancer risks. To elucidate the biological underpinnings of these cancer-specific signatures, we systematically profiled the systemic laboratory test markers associated with each of the 10 major cancer types (Figure 2F). This analysis revealed that each malignancy possesses a distinct systemic "fingerprint." While biologically related tumors shared some common features, distinct patterns emerged. For example, a strong inflammatory signature was a hallmark of lung cancer, whereas liver cancer was characterized primarily by markers in liver function enzymes.

Taken together, these findings show that Oncoformer constructs a highly structured latent space that is robustly integrated across modalities and is meaningful. It performs data-driven phenotyping, segregating patients into prognostically relevant subgroups defined by systemic biological signatures that align with known cancer biology. This provides a basis for downstream predictive tasks.

Oncoformer achieves high performance in current cancer diagnoses and future predictions

Building upon the structured latent space, we next evaluated Oncoformer's performance on the critical tasks of cancer diagnoses and the prediction of future cancer incidence. The model was trained on a rich longitudinal dataset that includes over 11.9 million clinical visits recorded months or even years prior to a formal cancer diagnosis, providing a robust foundation for predictive modeling (Figure 3A).

We first assessed how the model's predictive power is influenced by the richness and temporal depth of the input data, examining two key factors: the volume of data (number of visits) and the length of the historical observation window (Figure 3B). First, performance scaled directly with the number of total visits provided to the model, with both AUC and F1-score steadily increasing as more evidence was integrated (Figure 3B, top). Second, we found that a longer historical observation window significantly enhanced predictive accuracy. When stratifying patients by the total time span of their available health records, those with records spanning 10 years yielded better predictions than those with records spanning only 1 year (Figure 3B, bottom). This critical finding suggests that Oncoformer is not merely detecting acute, late-stage signals but is capable of learning from long-term, subtle drifts in a patient's physiological state that precede a cancer diagnosis by many years.

One of the primary goals of this work is to enable early, pre-symptomatic cancer detection. We therefore evaluated the model's ability to perform both concurrent diagnoses and future risk prediction in a pan-cancer setting on the independent COMPASS-Replication cohort. When tasked with diagnoses using data from patients who had already been diagnosed, Oncoformer achieved an AUROC of 0.956 (95% confidence interval [CI]: 0.955–0.958), confirming its ability to recognize overt cancer signals (Figure S3A, left). When using only pre-diagnostic data, Oncoformer could predict the future onset of cancer up to 1 year in advance with an AUROC of 0.869 (95% CI: 0.861–0.877). The predictive accuracy was even higher for shorter lead times, reaching an AUROC of 0.884 for a 6-month window (Figure S3A, right). This temporal performance gradient is clinically intuitive: the closer a patient is to diagnosis, the stronger the biological signals of nascent cancer become. Importantly, the distinction between these two tasks is defined by the temporal visibility of data: "diagnosis" utilizes data up to and including the index date, whereas "prediction" enforces a strict 90-day washout period that excludes all data within 90 days prior to diagnosis (see STAR Methods).

This high level of performance was robust across a wide range of individual malignancies. In the COMPASS-Replication cohort, the full multimodal Oncoformer maintained strong diagnostic and predictive AUROC values across all ten cancer types (Figure 3C). For instance, in predicting lung cancer, the model achieved a predictive AUROC of 0.901. Performance was particularly high for prostate cancer, with a predictive AUROC of 0.966 and a diagnostic AUROC of 0.977.

To rigorously benchmark Oncoformer against established methods, we compared its diagnostic performance on the COMPASS-Replication cohort with a panel of state-of-the-art baseline models (Table S3). For longitudinal laboratory and EHR data, we retrained XGBoost, Med-BERT,[27] and Delphi[26] from scratch on the COMPASS training set. For chest X-ray imaging, we benchmarked against ResNet-50[64] and DINO v2.[65] Across all ten cancer types, the unified Oncoformer architecture consistently outperformed each individual baseline, confirming that the observed performance gains stem from both architectural design and multimodal synergy. We further conducted an ablation analysis to evaluate the individual contributions of each data modality (Table S4). Five model configurations of increasing complexity were compared, ranging from a demographic baseline using only age and gender to lab test models trained with and without common serum tumor markers, an X-ray-only ablation model, and the full multimodal Oncoformer. After strictly excluding all tumor markers from the input, the lab test model still achieved a predictive AUROC of 0.818 for lung cancer, confirming that Oncoformer captures broad systemic physiological signals rather than relying on cancer-specific markers. Adding chest X-ray imaging yielded the largest and statistically significant gains for thoracic and intrathoracic-adjacent cancers (e.g., lung and esophageal cancer), consistent with the imaging modality's anatomical coverage, while contributing smaller incremental gains for other cancer types.

Beyond discrimination performance, we evaluated the clinical decision-making utility of Oncoformer on the COMPASS-Replication cohort. Decision curve analysis confirmed that Oncoformer provides a positive net benefit over both the "screen-all" and "screen-none" default strategies across a wide range of clinically relevant threshold probabilities, maintaining a net benefit advantage over XGBoost and DINO v2 (Figure 3D). Calibration analysis further demonstrated that Oncoformer's risk estimates closely tracked actual observed cancer frequencies, achieving a Brier score of 0.041 and an expected calibration error of 0.014, outperforming both XGBoost (Brier = 0.078, expected calibration error [ECE] = 0.065) and DINO v2 (Brier = 0.082, ECE = 0.072) (Figure S3B). Together, these results indicate that Oncoformer not only achieves high discrimination but also produces well-calibrated risk estimates suitable for informing clinical decision-making.

To validate the model's real-world utility, we conducted a prospective pilot study for colorectal cancer screening. We first performed a survey in an annual health check population in Wenzhou, China, who underwent comprehensive medical evaluations and routine clinical tests, including chest X-ray (CXR) and lab tests. From an initial pool of 38,059 consenting participants, a high-risk cohort of 3,025 individuals underwent a blinded study in which an Oncoformer model-based prediction and a routine colonoscopy were performed. In this screening setting, the Oncoformer achieved an AUROC of 0.927 (0.880–0.965). The model identified 54 out of 60 cancer cases, reaching a sensitivity of 90.0% with a specificity of 86.8%. These findings from a screening setting further underscore the potential of Oncoformer as a practical tool for non-invasive cancer risk stratification (Figure S4).

To further assess the generalizability and applicability of our framework, we evaluated its performance on the independent UK Biobank (UKB) cohort (Table S5), a predominantly White European population (94.0%) that is demographically and ethnically distinct from the Asian COMPASS cohorts. We compared the predictive power of features extracted by our pre-trained Oncoformer model against a UKB reference model trained from scratch using age, gender, and available routine laboratory tests. Even in this challenging external validation setting, Oncoformer demonstrated a consistent performance improvement over the baseline model across cancer types for both current diagnoses and future predictions (Table S6). For instance, in the task of predicting future lung cancer risk, leveraging Oncoformer's representations improved the AUROC by an absolute gain of 0.079 (from

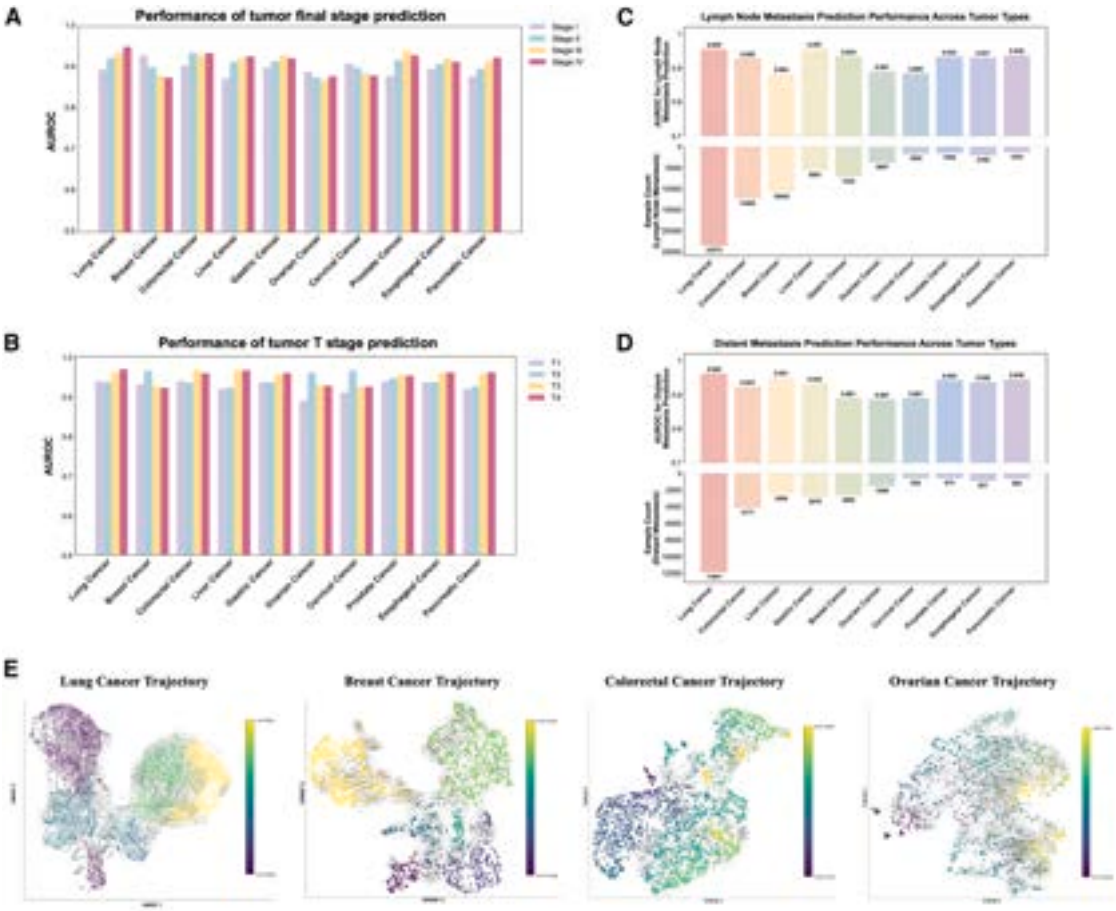


Figure 4. Oncoformer precisely infers tumor stage and maps disease progression trajectories from routine clinical data

(A–D) Performance of Oncoformer in inferring clinical tumor stage using only pre-diagnostic data. Bar plots show the AUROC values for predicting (A) overall tumor stage (I)–(IV), (B) primary tumor T stage (T1-T4), (C) lymph node metastasis (N stage), and (D) distant metastasis (M stage) across major cancer types. (E) Disease progression trajectories visualized in the UMAP latent space for representative cancers. Each point represents a single clinical visit from a patient who was eventually diagnosed with cancer. Points are colored by the final clinical stage at diagnosis (stages 1–2 as early stage and stages 3–4 as late stage). The arrows, generated using velocity analysis on the temporal sequence of visits, map out the inferred progression pathways from an early, pre-malignant state toward advanced disease.

0.708 to 0.787). Similarly, consistent performance gains were observed for other challenging malignancies such as pancreatic cancer (from 0.603 to 0.690). This consistent improvement in a completely different population confirms that Oncoformer learns fundamental, generalizable biological signals of developing cancer, establishing its potential as a feature extractor to enhance predictive modeling in diverse healthcare settings.

Oncoformer infers tumor stage and maps disease progression from pre-diagnostic data

A crucial determinant of cancer prognosis and treatment strategy is the tumor stage at diagnosis. We next investigated whether Oncoformer could infer the biological aggressiveness of a developing tumor—its clinical stage—using only routine data collected before a diagnosis was ever made, enabling non-invasive risk stratification for tumor invasiveness.

The model achieved high accuracy in clinical stage prediction across all major cancer types, with AUROC values consistently exceeding 0.90 for most malignancies in predicting overall staging (I–IV) (Figure 4A). The model's predictive capacity extended to the specific components of the tumor-node-metastasis (TNM) staging system. It accurately inferred the primary tumor's T stage (Figure 4B), the presence of lymph node metastasis (N stage) (Figure 4C), and, most critically, the likelihood of distant metastasis (M stage) (Figure 4D). For instance, it predicted metastatic potential with an AUROC of 0.923 for colorectal cancer and 0.943 for prostate cancer, suggesting that the systemic biological footprints of aggressive disease are detectable in routine data long before clinical manifestation. To visualize these progression dynamics, we mapped the inferred disease trajectories within the learned latent space, analogous to pseudotime analysis in single-cell genomics. By applying an adjusted RNA velocity algorithm to each patient's sequence of clinical visits, we constructed "disease progression maps" for individual cancers (Figure 4E). These maps revealed continuous trajectories from health toward malignancy, with clear, convergent pathways

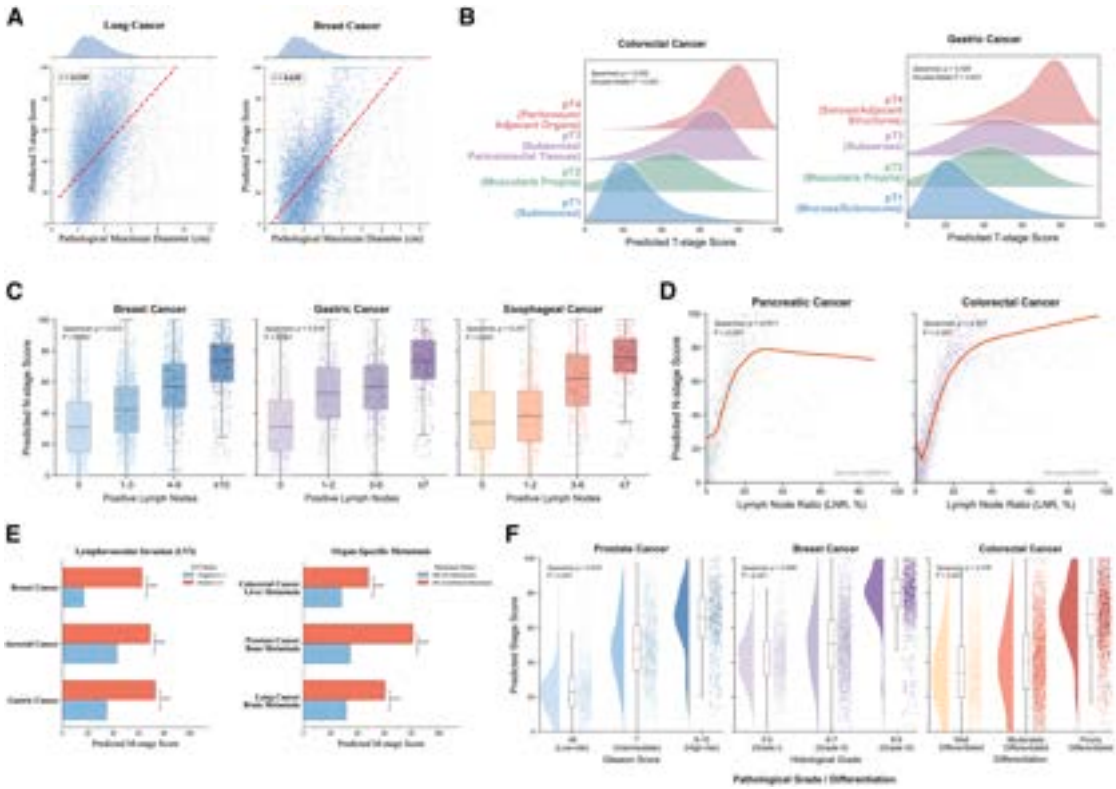


Figure 5. Independent validation of Oncoformer's pre-diagnostic staging predictions across multiple cancer types

(A) Scatterplots showing the correlation between Oncoformer's predicted T-stage score and pathological maximum tumor diameter (cm) for lung cancer and breast cancer. Red dashed lines indicate linear regression fits.

