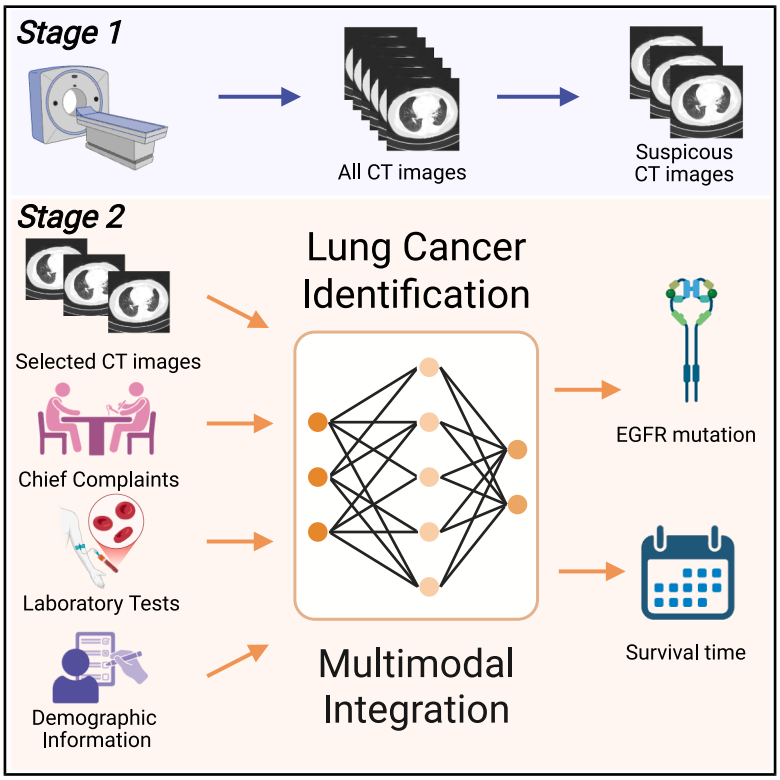
Cell Reports
Medicine

Article

AI-enabled molecular phenotyping and prognostic predictions in lung cancer through multimodal clinical information integration

Graphical abstract



Highlights

- LUCID integrates multimodal data to predict EGFR mutations and survival outcomes
- Evaluation on 5,175 samples achieves AUCs of 0.881 for EGFR mutation prediction
- LUCID predicts patients' survival outcome with AUCs up to 0.912
- External validation on 1,285 samples confirms the robustness of LUCID

Authors

Yuxing Lu, Fei Liu, Yunfang Yu, ..., Kang Zhang, Jinzhuo Wang, Xiaoying Huang

Correspondence

chengdi_wang@scu.edu.cn (C.W.), kang.zhang@gmail.com (K.Z.), wangjinzhuo@pku.edu.cn (J.W.), huangxiaoying@wzhospital.cn (X.H.)

In brief

Lu et al. present LUCID, a multimodal AI framework that integrates multimodal data to non-invasively predict *EGFR* mutations and survival outcomes in lung cancer patients. LUCID achieves high performance and robustness across 5,715 patients and 1,285 external validations, providing a scalable tool for precision oncology and personalized treatment planning.

Lu et al., 2025, Cell Reports Medicine 6, 102216
July 15, 2025
https://doi.org/10.1016/j.xcrm.2025.102216

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

AI-enabled molecular phenotyping and prognostic predictions in lung cancer through multimodal clinical information integration

Yuxing Lu,[1,2,10] Fei Liu,[2,3,4,10] Yunfang Yu,[4,5,10] Bojiang Chen,[6,10] Wenyao Yu,[2,7] Zixing Zou,[5] Kun Li,[2,5] Miao Man,[4] Caiwen Ou,[8] Chengdi Wang,[6,*] Kang Zhang,[2,4,5,*] Jinzhuo Wang,[1,*] and Xiaoying Huang[9,11,*]

[1]Department of Big Data and Biomedical AI, College of Future Technology, Peking University, Beijing 100871, China
[2]State Key Laboratory of Eye Health, Eye Hospital and Institute for Advanced Study on Eye Health and Diseases, Wenzhou Medical University, Wenzhou, China
[3]National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[4]Institute for AI in Medicine and Faculty of Medicine, Macau University of Technology, Macau, China
[5]Guangzhou National Laboratory, Guangzhou 510005, China
[6]Precision Medical Center, Precision Medicine Research Center, Department of Respiratory and Critical Care Medicine, West China Hospital of Sichuan University, Chengdu, China
[7]Department of Mathematics, University of California, San Diego, San Diego, CA, USA
[8]The Tenth Affiliated Hospital (Dongguan People's Hospital) Southern Medical University, Southern Medical University, Dongguan 523059, China
[9]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 325000, China
[10]These authors contributed equally
[11]Lead contact
*Correspondence: chengdi_wang@scu.edu.cn (C.W.), kang.zhang@gmail.com (K.Z.), wangjinzhuo@pku.edu.cn (J.W.), huangxiaoying@wzhospital.cn (X.H.)
https://doi.org/10.1016/j.xcrm.2025.102216