(B) Ridge density plots of predicted T-stage scores stratified by pathological invasion depth (pT1 through pT4) for colorectal cancer and gastric cancer. Anatomical invasion layers corresponding to each pT stage are annotated.

(C) Boxplots of predicted N-stage scores stratified by the absolute count of pathologically confirmed positive lymph nodes for breast cancer, gastric cancer, and esophageal cancer.

(D) Scatterplots with non-linear locally estimated scatterplot smoothing (LOESS) regression fits showing the relationship between predicted N-stage scores and the continuous lymph node ratio (LNR, defined as the proportion of positive lymph nodes among total examined nodes) for pancreatic and colorectal cancer.

(E) Validation of M-stage predictions. Left: predicted M-stage scores stratified by lymphovascular invasion (LVI) status (negative vs. positive) for breast, colorectal, and gastric cancer. Right: predicted M-stage scores stratified by confirmed organ-specific metastasis (M0 vs. M1) for colorectal cancer liver metastasis, prostate cancer bone metastasis, and lung cancer brain metastasis.

(F) Predicted overall stage scores stratified by established pathological grading systems: Gleason score categories in prostate cancer, histological grade in breast cancer, and tumor differentiation grade in colorectal cancer. Predictions were generated exclusively from pre-diagnostic clinical features. Ground-truth pathological variables, derived from postoperative surgical reports, were used strictly for independent validation and were never incorporated as model inputs.

originating from diffuse, early-stage-associated regions and progressing toward condensed late-stage regions, demonstrating that Oncoformer captures the dynamic, temporal evolution of cancer at a patient population level.

To independently validate that these staging predictions reflect genuine disease aggressiveness rather than statistical artifacts, we conducted an extensive retrospective analysis using deeply phenotyped sub-cohorts with matched postoperative pathology reports (Figure 5). Oncoformer's predicted T-stage scores correlated with pathological tumor diameter in lung cancer (Spearman $\rho = 0.519$) and breast cancer ($\rho = 0.635$) and accurately stratified pathological invasion depth (pT1 through pT4) in colorectal cancer ($\rho = 0.622$) and gastric cancer ($\rho = 0.526$) (Figures 5A and 5B). For lymph node involvement, predicted N-stage scores tracked the absolute count of positive lymph nodes across breast, gastric, and esophageal cancers ($\rho = \sim 0.437–0.518$) and correlated with the continuous lymph node ratio in pancreatic cancer ($\rho = 0.611$) and colorectal cancer ($\rho = 0.507$) (Figures 5C and 5D). The model also assigned significantly higher M-stage scores to tumors with confirmed lymphovascular invasion and to patients with organ-specific metastases, including colorectal cancer liver metastasis, prostate cancer bone metastasis, and lung cancer brain metastasis (all $p < 0.001$) (Figure 5E). Furthermore, predicted stage scores stratified established grading systems, including Gleason scores in prostate cancer ($\rho = 0.672$), histological grades in breast cancer ($\rho = 0.586$), and tumor differentiation grades in colorectal cancer ($\rho = 0.479$) (Figure 5F). These correlations, derived from pathological variables that were never exposed to the model during training, confirm that Oncoformer captures genuine biological signals of tumor aggressiveness from routine clinical data.

To further assess whether these clinical predictions correspond to established molecular mechanisms of tumor progression, we conducted a pathway-level analysis using independent tumor genomic profiling data. Approximately 520 genes from a targeted sequencing panel were organized into a four-level hierarchy of canonical cancer pathways. For each pathway, we computed the Oncoformer-Genomic Concordance (OGC), defined as the Spearman correlation between the Oncoformer staging inference score and the pathway-level somatic mutation burden (Figure S5A). While individual gene-level concordance was generally modest, reflecting the sparse and stochastic nature of single-gene mutations, aggregation into functional pathways revealed progressively stronger concordance, consistent with the pathway convergence phenomenon in cancer genomics and the demonstrated biological relevance of assembly-level somatic mutation burden in predicting clinical outcomes.[66,67] At the pathway level, three pathways achieved core status (OGC > 0.7, false discovery rate [FDR] < 0.01): cell cycle, *TP53* pathway, and DNA damage repair, while RAS/MAPK signaling and *PI3K/AKT/mTOR* showed moderate concordance (OGC 0.5–0.7). Gene-level analysis within these core pathways revealed that the highest concordance was driven by *TP53*, *CDKN1A*, and *MDM2* in the *TP53* pathway; *BRCA1*, *BRCA2*, *ATM*, and *ATR* in the DNA damage repair pathway; and *CDK4*, *CDK6*, *CCND1*, and *RB1* in the cell-cycle pathway (Figure S5B). These results provide independent genomic-level evidence that the systemic signals captured by Oncoformer from routine clinical data converge on well-characterized molecular pathways governing tumor progression.

We further performed Shapley additive explanations (SHAP) analysis to elucidate which routine clinical features drive the staging predictions and visualized the results as a pan-cancer biomarker network (Figure S6). The network reveals both shared and cancer-type-specific clinical signatures underlying staging inference. Features positioned within the inner ring, such as lactate dehydrogenase (LDH), D-dimer, C-reactive protein (CRP), and albumin (ALB), contribute to staging predictions across multiple cancer types, whereas features in the outer ring, such as prostate-specific antigen (PSA), neuron-specific enolase (NSE), CA19-9, and CYFRA21-1, are predominantly associated with a single malignancy.

Oncoformer predicts patient-specific responses to guide personalized therapies

An important challenge in oncology is predicting which patients will benefit from a specific treatment. We next investigated whether Oncoformer, using only pre-treatment clinical data, could forecast patient-specific responses to a wide range of anti-cancer therapies, thereby enabling customized treatment selection. We first constructed a dataset capturing the comprehensive landscape of modern cancer treatments, including diverse chemotherapy, targeted therapy, and immunotherapy agents (Figure 6A).

We then assessed whether the model's latent space, generated from pre-treatment data, contained information predictive of future therapeutic outcomes. We focused on a pan-therapy task of classifying patients as either "sensitive" or "resistant" based on their clinical response. A UMAP visualization of the patient embeddings showed clear separation between these two groups, indicating that the intrinsic features learned by Oncoformer are strongly associated with the determinants of treatment sensitivity and resistance (Figure 6B).

Beyond simple binary classification, we tasked Oncoformer with a more challenging quantitative goal: to predict the entire longitudinal trajectory of a patient's response to a specific therapy. We employed the model's dedicated treatment-response prediction module and evaluated this capability across six critical and diverse cancer-therapy scenarios on the COMPASS-Replication cohort. For patients with non-small cell lung cancer (NSCLC), the model predicted the tumor shrinkage trajectory in response to EGFR inhibitors (R^2 = 0.658) (Figure 6C). Similarly, it forecasted the response to endocrine therapy in breast cancer (R^2 = 0.702) (Figure 6D) and to cetuximab, an anti-EGFR antibody, in colorectal cancer (R^2 = 0.721) (Figure 6E). The model's predictive power extended beyond tumor size measurements to serum biomarker trajectories. In prostate cancer, it predicted the longitudinal decline of PSA levels following androgen deprivation therapy (ADT) (R^2 = 0.683) (Figure 6F), and in ovarian cancer, it forecasted response to olaparib, a poly(ADP-ribose) polymerase (PARP) inhibitor (R^2 = 0.704) (Figure 6G). Additionally, the model predicted the trajectory of alpha-fetoprotein (AFP) levels in patients with hepatocellular carcinoma treated with sorafenib (R^2 = 0.573) (Figure 6H), further demonstrating its ability to forecast serum biomarker dynamics across different cancer types and treatment modalities. In all six scenarios, Oncoformer not only distinguished between sensitive and resistant cohorts but also recapitulated the patient-specific dynamics of their response curves ($p < 0.001$ for all comparisons). These results demonstrate Oncoformer's capability to function as a predictive biomarker, integrating complex pre-treatment patient data to forecast therapeutic benefit across diverse treatment modalities.

Oncoformer enables dynamic surveillance and prognostic stratification

Beyond static predictions at discrete time points, a comprehensive cancer model should be capable of continuous, dynamic surveillance of disease states and provide robust prognostic information. We therefore evaluated Oncoformer's ability to forecast the future trajectory of key clinical variables and to stratify patients based on their long-term risk of recurrence. All analyses in this section were performed on the COMPASS-Replication cohort.

First, we tested the model's capacity for dynamically predicting the future values of critical, continuous variables. Oncoformer achieved test-set R^2 values of 0.641 for CA125, 0.603 for carcinoembryonic antigen (CEA), and 0.518 for AFP (Figure 7A). This predictive power extended to radiological measurements, where the model forecasted future changes in tumor size with a test-set R^2 of 0.571 (Figure 7B). These results, consistent across both validation and test sets, establish Oncoformer as a tool for non-invasive, continuous monitoring, potentially anticipating changes in disease burden before they are clinically measured.

Next, we assessed whether the holistic patient representation learned by Oncoformer could translate into a prognostic signature. We generated a single, continuous risk score for each patient using

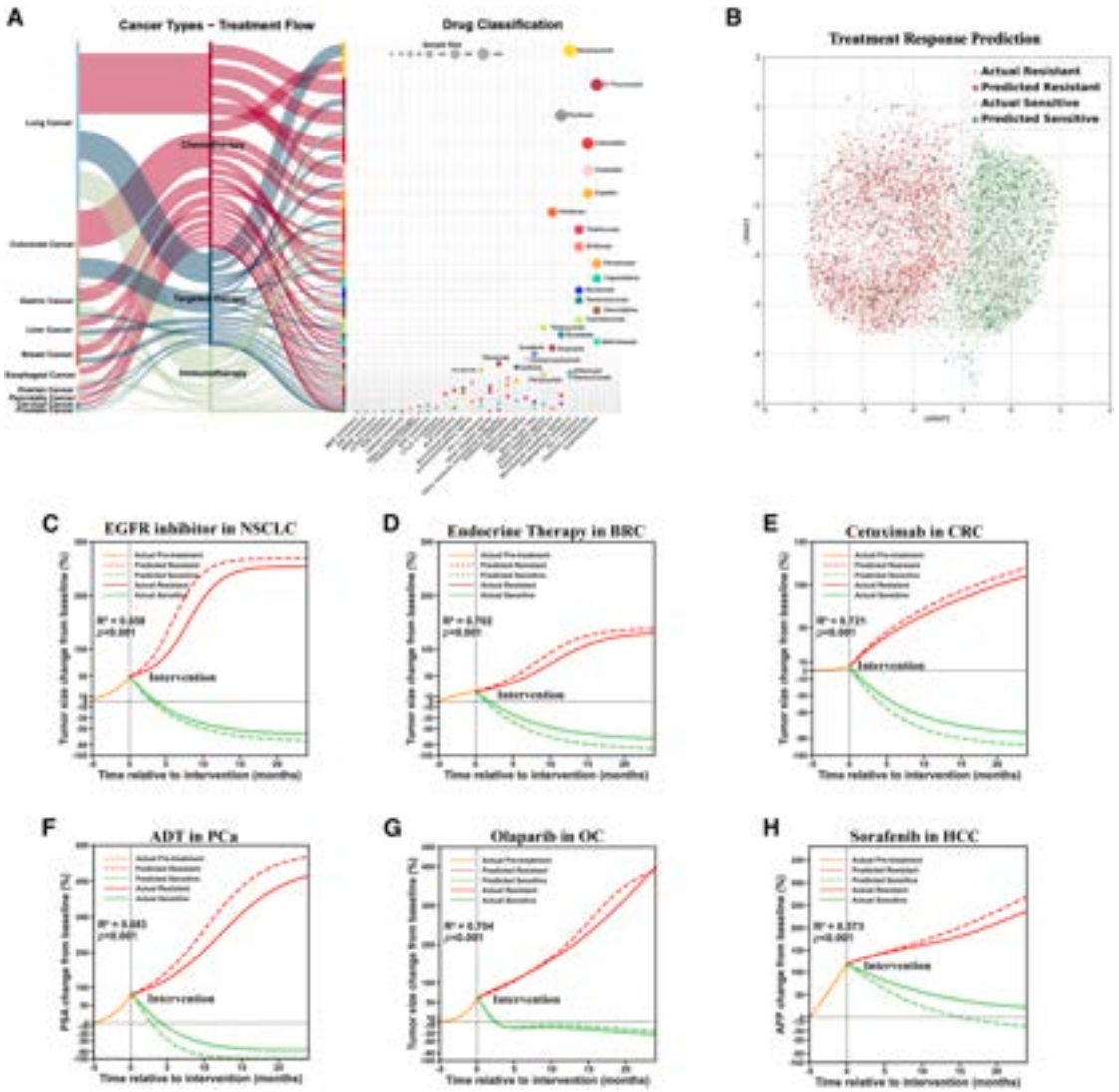


Figure 6. Oncoformer predicts patient-specific responses to guide personalized anti-cancer therapies

(A) Overview of anti-cancer therapies administered in the cohort. The alluvial plot (left) shows the flow from cancer types to major treatment modalities (chemotherapy, targeted therapy, immunotherapy). The lollipop plot (right) details the usage frequency of specific drug classes and individual drugs.

(B) UMAP visualization of patient representations prior to treatment initiation, colored by their predicted and actual response (sensitive vs. resistant). The clear separation demonstrates that the model learns features predictive of treatment outcome.

(C–H) Oncoformer's quantitative prediction of patient-specific treatment responses for representative cancer-therapy pairs. The model accurately forecasts the longitudinal trajectory of tumor size or serum markers (e.g., PSA) following intervention. Solid lines represent the mean actual response curves for patients with sensitive and resistant disease, while dashed lines represent the model's predictions. The R^2 value indicates the goodness of fit between predicted and actual post-treatment trajectories. The model demonstrates predictive accuracy for (C) epidermal growth factor receptor (EGFR) inhibitors in NSCLC, (D) endocrine therapy in breast cancer (BRC), (E) cetuximab in colorectal cancer (CRC), (F) ADT in prostate cancer (PCa), (G) olaparib in ovarian cancer (OC), and (H) sorafenib in hepatocellular carcinoma (HCC).

their pre-treatment data and used the median score to stratify them into high-risk and low-risk groups. This Oncoformer-derived risk score proved to be a significant prognostic factor for recurrence-free survival (RFS) across a wide array of malignancies (Figure 7C). The Kaplan-Meier survival curves revealed a statistically significant separation between the two risk groups for nearly all cancer types analyzed. Patients at high risk with colorectal cancer (HR = 2.86, 95% CI: 2.53–3.23, $p < 0.001$), gastric cancer (HR = 2.79, 95% CI: 2.46–3.17, $p < 0.001$), lung cancer (HR = 2.76, 95% CI: 2.54–3.01, $p < 0.001$), prostate cancer (HR = 2.66, 95% CI: 2.01–3.52, $p < 0.001$), breast cancer (HR = 2.62, 95% CI: 2.30–2.98, $p < 0.001$) and cervical cancer (HR = 2.58, 95% CI: 2.13–3.13, $p < 0.001$) had a worse prognosis compared with their low-risk counterparts. The risk score also provided significant stratification for liver cancer (HR = 2.46, 95% CI: 2.13–2.83), esophageal cancer (HR = 2.46, 95% CI: 1.95–3.11), ovarian cancer (HR = 1.97, 95% CI: 1.46–2.66), and pancreatic cancer (HR = 1.42, 95% CI: 1.10–1.83, $p = 0.008$) (all other $p < 0.001$).

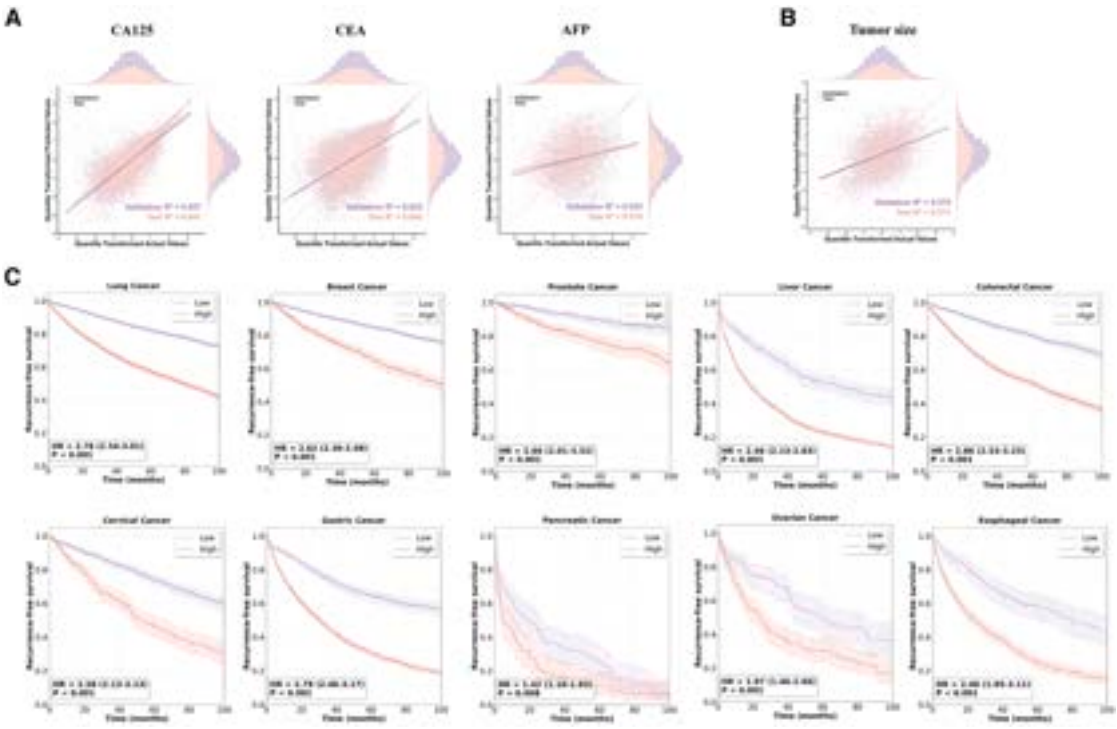


Figure 7. Oncoformer enables dynamic surveillance and prognostic stratification

(A and B) Performance of Oncoformer in dynamically predicting continuous clinical variables on both validation and test cohorts. Scatterplots show the high correlation (R^2) between model-predicted and actual values for (A) key serum tumor markers (CA125, CEA, and AFP) and (B) tumor size. Predictions were made for future time points using only prior clinical data.

(C) Oncoformer-derived risk scores effectively stratify patients for RFS across major cancer types. Kaplan-Meier curves show a significant separation in RFS between patients classified as high-risk vs. low-risk based on their pre-treatment Oncoformer representation. HRs and *p* values are shown for each cancer type.

To further quantify the discriminative power of this continuous risk score, we computed Harrell's concordance index (C-index) and benchmarked Oncoformer against three established survival models: Cox proportional hazards,[68] random survival forest,[69] and DeepHit[70] (Table S7). Oncoformer consistently achieved the highest C-index across all ten cancer types, including 0.836 (0.820–0.854) for prostate cancer, 0.809 (0.799–0.821) for breast cancer, and 0.780 (0.769–0.792) for colorectal cancer, confirming that the learned representation provides a prognostic risk score that outperforms conventional modeling approaches. This prognostic risk score was computed from data available before the initiation of first-line treatment, providing a baseline stratification at the point of the treatment decision. Because the temporal decoder operates under causal masking, an updated score can also be generated at any subsequent visit using only data available up to that point, allowing the same representation to function as a continuously updated longitudinal monitor during follow-up. SHAP analysis indicated that recurrence predictions were driven primarily by immune and metabolic markers, led by $CD4^+$ T-cell percentage and LDL-cholesterol (Figure S7).

DISCUSSION

In this study, we demonstrate that routine clinical data, when analyzed by a purpose-built transformer framework, can provide a longitudinal view of cancer-related clinical states. We established the COMPASS study, a large-scale real-world cohort, and developed Oncoformer, a unified multimodal model. This framework achieves high-accuracy pan-cancer detection and predictions up to a year before diagnosis, non-invasively infers tumor staging from pre-diagnostic data, predicts patient-specific responses to major therapies through its treatment-response prediction module, and provides prognostic stratification. Validation on large internal and external cohorts confirms the robustness and generalizability of our findings, supporting a data-driven approach.

Our work builds upon and extends a rich body of previous work leveraging computational models for cancer prediction. Previous efforts have demonstrated the power of EHRs alone, using transformer architectures for pan-cancer risk stratification[30] or predicting specific malignancies.[60,71,72] Other advanced models have shown the benefit of multimodality by integrating clinical data with specialized inputs like digital pathology or omics to refine prognosis.[47,48,73] Oncoformer differs from prior work in several ways. First, whereas most models are "point solutions" designed for a single task like diagnosis or outcome prediction, Oncoformer is designed as a unified, end-to-end framework that addresses the entire cancer continuum from a shared patient representation. Second, our multimodal approach effectively synergizes longitudinal EHRs with opportunistic chest X-rays—data modalities

that are routinely collected, low cost, and globally available, making our framework highly scalable for population-level surveillance in contrast to methods requiring specialized and costly data inputs or equipment. Finally, the development and validation of Oncoformer on a large-scale cohort (over 3.67 million individuals) provides a level of robustness and generalizability that is essential for real-world clinical translation.

These findings support the value of routine clinical data as a longitudinal source of cancer-associated signals. While targeted molecular biomarkers (e.g., *BRCA1/2* mutations or *EGFR* or *KRAS* mutations) have contributed substantially to personalized oncology, they are often static, costly, and provide a narrow view of a tumor's biology.[74–77] Similarly, emerging MCED tests, while promising, face challenges in cost and accessibility for population-scale screening.[10,78,79] In this context, routine-care data can be viewed as a longitudinal source of clinically relevant signals. Unlike snapshots from bespoke tests, these data capture the longitudinal evolution of a patient's health state, a concept consistent with recent longitudinal ctDNA monitoring studies showing that dynamic molecular trajectories can refine cancer risk stratification.[80] Our finding that a 10-year health record yields superior predictive power over a 1-year record illustrates this point, suggesting that the systemic "echo" of carcinogenesis is a long-term, evolving process. This predictive horizon, which remains robust to ~12–18 months and detectable until ~36 months before diagnosis (Figure S8), matches the preclinical detectable phase reported for many solid tumors.[81] Consistent with this, recent population-scale plasma proteomic analyses in the UK Biobank have identified tumor-promoting inflammatory signatures that predict lung cancer more than 5 years before diagnosis,[21] independently supporting the existence of a detectable, years-long systemic prelude to malignancy. It likely reflects the interval over which an occult, clonally evolving tumor engages the host through chronic inflammation, coagulation activation, and metabolic reprogramming,[82] the same axes that dominate the model's SHAP attributions (Figures S6 and S7) and converge on core cancer pathways (Figure S5). This notion contrasts with many existing AI models in oncology that often rely on a single data modality or a short time window around diagnosis.[19,38,40,83–85] By demonstrating that non-specific, system-wide changes, when integrated over time, can contribute to cancer prediction, our work suggests a scalable route toward risk-informed monitoring in routine care.

The Oncoformer model provides a shared representation for interpreting cancer-associated signals across multiple clinical tasks. Most AI applications in oncology are developed as "point solutions" for discrete tasks, such as classifying a pathology slide or detecting a nodule on a CT scan.[55,86,87] Oncoformer's architecture supports multiple related tasks—from pre-diagnostic risk estimation to post-treatment prognosis—using a shared patient representation. These results suggest that clinical signals across these different stages may be captured within a common computational framework. The performance gain from multimodal integration, where combining EHR and X-ray data consistently outperformed unimodal models, highlights the value of integrating complementary routine data streams. Furthermore, Oncoformer may help generate data-driven hypotheses about shared systemic pathways (such as inflammation and metabolism) that could be further investigated at the molecular level, linking large-scale data patterns with cancer biology.

A notable finding of our study is the ability of the unimodal X-ray model to predict cancers located far outside the thoracic cavity, such as cervical cancer (diagnostic AUROC = 0.905) and colorectal cancer (diagnostic AUROC = 0.896). This observation challenges the conventional view of medical imaging as a tool for detecting only local anatomical abnormalities. We hypothesize that this predictive power stems from the model's ability to learn subtle, systemic signatures of malignancy that manifest within the chest cavity. These could include alterations in the cardiomediastinal silhouette, changes in bone density indicative of metabolic shifts, or previously unrecognized patterns in the lung parenchyma reflecting a systemic inflammatory or immune response to a distant tumor. This suggests that a simple chest X-ray, when analyzed with a powerful deep learning model, may serve as a window into the patient's overall systemic health and their body's response to cancer, far beyond its traditional anatomical interpretation. Future research, including radiogenomics and biological validation, is warranted to uncover the specific pathophysiological mechanisms underlying these "long-distance" predictive signals.

The interpretability of deep learning models remains a key concern for clinical translation. Comprehensive SHAP analyses across six predictive tasks (Figure S7) reveal that Oncoformer does not rely on a single universal feature set but adapts its attention to task-relevant clinical markers. Albumin and neutrophil percentage were the most influential features for pan-cancer diagnosis, while LDH and ALP dominated staging predictions, and $CD4^+$ percentage and LDL-C were most important for recurrence prediction. Distinct feature sets also emerged for different treatment modalities, with indirect bilirubin and mean platelet volume ranking highest for EGFR inhibitor response and direct bilirubin and red cell distribution width for immunotherapy response. These task-specific profiles align with known cancer biology: the prominence of inflammatory and nutritional markers in diagnosis is consistent with the systemic inflammatory response to malignancy, the dominance of LDH and D-Dimer in staging parallels their established clinical roles as surrogate markers of tumor burden and hypercoagulability in advanced disease, and the emergence of immune cell markers for recurrence prediction reflects the recognized role of immune surveillance in preventing relapse. This biological coherence across tasks provides evidence that Oncoformer captures clinically meaningful signals rather than exploiting spurious correlations, partially addressing the inherent interpretability challenges of deep learning models in clinical settings.

Beyond the per-task feature rankings, the SHAP-derived biomarker network (Figure S6) is itself biologically informative through its two-tier structure. The shared, inner-ring markers correspond to convergent systemic responses common to malignancy irrespective of tissue of origin, with LDH reflecting glycolytic turnover and tumor burden, D-Dimer reflecting cancer-associated hypercoagulability, CRP reflecting systemic inflammation, and albumin reflecting catabolic and nutritional decline. By contrast, the outer-ring markers map onto lineage-specific secretory biology, exemplified by PSA, NSE, CA19-9, and CYFRA21-1 for prostate, neuroendocrine, pancreatobiliary,

and squamous tumors, respectively. The network thus recapitulates a fundamental dichotomy between the shared systemic footprint of malignancy and the tissue-specific identity of individual tumors and shows that Oncoformer's staging inference integrates both axes.