SUMMARY

Lung cancer remains the leading cause of cancer-related mortality worldwide. The need for cost-effective, non-invasive methods to detect specific gene mutations for targeted therapy and predict patient survival outcomes underscores the importance of advancing diagnostic and prognostic capabilities. Contemporary lung cancer diagnostic models often fail to integrate diverse patient data, leading to incomplete clinical assessments. To address these challenges, we propose LUCID, a multimodal data integration framework designed to predict epidermal growth factor receptor (EGFR) mutation status and survival outcomes in patients with lung cancer. Tailored for early-stage clinical assessment, LUCID leverages lung computed tomography (CT) images, chief complaints, laboratory test results, and demographic data to deliver comprehensive, non-invasive predictions. LUCID achieved strong performance in a retrospective cohort of 5,175 patients, with areas under the receiver operating characteristic curve (AUCs) ranging from 0.851 to 0.881 for EGFR mutation prediction and from 0.821 to 0.912 for survival time prediction. The model also demonstrated robustness across external validation cohorts and in scenarios with missing modalities.

INTRODUCTION

The global burden of lung cancer continues to present significant challenges in healthcare, remaining the leading cause of cancer-related mortality worldwide.[1,2] Recent advancements in artificial intelligence (AI) have opened the unprecedented opportunities for improving lung cancer diagnosis, treatment planning, and outcome prediction.[3–8] The integration of AI-driven solutions with clinical practice has demonstrated particular promise in enhancing diagnostic accuracy and treatment efficacy, marking a paradigm shift in oncological care.[9–13]

In lung cancer management, AI has achieved significant technological breakthroughs across multiple domains. Deep learning algorithms have demonstrated remarkable success in analyzing medical imaging data, including X-ray, computed tomography (CT) scans, positron emission tomography, and magnetic resonance imaging.[14–17] These algorithms can detect subtle patterns and features that might escape human observation, enabling earlier and more accurate diagnosis. Additionally, AI approaches have shown exceptional capability in processing complex molecular data, including genomic profiles and biomarker information, leading to more precise patient stratification and personalized treatment strategies.[18–20]

The inherent multimodal nature of medical data presents both challenges and opportunities in lung cancer care. By analyzing diverse data types simultaneously, multimodal AI approaches provide a more holistic understanding of the disease. The increasing availability of high-quality paired datasets has accelerated the development and application of sophisticated multimodal fusion techniques.[21–24] These multimodal models enhance the reliability and accuracy of AI systems across various applications, from initial diagnosis to treatment planning and prognosis prediction.[7] They also enable clinicians to make more informed decisions based on comprehensive, personalized recommendations that consider multiple aspects of patient data.[25,26]

Epidermal growth factor receptor (*EGFR*) gene mutation is a key driver of tumorigenesis and progression in certain types of non-small cell lung cancer.[27–30] These mutations lead to the overactivation of the *EGFR* signaling pathway, promoting uncontrolled tumor cell proliferation, survival, and metastasis. Specific *EGFR* mutations, such as exon 19 deletions and *L858R* mutations in exon 21, are well-established therapeutic targets for *EGFR* tyrosine kinase inhibitors (TKIs), which directly inhibit the aberrant *EGFR* activity. *EGFR*-TKIs have demonstrated significant clinical efficacy in patients harboring these mutations, offering improved outcomes and fewer side effects compared to traditional chemotherapy.

Beyond mutation identification, accurate survival time prediction remains a crucial component in optimizing lung cancer treatment strategies. This prognostic information enables clinicians to develop more tailored treatment approaches while providing essential insights to patients and their families for future planning.[31] The prediction process integrates multiple clinical parameters, including tumor characteristics, disease staging, patient demographics, and overall health status, all of which significantly influence survival outcomes.[32,33] Moreover, robust prediction models contribute to more efficient resource allocation, helping optimize interventions to enhance both quality of life and treatment effectiveness.