These findings suggest a route toward risk-adapted oncology workflows built on routine clinical data. Current screening programs, such as for breast or colorectal cancer, are typically age-based and follow a one-size-fits-all schedule. Our framework supports a risk-based monitoring model, where an individual's screening intensity and modality could be tailored to their updated risk score. In practice, Oncoformer is designed to operate within existing healthcare infrastructure: during routine annual health check-ups or regular clinical visits, the model processes already collected laboratory tests and opportunistic chest X-rays to generate a personalized cancer risk score, flagging individuals who exceed a clinical threshold for targeted confirmatory workup. This integration into routine-care pathways requires no additional specialized tests and could help optimize the allocation of healthcare resources by directing more intensive follow-up to those at highest risk. Furthermore, while our prospective pilot was initially evaluated in an asymptomatic high-risk cohort to optimize the positive predictive value, future large-scale prospective trials will expand to unselected, average-risk populations. In the therapeutic setting, Oncoformer moves beyond identifying prognostic markers to providing predictive insights. While genomic markers guide "what to target," our model addresses "who will respond" by integrating a patient's pre-treatment clinical state. This data-driven approach, which integrates longitudinal pre-treatment data to forecast therapeutic outcomes, could support clinicians and patients in shared decision-making, helping to select promising therapies and avoid the toxicity and cost of ineffective treatments.

Limitations of the study

Our study has several limitations that warrant discussion. First, its retrospective nature means that the clinical utility of Oncoformer must be validated in prospective, randomized controlled trials to assess its real-world impact on patient outcomes and healthcare costs. While our prospective colorectal cancer screening pilot provides initial evidence of clinical feasibility, large-scale prospective validation across multiple cancer types and diverse clinical settings remains essential before clinical deployment.

Second, although we employed adversarial de-biasing techniques and validated our model on an external UK cohort, the potential for unobserved confounding and dataset-specific biases inherent in real-world data cannot be entirely eliminated. The COMPASS cohorts were drawn from regional healthcare centers in China with a >99% Asian patient population, and while the UK Biobank validation demonstrates cross-population transferability, the UKB cohort itself is subject to well-documented volunteer selection bias. Continuously expanding our training corpus with multi-national, ethnically diverse cohorts remains a primary objective for future work.

Third, we observed that the model's performance varied across cancer types and clinical tasks. Predictive accuracy was lower for certain malignancies such as gastric cancer (predictive AUROC = 0.889) compared with prostate cancer (predictive AUROC = 0.966), likely reflecting differences in the strength of systemic prodromal signals across tumor biology. The model's performance on rare cancer types and fast-progressing malignancies that may lack a prolonged pre-diagnostic physiological footprint remains to be explored.

Fourth, patterns of missing data in routine EHRs present an ongoing challenge. While our adversarial missingness discriminator mitigates the influence of missing value patterns during training, the model's reliability may be compromised in healthcare settings with substantially different testing practices or extreme data sparsity compared with the training cohorts.

Finally, the interpretability of deep learning models, though partially addressed by our SHAP analyses and pathway-level molecular validation, warrants further investigation. The biological mechanisms linking routine clinical features to specific cancer predictions require additional experimental validation to fully translate computational findings into mechanistic understanding.

Conclusions

In conclusion, this study demonstrates that Oncoformer can use routinely collected clinical data to support multiple cancer-related prediction tasks. By integrating longitudinal health records and medical imaging at a scale of over 3.67 million individuals, the model captures clinically relevant signals across early prediction, staging inference, treatment-response prediction, and prognostic stratification. These findings suggest that longitudinal routine-care data may help inform risk-adapted cancer detection and treatment strategies, while prospective validation remains necessary before clinical deployment.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Kang Zhang (kang.zhang@gmail.com).

Materials availability

This study did not generate any new, unique reagents.

Data and code availability

All code used for the core analyses and implementation of the Oncoformer model is available at the GitHub repository (https://github.com/kaiwang13/Oncoformer) or the Macau University of Science and Technology AI-X code repository (https://must.digitwinmed.com/must/Oncoformer). Requests for access to de-identified data will be made to the corresponding author upon reasonable request and granted after approval by the data access committee. The UK Biobank data are an open access resource, and researchers can apply for access through the official access management system (www.ukbiobank.ac.uk).

CONSORTIA

The contributing International Consortium of Digital Twins in Healthcare and Medicine consortium members are Fei Liu, Kai Wang, Hui Xu, Cheng Tang, Xian Shen, Lei Yang, Li Yang, Li Liu, Changxi Hu, Gen Li, Wei Wu, Sian Liu, Jin Kang, Jungho Kong, Ting Li, Io Nam Wong, Charlotte L. Zhang, Caiwen Ou, Manson Fok, Taiwa Hou, Kanmin Xue, Yun Yin, Hao Zhu, Jonathan Gootenberg, Omar O. Abudayyeh, Michael Karin, Alexandre Loupy, John E.J. Rasko, Trey Ideker, Huiyan Luo, Eric Oermann, and Kang Zhang.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China(32470964, W2431057, 32100631, and 82188102); the Summit Advancement Disciplines of Zhejiang Province (Wenzhou Medical University - Pharmaceutics); the Macau Science and Technology Development Fund, Macao (0007/2020/AFJ, 0070/2020/A2, and 0003/2021/AKP); the Guangzhou National Laboratory, sponsored by Beijing Nova Program(20240484627); Capital's Funds for Health Improvement and Research (2024-4-40215); the China Postdoctoral Science Foundation (2023T160061); and the Macao Young Scholars Program (AM2023018).

We thank the Systemic and Tumor EHR and Image Reading and Evaluation Group (leaders: K.Z., H.L., and F.L.; senior physicians: Shancheng Ren, Jia Liu, Sujun Han, Fangming Wang, and Rui Chen; junior/evaluating physicians: J. Kang, S.L., K.X., M.F., T.H., J.E.J.R., A.L., E.O., and T.L.) and the Systemic and Tumor EHR and Image Reading and Evaluation Group (Wenkuan Wang, Dong Chen, Yajian Li, Bolin Jia, Mingshuai Wang, Shihao Zhang, Zhannan SI, Xinwei Wang, Duocai Li, Hui Sun, Jiakai He, Qihua Zhang, Baoyi Han, Ying Ma, Xuhang Wang, Yechi Zhang, Shuang Liu, Sai Pan, Chuyue Zhang, and Yue Niu).

AUTHOR CONTRIBUTIONS

K.Z. conceived the study, designed the overall research framework, coordinated the multi-institutional collaboration, and supervised the project. F.L. led the study, developed and implemented the predictive models, performed the statistical analyses and data-integrity review, and wrote the manuscript. K.W. co-led the study, contributing to model development and statistical analysis. F.L., K.W., H.X., C.T., X.S., M.W., Lei Yang, Li Yang, and L.L. contributed equally to this work. H.X., C.T., X.S., M.W., Lei Yang, Li Yang, and L.L. led cohort assembly, clinical data acquisition, harmonization, and validation at their respective participating centers and contributed to the interpretation of the clinical prediction results. C.H., G.L., W. Wu, Z.Z., B.L., S.L., J. Kang, T.L., I.N.W., X.H., G.C., W. Lu, I.Z., C.L.Z., Y.S., W. Lin, C.O., M.F., T.H., W. Wang, K.X., Y.Y., H.L., and E.O. contributed to clinical data acquisition, cohort assembly, data harmonization, quality control, clinical annotation, or validation across participating centers. C.H., G.L., W. Wu, Z.Z., B.L., W. Lu, Y.Y., and E.O. supported the computational analyses, model training and evaluation, and statistical validation. J. Kong, H.Z., J.G., O.O.A., M.K., A.L., J.E.J.R., T.I., H.L., E.O., and K.Z. provided interdisciplinary guidance on cancer biology, clinical translation, multimodal modeling, external validation, and interpretation of the findings. All authors critically reviewed the manuscript, discussed the results, approved the final version, and agree to be accountable for the work.

DECLARATION OF INTERESTS

K.W., F.L., and K.Z. filed a patent related to this work.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Multimodal data processing
  - Tabular data processing
  - Image data processing
  - Oncoformer architecture
  - Oncoformer architecture: visit encoder
  - Temporal embedding and decoder
  - Task-specific decoders
  - VAE-enhanced multimodal fusion
  - Training procedures
  - Adversarial missingness/bias control
  - Pretraining step
  - Disease prediction fine-tuning
  - Treatment Response Prediction Framework
  - Implementation details
  - Study Endpoints and Task Formulation
  - UK Biobank generalizability
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2026.07.009.

Received: September 22, 2025
Revised: April 11, 2026
Accepted: July 7, 2026
Published: July 27, 2026

REFERENCES

1. Siegel, R.L., Giaquinto, A.N., and Jemal, A. (2024). Cancer statistics, 2024. CA: Cancer J. Clin. *74*, 12–49. https://doi.org/10.3322/caac.21820.
2. Kurian, M., Adashek, J.J., and West, H.J. (2024). Cancer Care in the Era of Artificial Intelligence. JAMA Oncol. *10*, 683. https://doi.org/10.1001/jamaoncol.2023.7263.
3. Ngoi, N.Y.L., Gallo, D., Torrado, C., Nardo, M., Durocher, D., and Yap, T.A. (2025). Synthetic lethal strategies for the development of cancer therapeutics. Nat. Rev. Clin. Oncol. *22*, 46–64. https://doi.org/10.1038/s41571-024-00966-z.
4. Boire, A., Burke, K., Cox, T.R., Guise, T., Jamal-Hanjani, M., Janowitz, T., Kaplan, R., Lee, R., Swanton, C., Vander Heiden, M.G., et al. (2024). Why do patients with cancer die? Nat. Rev. Cancer *24*, 578–589. https://doi.org/10.1038/s41568-024-00708-4.
5. Jassim, A., Rahrmann, E.P., Simons, B.D., and Gilbertson, R.J. (2023). Cancers make their own luck: theories of cancer origins. Nat. Rev. Cancer *23*, 710–724. https://doi.org/10.1038/s41568-023-00602-5.
6. Rubinstein, W.S., Patriotis, C., Dickherber, A., Han, P.K.J., Katki, H.A., LeeVan, E., Pinsky, P.F., Prorok, P.C., Skarlupka, A.L., Temkin, S.M., et al. (2024). Cancer screening with multicancer detection tests: A translational science review. CA: Cancer J. Clin. *74*, 368–382. https://doi.org/10.3322/caac.21833.
7. Weeden, C.E., Hill, W., Lim, E.L., Grönroos, E., and Swanton, C. (2023). Impact of risk factors on early cancer evolution. Cell *186*, 1541–1563. https://doi.org/10.1016/j.cell.2023.03.013.
8. Zhou, R., Tang, X., and Wang, Y. (2024). Emerging strategies to investigate the biology of early cancer. Nat. Rev. Cancer *24*, 850–866. https://doi.org/10.1038/s41568-024-00754-y.
9. Crosby, D., Bhatia, S., Brindle, K.M., Coussens, L.M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R.C., Gambhir, S.S., Kuhn, P., et al. (2022). Early detection of cancer. Science *375*, eaay9040. https://doi.org/10.1126/science.aay9040.
10. Wan, J.C.M., Sasieni, P., and Rosenfeld, N. (2025). Promises and pitfalls of multi-cancer early detection using liquid biopsy tests. Nat. Rev. Clin. Oncol. *22*, 566–580. https://doi.org/10.1038/s41571-025-01033-x.
11. Zhou, H., Zhu, L., Song, J., Wang, G., Li, P., Li, W., Luo, P., Sun, X., Wu, J., Liu, Y., et al. (2022). Liquid biopsy at the frontier of detection, prognosis and progression monitoring in colorectal cancer. Mol. Cancer *21*, 86. https://doi.org/10.1186/s12943-022-01556-2.
12. Lawrence, R., Watters, M., Davies, C.R., Pantel, K., and Lu, Y.J. (2023). Circulating tumour cells for early detection of clinically relevant cancer. Nat. Rev. Clin. Oncol. *20*, 487–500. https://doi.org/10.1038/s41571-023-00781-y.
13. Malla, M., Loree, J.M., Kasi, P.M., and Parikh, A.R. (2022). Using Circulating Tumor DNA in Colorectal Cancer: Current and Evolving Practices. J. Clin. Oncol. *40*, 2846–2857. https://doi.org/10.1200/jco.21.02615.

14. Fitzgerald, R.C., Antoniou, A.C., Fruk, L., and Rosenfeld, N. (2022). The future of early cancer detection. Nat. Med. *28*, 666–677. https://doi.org/10.1038/s41591-022-01746-x.

15. Dowrick, A., Ziebland, S., Rai, T., Friedemann Smith, C., and Nicholson, B.D. (2024). A manifesto for improving cancer detection: four key considerations when implementing innovations across the interface of primary and secondary care. Lancet Oncol. *25*, e388–e395. https://doi.org/10.1016/s1470-2045(24)00102-5.

16. Tiwari, A., Mishra, S., and Kuo, T.R. (2025). Current AI technologies in cancer diagnostics and treatment. Mol. Cancer *24*, 159. https://doi.org/10.1186/s12943-025-02369-9.

17. Hippisley-Cox, J., and Coupland, C.A. (2025). Development and external validation of prediction algorithms to improve early diagnosis of cancer. Nat. Commun. *16*, 3660. https://doi.org/10.1038/s41467-025-57990-5.

18. Knevel, R., and Liao, K.P. (2023). From real-world electronic health record data to real-world results using artificial intelligence. Ann. Rheum. Dis. *82*, 306–311. https://doi.org/10.1136/ard-2022-222626.

19. Lipkova, J., Chen, R.J., Chen, B., Lu, M.Y., Barbieri, M., Shao, D., Vaidya, A.J., Chen, C., Zhuang, L., Williamson, D.F.K., et al. (2022). Artificial intelligence for multimodal data integration in oncology. Cancer Cell *40*, 1095–1110. https://doi.org/10.1016/j.ccell.2022.09.012.

20. Raab, R., Küderle, A., Zakreuskaya, A., Stern, A.D., Klucken, J., Kaissis, G., Rueckert, D., Boll, S., Eils, R., Wagener, H., et al. (2023). Federated electronic health records for the European Health Data Space. Lancet Digit. Health *5*, e840–e847. https://doi.org/10.1016/s2589-7500(23)00156-5.

21. Pandya, T., Zagorulya, M., Leung, M.M., Augustine, M., Liu, L.Y., Leppä, A.-M., Baruchel, U., Ng, S.W., Klockner, T., Mugabo, M., et al. (2026). Plasma signals of lung tumor promotion for molecular cancer prevention. Cell *189*, 3903–3921.e26. https://doi.org/10.1016/j.cell.2026.05.005.

22. Penberthy, L.T., Rivera, D.R., Lund, J.L., Bruno, M.A., and Meyer, A.M. (2022). An overview of real-world data sources for oncology and considerations for research. CA: Cancer J. Clin. *72*, 287–300. https://doi.org/10.3322/caac.21714.

23. Khan, M.S., Usman, M.S., Talha, K.M., Van Spall, H.G.C., Greene, S.J., Vaduganathan, M., Khan, S.S., Mills, N.L., Ali, Z.A., Mentz, R.J., et al. (2023). Leveraging electronic health records to streamline the conduct of cardiovascular clinical trials. Eur. Heart J. *44*, 1890–1909. https://doi.org/10.1093/eurheartj/ehad171.

24. Haue, A.D., Hjaltelin, J.X., Holm, P.C., Placido, D., and Brunak, S.R. (2024). Artificial intelligence-aided data mining of medical records for cancer detection and screening. Lancet Oncol. *25*, e694–e703. https://doi.org/10.1016/s1470-2045(24)00277-8.

25. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., and Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature *616*, 259–265. https://doi.org/10.1038/s41586-023-05881-4.

26. Shmatko, A., Jung, A.W., Gaurav, K., Brunak, S., Mortensen, L.H., Birney, E., Fitzgerald, T., and Gerstung, M. (2025). Learning the natural history of human disease with generative transformers. Nature *647*, 248–256. https://doi.org/10.1038/s41586-025-09529-3.

27. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit. Med. *4*, 86. https://doi.org/10.1038/s41746-021-00455-y.

28. Fahrner, L.J., Chen, E., Topol, E., and Rajpurkar, P. (2025). The generative era of medical AI. Cell *188*, 3648–3660. https://doi.org/10.1016/j.cell.2025.05.018.

29. Kraljevic, Z., Bean, D., Shek, A., Bendayan, R., Hemingway, H., Yeung, J.A., Deng, A., Baston, A., Ross, J., Idowu, E., et al. (2024). Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. Lancet Digit. Health *6*, e281–e290. https://doi.org/10.1016/s2589-7500(24)00025-6.

30. Jung, A.W., Holm, P.C., Gaurav, K., Hjaltelin, J.X., Placido, D., Mortensen, L.H., Birney, E., Brunak, S.R., and Gerstung, M. (2024). Multi-cancer risk stratification based on national health data: a retrospective modelling and validation study. Lancet Digit. Health *6*, e396–e406. https://doi.org/10.1016/s2589-7500(24)00062-1.

31. Bai, L., Lyu, J., Feng, J., Qiao, X., Qu, Y., Yang, G., Zhu, Y., Liao, L., Gao, H., Zang, A., et al. (2026). Cancer biomarkers discovered using pan-cancer plasma proteomic profiling. Nat. Biomed. Eng. *10*, 16–38. https://doi.org/10.1038/s41551-025-01448-y.

32. Chang, T.G., Park, S., Schäffer, A.A., Jiang, P., and Ruppin, E. (2025). Hallmarks of artificial intelligence contributions to precision oncology. Nat. Cancer *6*, 417–431. https://doi.org/10.1038/s43018-025-00917-2.

33. Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I.S., and van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. Nat. Med. *30*, 958–968. https://doi.org/10.1038/s41591-024-02902-1.

34. Sun, V.H., Heemelaar, J.C., Hadzic, I., Raghu, V.K., Wu, C.Y., Zubiri, L., Ghamari, A., LeBoeuf, N.R., Abu-Shawer, O., Kehl, K.L., et al. (2024). Enhancing Precision in Detecting Severe Immune-Related Adverse Events: Comparative Analysis of Large Language Models and International Classification of Disease Codes in Patient Records. J. Clin. Oncol. *42*, 4134–4144. https://doi.org/10.1200/jco.24.00326.

35. Rakaee, M., Tafavvoghi, M., Ricciuti, B., Alessi, J.V., Cortellini, A., Citarella, F., Nibid, L., Perrone, G., Adib, E., Fulgenzi, C.A.M., et al. (2025). Deep Learning Model for Predicting Immunotherapy Response in Advanced Non-Small Cell Lung Cancer. JAMA Oncol. *11*, 109–118. https://doi.org/10.1001/jamaoncol.2024.5356.

36. Sunami, K., Naito, Y., Saigusa, Y., Amano, T., Ennishi, D., Imai, M., Kage, H., Kanai, M., Kenmotsu, H., Komine, K., et al. (2024). A Learning Program for Treatment Recommendations by Molecular Tumor Boards and Artificial Intelligence. JAMA Oncol. *10*, 95–102. https://doi.org/10.1001/jamaoncol.2023.5120.

37. Bertsimas, D., Margonis, G.A., Sujichantararat, S., Koulouras, A., Ma, Y., Antonescu, C.R., Brennan, M.F., Martín-Broto, J., Tang, S., Rutkowski, P., et al. (2024). Interpretable artificial intelligence to optimise use of imatinib after resection in patients with localised gastrointestinal stromal tumours: an observational cohort study. Lancet Oncol. *25*, 1025–1037. https://doi.org/10.1016/s1470-2045(24)00259-6.

38. Lång, K., Josefsson, V., Larsson, A.M., Larsson, S., Högberg, C., Sartor, H., Hofvind, S., Andersson, I., and Rosso, A. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. *24*, 936–944. https://doi.org/10.1016/s1470-2045(23)00298-x.

39. Osipov, A., Nikolic, O., Gertych, A., Parker, S., Hendifar, A., Singh, P., Filippova, D., Dagliyan, G., Ferrone, C.R., Zheng, L., et al. (2024). The Molecular Twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients. Nat. Cancer *5*, 299–314. https://doi.org/10.1038/s43018-023-00697-7.

40. Pan, X., AbdulJabbar, K., Coelho-Lima, J., Grapa, A.I., Zhang, H., Cheung, A.H.K., Baena, J., Karasaki, T., Wilson, C.R., Sereno, M., et al. (2024). The artificial intelligence-based model ANORAK improves histopathological grading of lung adenocarcinoma. Nat. Cancer *5*, 347–363. https://doi.org/10.1038/s43018-023-00694-w.

41. Christiansen, F., Konuk, E., Ganeshan, A.R., Welch, R., Palés Huix, J., Czekierdowski, A., Leone, F.P.G., Haak, L.A., Fruscio, R., Gaurilcikas, A., et al. (2025). International multicenter validation of AI-driven ultrasound detection of ovarian cancer. Nat. Med. *31*, 189–196. https://doi.org/10.1038/s41591-024-03329-4.

42. Foersch, S., Glasner, C., Woerl, A.C., Eckstein, M., Wagner, D.C., Schulz, S., Kellers, F., Fernandez, A., Tserea, K., Kloth, M., et al. (2023). Multistain deep learning for prediction of prognosis and therapy response in

colorectal cancer. Nat. Med. *29*, 430–439. https://doi.org/10.1038/s41591-022-02134-1.

43. Hoang, D.T., Shulman, E.D., Turakulov, R., Abdullaev, Z., Singh, O., Campagnolo, E.M., Lalchungnunga, H., Stone, E.A., Nasrallah, M.P., Ruppin, E., et al. (2024). Prediction of DNA methylation-based tumor types from histopathology in central nervous system tumors with deep learning. Nat. Med. *30*, 1952–1961. https://doi.org/10.1038/s41591-024-02995-8.

44. Ogier du Terrail, J., Leopold, A., Joly, C., Béguier, C., Andreux, M., Maussion, C., Schmauch, B., Tramel, E.W., Bendjebbar, E., Zaslavskiy, M., et al. (2023). Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nat. Med. *29*, 135–146. https://doi.org/10.1038/s41591-022-02155-w.

45. Salim, M., Liu, Y., Sorkhei, M., Ntoula, D., Foukakis, T., Fredriksson, I., Wang, Y., Eklund, M., Azizpour, H., Smith, K., et al. (2024). AI-based selection of individuals for supplemental MRI in population-based breast cancer screening: the randomized ScreenTrustMRI trial. Nat. Med. *30*, 2623–2630. https://doi.org/10.1038/s41591-024-03093-5.

46. Placido, D., Yuan, B., Hjaltelin, J.X., Zheng, C., Haue, A.D., Chmura, P.J., Yuan, C., Kim, J., Umeton, R., Antell, G., et al. (2023). A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat. Med. *29*, 1113–1122. https://doi.org/10.1038/s41591-023-02332-5.

47. Yuan, W., Chen, Y., Zhu, B., Yang, S., Zhang, J., Mao, N., Xiang, J., Li, Y., Ji, Y., Luo, X., et al. (2025). Pancancer outcome prediction via a unified weakly supervised deep learning model. Signal Transduct. Target. Ther. *10*, 285. https://doi.org/10.1038/s41392-025-02374-w.

48. Mataraso, S.J., Espinosa, C.A., Seong, D., Reincke, S.M., Berson, E., Reiss, J.D., Kim, Y., Ghanem, M., Shu, C.H., James, T., et al. (2025). A machine learning approach to leveraging electronic health records for enhanced omics analysis. Nat. Mach. Intell. *7*, 293–306. https://doi.org/10.1038/s42256-024-00974-9.

49. Jiang, X., Hoffmeister, M., Brenner, H., Muti, H.S., Yuan, T., Foersch, S., West, N.P., Brobeil, A., Jonnagaddala, J., Hawkins, N., et al. (2024). End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. Lancet Digit. Health *6*, e33–e43. https://doi.org/10.1016/s2589-7500(23)00208-x.

50. Zhang, P., Zhang, Q., and Li, S. (2024). Advancing Cancer Prevention through an AI-Based Integration of Traditional and Western Medicine. Cancer Discov. *14*, 2033–2036. https://doi.org/10.1158/2159-8290.Cd-24-0832.

51. He, J.C., Moffat, G.T., Podolsky, S., Khan, F., Liu, N., Taback, N., Gallinger, S., Hannon, B., Krzyzanowska, M.K., Ghassemi, M., et al. (2024). Machine Learning to Allocate Palliative Care Consultations During Cancer Treatment. J. Clin. Oncol. *42*, 1625–1634. https://doi.org/10.1200/jco.23.01291.

52. Neri, P., Lee, H., and Bahlis, N.J. (2024). Artificial Intelligence Individualized Risk Classifier in Multiple Myeloma. J. Clin. Oncol. *42*, 1207–1210. https://doi.org/10.1200/jco.23.02781.

53. Bravo González-Blas, C., Matetovici, I., Hillen, H., Taskiran, I.I., Vandepoel, R., Christiaens, V., Sansores-García, L., Verboven, E., Hulselmans, G., Poovathingal, S., et al. (2024). Single-cell spatial multi-omics and deep learning dissect enhancer-driven gene regulatory networks in liver zonation. Nat. Cell Biol. *26*, 153–167. https://doi.org/10.1038/s41556-023-01316-4.

54. Amgad, M., Hodge, J.M., Elsebaie, M.A.T., Bodelon, C., Puvanesarajah, S., Gutman, D.A., Siziopikou, K.P., Goldstein, J.A., Gaudet, M.M., Teras, L.R., et al. (2024). A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. Nat. Med. *30*, 85–97. https://doi.org/10.1038/s41591-023-02643-7.

55. Saldanha, O.L., Quirke, P., West, N.P., James, J.A., Loughrey, M.B., Grabsch, H.I., Salto-Tellez, M., Alwers, E., Cifci, D., Ghaffari Laleh, N., et al. (2022). Swarm learning for decentralized artificial intelligence in cancer histopathology. Nat. Med. *28*, 1232–1239. https://doi.org/10.1038/s41591-022-01768-5.

56. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al. (2024). A foundation model for clinical-grade computational pathology and rare cancers detection. Nat. Med. *30*, 2924–2935. https://doi.org/10.1038/s41591-024-03141-0.

57. Widman, A.J., Shah, M., Frydendahl, A., Halmos, D., Khamnei, C.C., Øgaard, N., Rajagopalan, S., Arora, A., Deshpande, A., Hooper, W.F., et al. (2024). Ultrasensitive plasma-based monitoring of tumor burden using machine-learning-guided signal enrichment. Nat. Med. *30*, 1655–1666. https://doi.org/10.1038/s41591-024-03040-4.

58. Volinsky-Fremond, S., Horeweg, N., Andani, S., Barkey Wolf, J., Lafarge, M.W., de Kroon, C.D., Ørtoft, G., Høgdall, E., Dijkstra, J., Jobsen, J.J., et al. (2024). Prediction of recurrence risk in endometrial cancer with multimodal deep learning. Nat. Med. *30*, 1962–1973. https://doi.org/10.1038/s41591-024-02993-w.

59. Hollon, T.C., Pandian, B., Adapa, A.R., Urias, E., Save, A.V., Khalsa, S.S.S., Eichberg, D.G., D'Amico, R.S., Farooq, Z.U., Lewis, S., et al. (2020). Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. Nat. Med. *26*, 52–58. https://doi.org/10.1038/s41591-019-0715-9.

60. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. *25*, 954–961. https://doi.org/10.1038/s41591-019-0447-x.

61. Carrasco-Zanini, J., Pietzner, M., Davitte, J., Surendran, P., Croteau-Chonka, D.C., Robins, C., Torralbo, A., Tomlinson, C., Grünschläger, F., Fitzpatrick, N., et al. (2024). Proteomic signatures improve risk prediction for common and rare diseases. Nat. Med. *30*, 2489–2498. https://doi.org/10.1038/s41591-024-03142-z.

62. Calderaro, J., Žigutytė, L., Truhn, D., Jaffe, A., and Kather, J.N. (2024). Artificial intelligence in liver cancer - new tools for research and patient management. Nat. Rev. Gastroenterol. Hepatol. *21*, 585–599. https://doi.org/10.1038/s41575-024-00919-y.

63. Guo, R., Cheng, L., Li, J., Hahn, P.R., and Liu, H. (2020). A survey of learning causality with data: Problems and methods. ACM Comput. Surv. *53*, 1–37. https://doi.org/10.1145/3397269.

64. He, K., Zhang, X., Ren, S., and Sun, J. (2016) Deep Residual Learning for Image Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

65. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2024). DINOv2: Learning Robust Visual Features without Supervision. Trans. Mach. Learn. Res. https://openreview.net/forum?id=a68SUt6zFt.

66. Park, S., Silva, E., Singhal, A., Kelly, M.R., Licon, K., Panagiotou, I., Fogg, C., Fong, S., Lee, J.J.Y., Zhao, X., et al. (2024). A deep learning model of tumor cell architecture elucidates response and resistance to CDK4/6 inhibitors. Nat. Cancer *5*, 996–1009. https://doi.org/10.1038/s43018-024-00740-1.