To address these critical needs in lung cancer care, we introduce LUCID, a novel 2-stage multimodal integration model designed to predict both *EGFR* mutation types and patient survival times with enhanced accuracy. Our model leverages a comprehensive dataset comprising 5,175 patients, incorporating multiple data modalities including CT images, patient-reported symptoms, laboratory findings, and demographic information. LUCID's integrated approach not only demonstrates superior predictive performance compared to conventional models but also highlights AI's transformative potential in healthcare through its ability to facilitate more personalized treatment strategies.

The key contribution of our study lies in its multimodal approach, which leverages a custom-designed transformer framework to analyze four distinct data modalities for predicting both *EGFR* mutation types and survival times in lung cancer patients. Extensive experimental validation demonstrates LUCID's superior performance compared to traditional single-modality approaches. The model achieves areas under the receiver operating characteristic curve (AUC) scores of 0.851–0.881 for *EGFR* mutation prediction and 0.821–0.912 for survival time prediction. External validation on an independent dataset also shows robust performance. Notably, while LUCID is trained on paired multimodal data, it maintains robust performance even in scenarios with single-modality input or missing modalities. These results represent a significant advancement in the field of AI-driven lung cancer diagnostics and prognostics.

RESULTS

Patient characteristics

In this study, we utilized a well-established multimodal lung cancer dataset from West China Hospital. This dataset comprises 8,162 lung cancer patients1 since 2009. After applying the eligibility criteria, 5,175 cases were enrolled for training and internal validation of our study. For external validation, we used a separate dataset of 1,285 patients from Sun Yat-sen Memorial Hospital. Male patients constituted 48.95% of the training and internal validation dataset, with the percentage in the external validation dataset being 57.51%. Each patient underwent chest CT scans, and we extracted detailed demographic metadata from their electronic health records, including demographic information, cancer subtypes, staging details, and textual chief complaints. Blood samples were collected after fasting, followed by laboratory tests and medical follow-ups. Additionally, we conducted tests to determine the *EGFR* mutation types of the patients. The clinical characteristics of the patients from each dataset used in our model development and validation are shown in Tables 1 and 2.

A two-stage deep learning framework for multimodal lung cancer diagnosis

Here we introduce LUCID, a multimodal integration model for lung cancer identification that includes *EGFR* gene mutations and patients' expected survival time. An overall illustration of LUCID is shown in Figure 1. Unlike conventional end-to-end models, LUCID operates in two distinct stages.

In the first stage (Figure 1A, see STAR Methods for details), we implemented a Vision Transformer (ViT)-based classification model[34] that was pretrained on the ImageNet-1000 dataset.[35] We fine-tuned this model using an independent dataset from West China Hospital, which includes all CT scans from 56 patients, totaling 4,099 lung CT images. Among these, 447 CT images were annotated with lesions by expert radiologists. The ViT model generates representional vectors for each CT image and classified them via a multilayer perceptron model,[36] distinguishing between normal CT images and those with potential lesions. These selected images are then processed in the second stage for further multimodal analysis and prediction.

In the second stage (Figure 1B, see STAR Methods for details), we employed a multimodal fusion approach using a specially designed transformer[17,37] architecture that incorporates both cross-attention and joint-attention modules to encode different data modalities hierarchically (Figure 3). Each CT image was divided into 16 $\times$ 16 pixel patches, and a convolutional neural network (CNN)[38] was used for feature extraction. To encode the textual chief complaints associated with the CT scans, we used the BERT[39] autoencoder large language model (LLM). The complete blood count, consisting of 92 indicators, was also encoded for integration, along with the patients' demographic information. The multimodal interaction attention mechanism in LUCID, through both the cross-attention and

Table 1. Clinical characteristics of the internal training and validation cohort, differentiated by *EGFR* types (*EGFR*-wild type and *EGFR*-sensitive type)

	EGFR-wild type	*EGFR*-sensitive type
Subjects (*n*)	2,047	3,128
Age (years)	58.4 (±11.28)	58.76 (±10.34)
Sex		
Male	670 (32.73)	1,863 (59.56)
Female	1,377 (67.27)	1,265 (40.44)
Smoking history		
Yes	1,219 (59.55)	846 (27.05)
No	828 (40.45)	2,282 (72.95)
Stage of disease		
Stage I	519 (25.35)	1,189 (38.01)
Stage II	218 (10.65)	256 (8.18)
Stage III	596 (29.12)	519 (16.59)
Stage IV	714 (34.88)	1,164 (37.21)
Pathology type		
ADENO-CA	1,353 (66.10)	2,970 (94.95)
SQ-CA	550 (26.87)	71 (2.27)
Others	144 (7.03)	87 (2.78)
EGFR mutations		
Common mutation	–	2,903 (92.80)
Rare mutation	–	160 (5.13)
20-INS	–	63 (2.01)
T790M	–	2 (0.06)