67. Kong, J., Zhao, X., Singhal, A., Park, S., Bachelder, R., Shen, J., Zhang, H., Moon, J., Ahn, C., Ock, C.-Y., et al. (2024). Prediction of immunotherapy response using mutations to cancer protein assemblies. Sci. Adv. *10*, eado9746. https://doi.org/10.1126/sciadv.ado9746.

68. Cox, D.R. (1972). Regression Models and Life-Tables. J. R. Stat. Soc., B (Methodol.) *34*, 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

69. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., and Lauer, M.S. (2008). Random survival forests. Ann. Appl. Stat. *2*, 841–860. https://doi.org/10.1214/08-AOAS169.

70. Lee, C., Zame, W., Yoon, J., and van der Schaar, M. (2018). DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. Proceedings of the AAAI Conference on Artificial Intelligence *32*. https://doi.org/10.1609/aaai.v32i1.11842.

71. Wang, L., Yin, Y., Glampson, B., Peach, R., Barahona, M., Delaney, B.C., and Mayer, E.K. (2024). Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data. EBiomedicine *110*, 105442. https://doi.org/10.1016/j.ebiom.2024.105442.

72. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Satuluru, S., Kim, T., Banerjee, I., Gichoya, J., Trivedi, H., Lehman, C.D., et al. (2022). Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. J. Clin. Oncol. *40*, 1732–1740. https://doi.org/10.1200/jco.21.01337.

73. Liu, F., Beck, S., Yang, L., Luo, H., and Zhang, K. (2026). Advancing AI for multi-omics and clinical data integration in basic and translational cancer research. Nat. Rev. Cancer *26*, 497–512. https://doi.org/10.1038/s41568-026-00922-2.

74. Naqvi, S.A.A., Riaz, I.B., Bibi, A., Khan, M.A., Imran, M., Khakwani, K.Z.R., Raina, A., Anjum, M.U., Cobran, E.K., Warner, J.L., et al. (2025). Heterogeneity of the Treatment Effect with PARP Inhibitors in Metastatic Castration-resistant Prostate Cancer: A Living Interactive Systematic Review and Meta-analysis. Eur. Urol. *87*, 626–640. https://doi.org/10.1016/j.eururo.2024.12.007.

75. Bedrosian, I., Somerfield, M.R., Achatz, M.I., Boughey, J.C., Curigliano, G., Friedman, S., Kohlmann, W.K., Kurian, A.W., Laronga, C., Lynce, F., et al. (2024). Germline Testing in Patients With Breast Cancer: ASCO-Society of Surgical Oncology Guideline. J. Clin. Oncol. *42*, 584–604. https://doi.org/10.1200/jco.23.02225.

76. Cheng, H.H., Shevach, J.W., Castro, E., Couch, F.J., Domchek, S.M., Eeles, R.A., Giri, V.N., Hall, M.J., King, M.C., Lin, D.W., et al. (2024). BRCA1, BRCA2, and Associated Cancer Risks and Management for Male Patients: A Review. JAMA Oncol. *10*, 1272–1281. https://doi.org/10.1001/jamaoncol.2024.2185.

77. Singhal, A., Li, B.T., and O'Reilly, E.M. (2024). Targeting KRAS in cancer. Nat. Med. *30*, 969–983. https://doi.org/10.1038/s41591-024-02903-0.

78. Guerra, C.E., Sharma, P.V., and Castillo, B.S. (2024). Multi-Cancer Early Detection: The New Frontier in Cancer Early Detection. Annu. Rev. Med. *75*, 67–81. https://doi.org/10.1146/annurev-med-050522-033624.

79. Xu, Y., Zhu, S., Xia, C., Yu, H., Shi, S., Chen, K., He, Y., Deng, C., Jin, H., Liu, J., et al. (2025). Liquid biopsy-based multi-cancer early detection: an exploration road from evidence to implementation. Sci. Bull. (Beijing) *70*, 2852–2867. https://doi.org/10.1016/j.scib.2025.06.030.

80. Black, J.R.M., Karasaki, T., Abbott, C.W., Li, B., Veeriah, S., Al Bakir, M., Liu, W.K., Huebner, A., Martínez-Ruiz, C., Pawlik, P., et al. (2025). Longitudinal ultrasensitive ctDNA monitoring for high-resolution lung cancer risk prediction. Cell *188*, 7083–7098.e18. https://doi.org/10.1016/j.cell.2025.10.020.

81. Papier, K., Atkins, J.R., Tong, T.Y.N., Gaitskell, K., Desai, T., Ogamba, C.F., Parsaeian, M., Reeves, G.K., Mills, I.G., Key, T.J., et al. (2024). Identifying proteomic risk factors for cancer using prospective and exome analyses of 1463 circulating proteins and risk of 19 cancers in the UK Biobank. Nat. Commun. *15*, 4010. https://doi.org/10.1038/s41467-024-48017-6.

82. Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discov. *12*, 31–46. https://doi.org/10.1158/2159-8290.Cd-21-1059.

83. Song, A.H., Williams, M., Williamson, D.F.K., Chow, S.S.L., Jaume, G., Gao, G., Zhang, A., Chen, B., Baras, A.S., Serafin, R., et al. (2024). Analysis of 3D pathology samples using weakly supervised AI. Cell *187*, 2502–2520.e17. https://doi.org/10.1016/j.cell.2024.03.035.

84. Chen, L., Wang, Y., Cai, C., Ding, Y., Kim, R.S., Lipchik, C., Gavin, P.G., Yothers, G., Allegra, C.J., Petrelli, N.J., et al. (2024). Machine Learning Predicts Oxaliplatin Benefit in Early Colon Cancer. J. Clin. Oncol. *42*, 1520–1530. https://doi.org/10.1200/jco.23.01080.

85. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., and Zou, J. (2023). A visual-language foundation model for pathology image analysis using medical Twitter. Nat. Med. *29*, 2307–2316. https://doi.org/10.1038/s41591-023-02504-3.

86. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell *172*, 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.

87. Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., et al. (2020). Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell *181*, 1423–1433.e11. https://doi.org/10.1016/j.cell.2020.04.045.

88. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. *51*, D638–D646. https://doi.org/10.1093/nar/gkac1000.

89. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (Curran Associates Inc), pp. 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

90. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. 9th Int. Conf. Learn. Represent. https://openreview.net/forum?id=YicbFdNTTy.

91. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery), pp. 785–794. https://doi.org/10.1145/2939672.2939785.

92. Wang, K., Liu, F., Wu, W., Hu, C., Shen, X., Wang, M., Li, G., Zeng, F., Liu, L., Wong, I.N., et al. (2025). A full life cycle biological clock based on routine clinical data and its impact in health and diseases. Nat. Med. *31*, 4225–4235. https://doi.org/10.1038/s41591-025-04006-w.

93. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019) Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, pp. 4171–4186.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
UK Biobank dataset	UK Biobank	https://www.ukbiobank.ac.uk
STRING protein-protein interaction database v12.0	Szklarczyk et al.[88]	https://string-db.org; RRID:SCR_005223
Software and algorithms		
Python 3.10	Python Software Foundation	https://www.python.org; RRID:SCR_008394
PyTorch 2.6	PyTorch Foundation	https://pytorch.org; RRID:SCR_018536
scikit-learn 1.2	scikit-learn Community	https://scikit-learn.org; RRID:SCR_002577
lifelines 0.27	Journal of Open Source Software	https://github.com/CamDavidsonPilon/lifelines
Statsmodels	Seabold and Perktold	https://www.statsmodels.org; RRID:SCR_016074
SHAP	Lundberg and Lee[89]	https://github.com/shap/shap
DINOv2	Oquab et al.[65]	https://github.com/facebookresearch/dinov2
Vision Transformer (ViT)	Dosovitskiy et al.[90]	https://github.com/google-research/vision_transformer
XGBoost	Chen and Guestrin[91]	https://github.com/dmlc/xgboost; RRID:SCR_016787
Med-BERT	Rasmy et al.[27]	https://github.com/ZhiGroup/Med-BERT
Delphi	Shmatko et al.[26]	N/A
DeepHit	Lee et al.[70]	N/A
Assembly-level mutation burden analysis	Kong et al.[67]	https://github.com/idekerlab/drug-NeST
NeST-VNN (hierarchical visible neural network)	Park et al.[66]	https://github.com/idekerlab/NeST-VNN
Oncoformer source code	This paper	GitHub: https://github.com/kaiwang13/Oncoformer; GitLab: https://must.digitwinmed.com/must/Oncoformer

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study utilized data from multiple cohorts sourced from medical centers in China and the UK Biobank, following a data curation process that excluded records with low-quality or unreadable data (Figure S1).

The primary retrospective cohorts were drawn from the China Oncology Multimodal Prediction and Surveillance Study (COMPASS). The development cohort, designated COMPASS-Main, was established using de-identified electronic health records from two tertiary hospitals in Wenzhou, China: The First Affiliated Hospital of Wenzhou Medical University and The Second Affiliated Hospital of Wenzhou Medical University. This cohort comprised 2,810,742 patients with 11,934,576 clinical visits and was further partitioned into training (n=2,248,594), tuning (n=281,074), and internal validation (n=281,074) sets. To assess the model's generalizability within China, an independent external validation cohort (COMPASS-Replication) was established using data from four geographically and institutionally distinct hospitals: Nanfang Hospital, Southern Medical University; Dongguan Hospital of Southern Medical University; the First Affiliated Hospital, Harbin Medical University and The First Affiliated Hospital of Chongqing Medical University. This cohort included 862,247 patients with 5,813,758 clinical visits. In total, the curated COMPASS dataset encompasses 17,748,334 clinical visits from 3,672,989 unique individuals.

To evaluate cross-population generalizability, a third cohort was derived from the UKB, a large-scale prospective study of 502,665 participants (Table S5). From this resource, we identified 44,275 participants with a cancer diagnosis and constructed a case-control cohort with appropriately matched non-cancer controls.

To further evaluate the Oncoformer model's performance in a real-world clinical screening setting, we conducted a prospective colorectal cancer screening study in an annual health check population in Wenzhou, China. The process began with 38,059 participants who provided written informed consent. From this group, individuals were assessed for high-risk criteria using a detailed questionnaire. The study design and specific criteria for defining the high-risk cohort were adapted from the methodology established in our previous work.[92] The high-risk cohort for colorectal cancer is defined by a constellation of demographic, clinical, and genetic determinants that increase an individual's lifetime risk above that of the average-risk population. Key non-modifiable criteria for inclusion in this group are an age of 50 years or older, a personal medical history of colorectal polyps, adenomas, hematochezia, or inflammatory bowel disease (IBD), and a significant family history of CRC, particularly in a first-degree relative diagnosed before

the age of 60. This risk is further compounded by lifestyle factors, including obesity, a sedentary lifestyle, smoking, excessive alcohol consumption, and a diet rich in red and processed meats. While all enrolled individuals were strictly asymptomatic (aligning with the goal of early detection), this initial pilot purposefully utilized a risk-enrichment strategy. Consistent with established early-stage validation paradigms for non-invasive tests, this approach ensured an adequate disease prevalence (~2.0%) for statistical evaluation, while ethically avoiding unnecessary invasive colonoscopies in completely low-risk populations. After applying exclusion criteria, a final cohort of 3,025 high-risk individuals who underwent comprehensive medical evaluation and routine clinical tests (including chest X-rays and laboratory tests) were included for analysis. Each participant received a risk prediction from the Oncoformer model and subsequently underwent a routine colonoscopy with pathological confirmation, which served as the gold standard for diagnosis. The model's predictions were compared against the colonoscopy and pathology results to determine diagnostic accuracy.

The institutional review board of each participating center approved the study. The study was registered at clinicaltrials.gov (NCT06791473). All work was conducted in compliance with the principles of the Declaration of Helsinki, as well as all applicable national and local laws and regulations regarding patient privacy and data protection in China.

METHOD DETAILS

Multimodal data processing

We structured the patient's clinical history as a chronological sequence of multimodal clinical visits. Each patient's longitudinal record is represented as a time-ordered sequence

$$S = \{(X_0, I_0, T_0), (X_1, I_1, T_1), \ldots, (X_L, I_L, T_L)\}$$

where: X_i denotes the vector of tabular clinical variables (including continuous and categorical laboratory test results and clinical measurements) collected at the i-th visit. I_i denotes the chest X-ray image associated with the i-th visit. If no image was taken during a visit, I_i is treated as a null input. T_i represents the time elapsed (in days) since the initial visit, with $T_0 = 0$ by definition, L is the total number of visits for the patient.

Tabular data processing

The tabular clinical variables in X_i were processed as follows to handle outliers, perform normalization, and encode missing values. Continuous variables were quantized using the formula $\mathrm{D}(x) = \lfloor (x - X_{min})/(X_{max} - X_{min}) * d_{cont} \rfloor$, where $\lfloor X \rfloor$ represents the floor function, X_{max} the maximum value of feature x, X_{min} the minimum value of feature x, and d_{cont} is the number of discrete bins. This method discretizes continuous values into a fixed number of integer bins, preserving their relative distribution. Missing values for all features were encoded with a distinct integer token (-1) to be treated as an informative category by the model. Categorical variables were mapped to unique integer indices. The final tabular feature vector X_i for each visit is a concatenation of categorical variables and the discretized continuous variables.

Image data processing

The imaging pipeline for all chest X-ray images I_i involved several sequential steps to convert raw medical images into a model-ready format. The pipeline began with a rigorous de-identification process where all Protected Health Information was scrubbed from the DICOM metadata of each image. Following de-identification, the DICOM files were parsed and converted into a numerical tensor format suitable for deep learning frameworks. Each image tensor was then resized to a uniform resolution of 518×518 pixels to ensure consistent input dimensions. Finally, the pixel intensities were normalized based on the mean and standard deviation of the whole training dataset, and standard data augmentation techniques, including random rotations, flips, and brightness adjustments, were applied during the training phase to enhance model robustness and prevent overfitting.

Oncoformer architecture

We developed Oncoformer, an encoder-decoder style transformer architecture specifically designed to process longitudinal EHR data. The model comprises three key components: an examination encoder, a temporal embedding, and task-specific decoder heads.

Oncoformer architecture: visit encoder

After preprocessing each patient's longitudinal data, we created a unified multimodal representation for each clinical visit through our multimodal examination encoder. For tabular data, we employed a dual-embedding strategy where discretized categorical and continuous features were processed through separate token and type embeddings. Categorical and continuous features were first discretized and then encoded using shared token embeddings to represent their magnitude, while separate type embeddings were assigned to each feature position to denote the specific clinical variable category. For medical imaging data, when available, images were processed through a pre-trained Vision Transformer[90] backbone using DINOv2[65] self-supervised learning framework to generate dense image feature vectors with dimension matching the Transformer hidden size. The visit-level encoding employs a BERT-like[93] Transformer encoder that processes embedded features through multiple self-attention layers, with a special

classification token prepended to each visit sequence. Missing values are handled through designated padding tokens, enabling the model to differentiate between absent tests and actual clinical observations. The final encoding process generates contextualized visit-level representations $E_{visit} = \text{Encoder}(\text{Embed}(X_i, I_i))$, where Encoder is a Transformer encoder that capture both intra-visit feature dependencies and cross-modal relationships, forming the foundation for subsequent temporal modeling where visit embeddings are enhanced with temporal positional encodings to capture disease progression patterns over time.

Temporal embedding and decoder

To model disease progression and capture the longitudinal nature of patient trajectories, we implemented a temporal embedding to capture the relative time between visits. From the examination encoder output we retrieve a visit level embedding, E_{visit}. To model temporal relationships, we used days elapsed since the initial visit as a linear positional embedding $\text{TimeEmbed}(T)$ to enable the architecture to learn time-dependent patterns in longitudinal EHR data. To create a longitudinal patient-level representation, we passed visit embeddings E_{visit} augmented with time information through a Transformer decoder: $E_{patient} = \text{Decoder}(E_{visit} + \text{TimeEmbed}(T))$, where causal masking ensures unidirectional information flow in this autoregressive process.

Task-specific decoders

Following the established patient-level longitudinal representation $E_{patient}$, we designed a task-specific decoder with separate pathways for discrete outcomes (e.g., diagnosis prediction) and continuous measurements (e.g., biomarker estimation and BA prediction). Each pathway applies a projection layer followed by ReLU activation, formalized as $y_i = ReLU(W_i^T E_{patient,i})$, where $E_{patient,i}$ represents a patient's digital representation derived from 1st to i-th visit. Importantly, causal masking prevents information from future visits from influencing predictions at i-th visit, ensuring fairness by restricting the model to only information available in real clinical scenarios. This architecture also enables simultaneous handling of diverse clinical prediction tasks while facilitating knowledge transfer between related objectives through jointly optimized parameters.

VAE-enhanced multimodal fusion

To achieve more robust cross-modal alignment and representation learning, we extended the multimodal examination encoder with a variational autoencoder (VAE) framework. In this enhanced architecture, dedicated VAE encoders parameterize latent distributions for laboratory test and imaging modalities respectively, where each encoder outputs both mean and log-variance parameters of D-dimensional multivariate normal distributions:$z_{tab} \sim \mathcal{N}(\mu_{tab}, \sigma^2_{tab})$ and $z_{img} \sim \mathcal{N}\left(\mu_{img}, \sigma^2_{img}\right)$. Latent samples are obtained via the reparameterization trick, and cross-modal coherence is enforced through a symmetric KL divergence penalty: $d(Z_{tab}, Z_{img}) = \text{KL}(q(z_{tab})|q(z_{img})) + \text{KL}(q(z_{img})|q(z_{tab}))$. The unified representation $Z_{fuse} = (Z_{tab} + Z_{img})/2$ is decoded back to the transformer embedding space, enabling the model to learn distributional rather than deterministic embeddings while maintaining modality alignment.

Training procedures

Our training procedure consisted of two stages: pretraining and finetuning. During pretraining, we employed self-supervised learning on unlabeled longitudinal EHR data to develop robust clinical representations. The subsequent finetuning stage adapted these representations for specific prediction tasks. This approach leverages generalizable patterns from large-scale unlabeled data before doing specializing downstream applications. Both stages utilized specialized loss functions and incorporated strategies to mitigate dataset-specific biases.

Adversarial missingness/bias control

Missing values in EHRs lead to incomplete or biased digital representations, as models may inadvertently learn to rely on the missing-state biases rather than the true clinical meaning of the feature expression values. Drawing inspiration examples from the concept of adversarial learning in other domains, we implemented a missingness discriminator output head. Concurrently, the missingness discriminator is designed to determine whether a specific feature value is missing or not. We implemented a gradient reversal layer (GRL) between the feature encoder and the missingness discriminator. During backpropagation, the GRL inverts the gradient, compelling the feature encoder to produce representations that are independent of the missingness status. This forces the encoder to focus on encoding the inherent clinical significance of the features rather than being influenced by whether a value is present or absent. The loss function for the missingness discrimination task is defined as: $\mathcal{L}_{missing} = -\frac{1}{M}\sum_{j=1}^{M}\sum_{m=0}^{1} z_{j,m}\, log\, \hat{z}_{j,m}$, where $z_{j,m}$ is a binary indicator denoting whether the feature value of sample$j$ has a missing status m (0 for present, 1 for missing),$\hat{z}_{j,m}$ is the predicted probability, M is the number of samples with potential missing values. By minimizing this loss, the model is encouraged to learn missing-invariant representations. This approach enables the creation of more robust digital representations that can better generalize across datasets with different missing value patterns, ultimately improving the accuracy and reliability of clinical outcome predictions.

Cohort bias, also known as a batch effect, is a significant challenge in multi-center data studies. Clinical data collected from different hospitals often exhibit systematic variations due to differences in patient demographics, practice patterns, and measurement protocols, potentially leading to biased models. Similar to the missingness discriminator, we design a cohort discriminator

that aims to identify the cohort label of each sample, while the encoder is forced to suppress cohort-specific information. The cohort classification loss is formulated as: $\mathcal{L}_{cohort} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} y_{i,d}\log(\widehat{y}_{i,d})$, where $y_{i,d}$ is a binary indicator of whether sample i belongs to domain d, $\widehat{y}_{i,d}$ is the predicted probability, N is the number of samples, and D is the number of domains (clinical cohorts). This approach encourages the model to learn cohort-invariant representations that generalize across healthcare settings while maintaining predictive performance for clinical outcomes.

Pretraining step

We employed a self-supervised pretraining approach with multiple complementary objectives to enable our model to learn comprehensive representations of EHR data. We randomly masked 50% of the valid test results in the current examination and trained the model to predict these masked values in both the current and the subsequent examinations. The masked language modeling (MLM) loss function is defined as: $\mathcal{L}_{MLM} = \frac{1}{|\mathcal{M}|}\sum_{i=0}^{N}\sum_{j\in\mathcal{M}_i}\mathcal{L}_{MSE}(\widehat{v}_{i,j}, v_{i,j})$, where $\mathcal{M}_i$ is the set of masked indices in examination event s_i and next examination after s_i, $|\mathcal{M}|$is the total number of masked tokens across all examination events, $v_{i,j}$ is the true value of the j-th test in examination s_i and next examination after s_i, and $\widehat{v_{i,j}}$ is the predicted value. The VAE regularization term is defined as: $\mathcal{L}_{\text{VAE}} = \lambda_1\sum_{m\in\{tab,img\}} KL(q(z_m)|\mathcal{N}(0,I)) + \lambda_2\, d(Z_{tab}, Z_{img})$, where the hyperparameters λ_1 and λ_2 control the relative importance of latent regularization and cross-modal alignment, respectively, balancing reconstruction fidelity against representational coherence in the learned latent space. Additionally, we incorporated the domain adversarial loss $\mathcal{L}_{domain}$ and $\mathcal{L}_{missing}$to promote cohort-invariant and missing-invariant representations. Therefore, the final pretraining objective combined these components with appropriate weighting coefficients: $\mathcal{L}_{pretrain} = \alpha_1\mathcal{L}_{MLM} + \alpha_2\mathcal{L}_{\text{VAE}} - \alpha_3\mathcal{L}_{cohort} - \alpha_4\mathcal{L}_{missing}$where the negative sign reflects the gradient reversal mechanism.

Disease prediction fine-tuning

We implemented two distinct disease prediction tasks that reflect different clinical scenarios: first occurrence disease diagnosis and future disease prediction. For first occurrence disease diagnosis, we trained the model to identify the first occurrence of specific diseases, excluding subsequent visits after initial diagnosis to capture true onset patterns rather than disease management. Formally, for a patient with longitudinal sequence S with a length of L, and $l_{i,d}$ represents whether this patient was diagnosed as positive for disease d at i-th visit, the label $c_{i,d}$ of first occurrence diagnosis task is defined as:

$$c_{i,d} = \begin{cases} 1, \textit{if } l_{i,d} = 1 \textit{ and } l_{j,d} = 0 \textit{ for all } j < i \\ 0, \textit{if } l_{i,d} = 0 \textit{ and } l_{j,d} = 0 \textit{ for all } j \in \{0, 1, \ldots, L\}. \end{cases}$$

Therefore, the loss function is: $\mathcal{L}_{first} = \frac{1}{N}\sum_{i=0}^{N}\sum_{d=1}^{D}\mathcal{L}_{BCE}(\widehat{c}_{i,d}, c_{i,d})$, where $\widehat{c}_{i,d}$ is the predicted probability of disease d, D is the total number of diseases considered, and $\mathcal{L}_{BCE}$ is the binary cross-entropy loss.

To explicitly mitigate the learning bias caused by the significant class imbalance between patients with and without cancer, we implemented a class-weighted strategy for the binary cross-entropy loss ($\mathcal{L}_{BCE}$). By assigning a proportionally higher penalty weight to the minority positive class (cancer cases) during the fine-tuning stage, we prevented the model from defaulting to majority-class predictions, ensuring robust sensitivity without compromising specificity.

For the future disease prediction task, we developed a more nuanced labeling strategy to identify patients at risk before disease manifestation. Formally, for a patient with longitudinal sequence S with a length of L, the label $f_{i,d}$ of future prediction task for disease d at i-th visit is defined as:

$$f_{i,d} = \begin{cases} 1, \textit{if } l_{j,d} = 0 \textit{ for } j \le i \textit{ and } \exists k > i \textit{ such that } l_{k,d} = 1 \\ 0, \textit{if } l_{i,d} = 0 \textit{ and } l_{j,d} = 0 \textit{ for all } j \in \{0, 1, \ldots, L\}. \end{cases}$$

The future state prediction loss is: $\mathcal{L}_{future} = \frac{1}{N}\sum_{i=0}^{N}\sum_{d=1}^{D}\mathcal{L}_{BCE}(\widehat{f}_{i,d}, f_{i,d})$, where $\widehat{f}_{i,d}$ is the predicted probability of future development of disease d based on the i-th examination event (X_i, I_i, T_i). The total loss for disease prediction tasks is $\mathcal{L}_{disease} = \mathcal{L}_{first} + \mathcal{L}_{future}$, balancing the model's ability to recognize both existing conditions and identify patients at risk for future disease development.

Treatment Response Prediction Framework

The prediction of treatment response was framed as a supervised regression task. The goal was to predict the post-treatment trajectory of a patient's disease marker (such as tumor size or serum biomarker levels) given a specific therapy A, using only data available prior to treatment initiation.

To achieve this, Oncoformer first generates a comprehensive, pre-treatment patient representation, Z_i, from their entire historical data. This high-dimensional embedding Z_i serves as a rich summary of the patient's baseline state, capturing clinical, laboratory, and imaging features accumulated over the longitudinal observation period. A task-specific regression head is then trained to model the conditional outcome, $P(Y \mid Z_i, A)$, where Y is the post-treatment trajectory and A denotes the administered therapy.

A key assumption underlying this approach is that the pre-treatment representation Z_i captures a sufficiently comprehensive set of patient characteristics to account for systematic differences in treatment assignment across patient subgroups. While this assumption cannot be formally verified with observational data, the depth and breadth of the longitudinal, multimodal data integrated by

Oncoformer are intended to create a representation that minimizes the influence of unmeasured patient-level factors on the predicted treatment trajectories.

Implementation details

We implemented our Oncoformer architecture using a combination of transformer models. Specifically, we utilized a 24-layer transformer encoder with a hidden dimension of 1024 as the examination encoder to process individual examination events, and a 12-layer auto-regressive transformer decoder with a hidden dimension of 768 as the temporal encoder to capture longitudinal patterns across the sequence of examinations. This design leverages the attention capabilities of the multi-headed self-attention mechanism for understanding relationships between clinical measurements within each examination while employing the causal masked attention mechanism to model the temporal progression of patient health.

The model was implemented using PyTorch and trained using a two-stage approach. For the pretraining phase, we trained the model for 200 epochs using the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-6}. The subsequent fine-tuning phase for the downstream tasks was conducted for 100 epochs using the Adam optimizer with a reduced learning rate of 10^{-4} while maintaining the same weight decay of 10^{-6}.

The Oncoformer model was developed and validated using a rigorous, multi-stage process involving distinct, non-overlapping patient sets, as outlined in Figure S1. The entire development process, including pre-training and fine-tuning, was conducted exclusively on the COMPASS-Main cohort, which was randomly split at the patient-level in an 8:1:1 ratio to generate a training set (n=2,248,594) for model parameter training, a tuning set (n=281,074) for hyperparameter optimization and model selection, and an internal validation set (n=281,074) for an initial, unbiased performance assessment. Subsequently, the final, optimized model was subjected to two rounds of stringent external validation on completely independent cohorts: first on the COMPASS-Replication cohort (n=862,247) to test for cross-institution generalizability, and second on the constructed case-control cohort from the UK Biobank (n= 502,665) to test for cross-population generalizability. Crucially, all reported performance metrics in this study, unless otherwise specified, are from these designated validation sets, which were never seen by the model during any training or tuning phase.

Study Endpoints and Task Formulation

Cancer Diagnosis and Prediction

The primary tasks involved both the concurrent diagnosis and future prediction of pan-cancer incidence. Cancer cases were identified through electronic pathology records confirming a malignant neoplasm, cross-referenced with a corresponding first-ever ICD-10 diagnosis code (C00-C97), which defined the index date of diagnosis. All cancer diagnoses and tumor classifications in this study are based on the primary site of the malignancy, as defined by the corresponding ICD-10 codes. Metastatic or secondary tumor deposits were not classified as separate cancer cases. To create a robust and clinically relevant control group, control subjects were selected from individuals with no history of any malignant neoplasm diagnosis throughout their entire observation period, matched to the case population on baseline age and sex. This control cohort naturally included subjects with a variety of comorbidities and known cancer risk factors, ensuring that the model learns to distinguish true cancer signals from those of other common health conditions.

For each clinical visit, a diagnosis label was assigned as 1 only for visits occurring on the index date of diagnosis. The prediction label was assigned as 1 for pre-diagnostic visits of a future cancer patient. All visits for control subjects were labeled as 0.

Importantly, we deliberately maintained a natural class imbalance in our cohorts. This distribution closely aligns with the broad population epidemiology observed in our external UK Biobank cohort. Preserving this natural ratio, rather than artificially balancing the dataset, is critical for simulating real-world screening scenarios. It ensures that predictive metrics reliably reflect true clinical utility and avoid the severe overestimation of performance often seen in artificially balanced models. This distribution closely aligns with the cancer prevalence observed in our independent UK Biobank validation cohort (8.8%), confirming that our training data reflects real-world epidemiology.

Predictive leakage mitigation

To prevent information leakage from the final diagnostic work-up, a strict "washout" period was enforced for all future prediction tasks. Specifically, all clinical data recorded within the 90 days immediately preceding the index date of diagnosis were excluded from the input for the predictive models. Therefore, the "1-year prediction" task, for example, used data from 365 days to 91 days before diagnosis to predict a cancer that would be formally diagnosed within the next year. This ensures that the model's predictive performance relies on long-term, subtle signals rather than overt, late-stage diagnostic clues.

To empirically validate this washout design, we systematically extended the data exclusion window from the index date (0 days) up to 36 months prior to diagnosis across all cancer types (Figure S8). The performance trajectories show a consistent decline between 0 and 3 months, reflecting the clinical diagnostic workup process during which disease-directed tests are progressively ordered before formal diagnosis. This pattern confirms that the 90-day cutoff effectively excludes data influenced by the diagnostic evaluation. Beyond the 3-month mark, the curves transition into a gradual decline, maintaining predictive accuracy above 0.85 for most cancer types at 12 months, demonstrating that the model's sustained performance relies on long-term physiological signals independent of acute diagnostic clues. Using the same 90-day washout definition, subgroup-level discrimination and

calibration were further estimated by sex, age, and ethnicity in COMPASS-Replication and the UK Biobank where sample size permitted (Figure S9).

Identification of patients with utilization spikes

For each patient we computed the monthly rate of clinical encounters, laboratory-test orders, and imaging studies. A within-patient baseline rate was defined as the mean monthly rate over 12–36 months before the index (diagnosis) date, and a recent rate as the mean monthly rate over the 3–12 month pre-index window. A patient was flagged as exhibiting a utilization spike if the encounter rate, the laboratory-order rate, or the imaging rate in the recent window was $\geq$2-fold higher than the corresponding within-patient baseline. Because imaging is comparatively sparse, patients with no imaging during the baseline period but $\geq$1 imaging study during the recent window were also flagged. Patients with <12 months of pre-index history before the recent window were excluded, as a baseline rate could not be reliably estimated. The full Oncoformer was then re-evaluated on the remaining (non-spike) patients; robustness was confirmed with 1.5$\times$ and 3$\times$ thresholds. The results of these healthcare-utilization sensitivity analyses are summarized in Table S8.

Unlike the standalone single-modality encoders used for benchmarking in Table S3, the reduced-modality models in Table S4 were retrained within the ablation framework. Furthermore, to assess the model's reliance on specific, highly indicative tests, we conducted a key feature ablation study. We retrained and evaluated the predictive model after completely removing a panel of common serum tumor markers (including PSA, CEA, CA125, et al.) from the input data. The sustained high performance in this ablation analysis confirms that Oncoformer can capture robust cancer signals from a wide array of non-specific laboratory and imaging data, rather than depending on a few highly specific markers ordered during a diagnostic work-up.

Tumor Stage Inference

This task aimed to predict the clinical stage of a tumor at its future time of diagnosis, using only pre-diagnostic data. The gold-standard staging information, including Overall Stage (I-IV) and individual TNM components, was extracted from comprehensive surgical pathology reports and initial clinical staging workups documented at the time of first diagnosis. This was formulated as a multi-class classification problem where the model, given a pre-diagnostic visit, was trained to predict the stage that would be formally assigned at the future point of diagnosis. Crucially, these pathology-derived labels were used only as prediction targets and were never provided to the model as input; the model received exclusively routine pre-diagnostic laboratory tests and chest X-rays.

Pathological staging validation

To validate the clinical relevance of Oncoformer's pre-diagnostic staging predictions, we conducted a retrospective analysis using deeply phenotyped sub-cohorts with matched postoperative pathology reports. Gold-standard pathological variables, including maximum tumor diameter, pathological invasion depth (pT1-pT4), number of positive lymph nodes, Lymph Node Ratio (LNR), lymphovascular invasion (LVI) status, organ-specific metastasis status (M0 vs. M1), and histological grading (Gleason score for prostate cancer, Nottingham grade for breast cancer, differentiation grade for gastrointestinal cancers), were extracted from surgical pathology reports. Spearman rank correlations were computed between Oncoformer's pre-diagnostic staging scores and each pathological variable. Group comparisons for categorical pathological variables were assessed using the Kruskal-Wallis test (for ordered categories) or the Mann-Whitney U test (for binary comparisons such as LVI status and metastasis status).

Disease Progression Trajectory Analysis

To visualize cancer progression trajectories within the learned latent space, we adapted the RNA velocity framework from single-cell genomics. Each patient's sequence of clinical visits was treated as a temporal series of latent-space embeddings. For each consecutive pair of visits, a velocity vector was computed as the difference in embedding coordinates, representing the directionality and magnitude of change in the patient's health state. These velocity vectors were then aggregated and projected onto the two-dimensional UMAP space using a kernel-based smoothing approach, generating streamline plots that depict the inferred progression pathways from early-stage-associated regions toward late-stage regions of the latent space.

Treatment Response Prediction

This task evaluated the model's ability to predict patient-specific responses to various anti-cancer therapies. To construct the cohorts for this task, we first identified the precise start date of a specific line of therapy for each patient from medication administration records and physician's orders. The "pre-treatment" data window was then rigorously defined as all available clinical data (EHR and imaging) for that patient from their first recorded visit up to and including the day immediately prior to the therapy initiation date. Any data recorded on or after the start date of the specified therapy was strictly excluded from the model's input to prevent information leakage.

A binary response label ("Sensitive" vs. "Resistant") was created based on the first formal treatment assessment. For solid tumors, this was determined by the first post-treatment radiological evaluation (typically conducted 8–12 weeks after therapy initiation) according to RECIST 1.1 criteria: "Sensitive" was defined as a Complete or Partial Response (CR/PR), while "Resistant" included Stable or Progressive Disease (SD/PD). For prostate cancer treated with ADT, sensitivity was defined by a PSA decline of $\geq$50% from the pre-treatment baseline within the first 6 months of therapy. For quantitative prediction of response curves, the model's regression head was trained to forecast the entire time-series of post-intervention measurements (e.g., percentage change in tumor size from baseline) using only the strictly defined pre-treatment data.

Recurrence-free survival analysis

This task assessed the prognostic power of the Oncoformer-derived representation. The recurrence-free survival analysis was restricted to patients who received first-line treatment. For each such patient, a continuous prognostic risk score was generated using only the clinical data available at the treatment-naive baseline, that is, before the start of first-line therapy. Patients were then

stratified into "High risk" and "Low risk" groups based on the median score. Recurrence-free survival was defined as the time from the initiation of first-line therapy to the first documented event, which included disease recurrence (local or distant) or death from any cause. Patients who were alive and recurrence-free were censored at the date of their last known follow-up or the study's data cutoff date. Because the temporal decoder uses causal masking, this risk score can also be recomputed at any subsequent visit using only data available up to that point, enabling longitudinal updating during follow-up.

Feature Importance Analysis

To interpret the clinical features driving Oncoformer's predictions, we applied a SHAP (SHapley Additive exPlanations) analysis to quantify each input feature's contribution to the model output. SHAP values were computed on the COMPASS-Replication cohort for six predictive tasks: pan-cancer diagnosis, staging, metastasis prediction, EGFR treatment response, immunotherapy response, and pan-cancer recurrence (Figure S7). For each task, the top 15 features were ranked and visualized according to their mean absolute SHAP values, highlighting the clinical variables with the greatest overall contribution to model prediction.

To construct the cancer-type-specific staging biomarker network (Figure S6), a SHAP analysis was performed separately for each of the ten cancer types in the staging prediction task. For each cancer type, the top 20 features ranked by a mean absolute SHAP value were extracted. The resulting network was visualized as a bipartite graph in which large colored nodes represent the ten cancer types, with the node diameter scaled proportionally to cohort sample size, and small peripheral nodes represent individual clinical biomarkers, with node size reflecting aggregate feature importance (sum of mean absolute SHAP values across all connected cancer types). Edges connect each biomarker to the cancer type(s) for which it ranks among the top 20 staging predictors, with edge width proportional to the mean absolute SHAP value of that specific feature-cancer pair. Biomarkers connected to multiple cancer types were positioned within the inner ring of the network to highlight shared systemic signatures, whereas biomarkers predominantly associated with a single malignancy were distributed in the outer ring.

Pathway-level Molecular Validation

To assess whether Oncoformer's staging inference scores correspond to established cancer genomic alterations, we conducted a pathway-level concordance analysis. The genomic profiling used for this analysis was obtained from an independent dataset and was never provided to Oncoformer as input; the staging inference scores were generated solely from routine pre-diagnostic laboratory tests and chest X-rays. The analytical framework was inspired by the hierarchical visible neural network approach developed by Park et al.,[66] which demonstrated that mapping genetic alterations onto structured protein assembly hierarchies can reveal convergent biological mechanisms underlying therapeutic response. The rationale for aggregating mutation information at the pathway level rather than analyzing individual genes in isolation is further supported by Kong et al.,[67] who demonstrated that assembly-level somatic mutation burden provides a biologically meaningful representation for predicting immunotherapy response. Approximately 520 cancer-related genes from a targeted sequencing panel were organized into a four-level hierarchy: individual genes, functional sub-pathways, major cancer pathways, and three overarching biological processes (Proliferative Signaling, Genome Stability, and Regulation & Others). For each pathway, the Oncoformer-Genomic Concordance (OGC) was defined as the Spearman rank correlation between the Oncoformer staging inference score and the pathway-level somatic mutation burden (defined as the fraction of genes within the pathway harboring non-synonymous mutations). Pathways with OGC $\geq$ 0.7 and false discovery rate (FDR) $<$ 0.01 (Benjamini-Hochberg correction) were designated as Core pathways. Pathways with OGC between 0.5 and 0.7 were classified as showing moderate concordance. Gene-level protein association networks within Core pathways were constructed using known and predicted protein-protein interaction data from the STRING database (version 12.0; https://string-db.org).[88]

UK Biobank generalizability

To evaluate the cross-population generalizability of our framework, we used the UK Biobank (UKB) cohort as a fully independent external validation set (demographic characteristics provided in Table S5). The UKB, a large-scale prospective study of over 500,000 participants, provided the basis from which we constructed a case-control cohort by identifying all participants with a first-ever cancer diagnosis (ICD-10 coded) and matching them with non-cancer controls based on age, sex, and follow-up duration. To quantify the value of our pre-trained model on this cohort, we established a head-to-head comparison between three modeling approaches, with results presented in Table S6: (i) a UKB-specific baseline model trained from scratch using only available UKB data (demographics and standard laboratory tests); (ii) the Oncoformer zero-shot model, where the model trained exclusively on the COMPASS cohort was directly applied to the UKB data without any parameter updates or exposure to UKB outcomes, to evaluate intrinsic cross-population generalizability; and (iii) the Oncoformer fine-tuned model, where our pre-trained model was used as a feature extractor to generate representations upon which a new, lightweight prediction head was fine-tuned on a training partition of the UKB data. This fine-tuning regimen on the UK Biobank explicitly functions as 'further training' to adapt the model to a demographically, ethnically (predominantly White European), and environmentally distinct background, thereby rigorously stress-testing and improving the model's global robustness and cross-population generalizability.

This experimental design allowed us to directly assess whether the knowledge learned by Oncoformer from the COMPASS cohorts could be effectively transferred to improve predictive performance in a distinct healthcare and population context. For the cancer diagnosis task, we defined positive cases as individuals who had a confirmed cancer diagnosis recorded in hospital episode statistics, cancer registries, or primary care records either before the baseline assessment visit (2006–2010) or within the first 6 months after the baseline assessment. These cases represent the prevalent cancer population at baseline and early post-baseline period. For the cancer prediction task, we defined positive cases as individuals who developed cancer during the 5-year follow-up period,

with the first cancer diagnosis occurring between 6 months and 5 years after the baseline assessment visit. We excluded incident cases recorded within the first 6 months of follow-up to minimize the potential for reverse causality, where undiagnosed cancer at baseline might influence the predictor variables. We defined controls as individuals who did not develop any cancer. Cancer definitions integrated data from hospital episode statistics, cancer registries, death certificates, and primary care records where available.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were conducted using Python (version 3.10) and its associated scientific computing libraries, including PyTorch (version 2.60) for model development, scikit-learn (version 1.2) for performance metrics, stats models for statistical tests, and lifelines (version 0.27) for survival analysis. A two-sided P-value of less than 0.05 was considered statistically significant across all analyses. The performance of binary classification tasks—including cancer diagnosis, future risk prediction, and individual components of tumor staging—was primarily evaluated by the Area Under the Receiver Operating Characteristic curve (AUROC). We also reported the F1-score, the harmonic mean of precision and recall, as a supplementary metric robust to class imbalance. 95% CIs for AUROC estimates were calculated using DeLong standard errors, unless otherwise specified. For quantitative regression tasks, such as predicting the continuous trajectory of treatment response (e.g., tumor size change) or future biomarker levels, we used the coefficient of determination (R^2) to assess the goodness-of-fit and the proportion of variance explained by the model. Prognostic analyses for recurrence-free survival (RFS) were performed using standard survival analysis techniques. Kaplan-Meier curves were generated to visualize survival probabilities over time for different risk groups. The statistical significance of the difference between these curves was assessed using the log-rank test. Hazard ratios (HRs) and their corresponding 95% CIs were calculated using a multivariable Cox proportional hazards model, which was adjusted for potential confounding variables including patient age and sex at the time of diagnosis. The proportional hazards assumption was verified for all Cox models. Model calibration was assessed using the Brier score and Expected Calibration Error (ECE), computed on the COMPASS-Replication cohort by comparing predicted cancer probabilities against observed cancer frequencies across ten equal-width probability bins. Clinical utility was evaluated using Decision Curve Analysis (DCA), which quantifies the net benefit of the model across a range of threshold probabilities compared to default strategies of screening all or no patients.