Data include age, gender, smoking history, stage of disease at diagnosis, pathological type, and *EGFR* mutation types. Values are expressed as the mean ± standard deviation or as counts (with percentages) as appropriate. *EGFR* stands for epidermal growth factor receptor. ADENO-CA and SQ-CA refer to adenocarcinoma and squamous cell carcinoma, respectively.

Data are presented as the mean ± SD, or *n* (%), unless otherwise stated. *EGFR*, epidermal growth factor receptor.

Common Mutations include *Exon 19* deletions and *L858R* mutations in *Exon 21*, which are the most frequent *EGFR* mutations sensitive to *EGFR*-TKI therapies.

Rare mutations refer to less common *EGFR* mutations, which include *G719X*, *S768I*, *L861Q*, and other rare mutations.

Table 2. Clinical characteristics of the external validation cohort

	EGFR-wild type	*EGFR*-mutant type
Subjects (*n*)	551	734
Age (years)	58.68 (±12.06)	58.96 (±10.81)
Sex		
Male	341 (61.89)	398 (54.22)
Female	210 (38.11)	336 (45.78)
Stage of disease		
Stage I	119 (21.60)	199 (27.11)
Stage II	142 (25.77)	227 (30.93)
Stage III	123 (22.32)	107 (14.58)
Stage IV	167 (30.31)	201 (27.38)

Values are expressed as the mean ± standard deviation or as counts (with percentages) as appropriate.

Data are presented as the mean ± SD, or *n* (%), unless otherwise stated. *EGFR*, epidermal growth factor receptor.

joint-attention modules, enabled the comprehensive fusion of CT images, textual chief complaints, and laboratory test results, capturing shared and complementary information. The fused multimodal representations were then combined with patients' demographic information to perform subsequent classification and prediction tasks.

In this study, we employed the LUCID stage-1 model to filter suspicious CT images from CT scans and use the LUCID stage-2 model to predict *EGFR* mutation types and survival outcomes in lung cancer patients. The model was trained and tested on the West China Hospital dataset and externally validated using the Sun Yat-sen Memorial Hospital dataset.

Binary classification model helps filter CT images

Consistent with the methodologies employed by healthcare professionals, the first step in analyzing a patient's CT images involves selecting those CT images that may contain abnormal patterns or lesions from the entire image set of CT scan. This crucial task, while essential, is often time consuming and demands considerable effort from radiologists. In our study, we utilized the LUCID stage-1 model to efficiently filter these suspicious CT images.

We conducted a 5-fold cross validation on the annotated CT image dataset from West China Hospital and plotted the AUC curve for each fold, as shown in Figure 2B. LUCID's stage-1 model demonstrates impressive classification performance. The model exhibited strong and consistent performance across all five validation folds, achieving an AUC of 0.957 (95% confidence interval [CI], 0.933–0.981) and an F1 score of 0.878 (95% CI, 0.842–0.914). This indicates its effectiveness in identifying the majority of lung images with lesions. Empirically, we took out all lung CT images of a single patient (ID: 0019770975) who presented a lesion in the lower-left region of the lungs (Figure 2C). As highlighted by the red border, the stage-1 model efficiently identified most of the CT images displaying lower-left corner lesions, demonstrating its practical application in real-world scenarios.

In addition, we witnessed a scaling law in the stage-1 ViT classification models in LUCID. As illustrated in Figure 2A, we compared the accuracy, AUC, and F1 metrics of ViT models with different parameter sizes (86 M for ViT-b, 307 M for ViT-l, and 632 M for ViT-h). The results confirmed that ViT models with larger parameters yield superior image classification results, aligning with findings observed in the general domain.[40] We further implemented the ViT-h-based stage-1 model on the whole lung cancer dataset for unlabeled CT images selection. The model selected 24,773 highly suspicious CT images from 375,660 lung CT images of 5,175 patients (6.5%), averaging 4.78 CT images per patient in the West China Hospital dataset, and 6,228 highly suspicious CT images from 87,626 lung CT images of 1,285 patients (7.11%), averaging 4.85 CT images per patient in the external validation dataset. Most of these CT images contain lesions or nodules. The filtered CT images were used for inputs of stage-2 models.

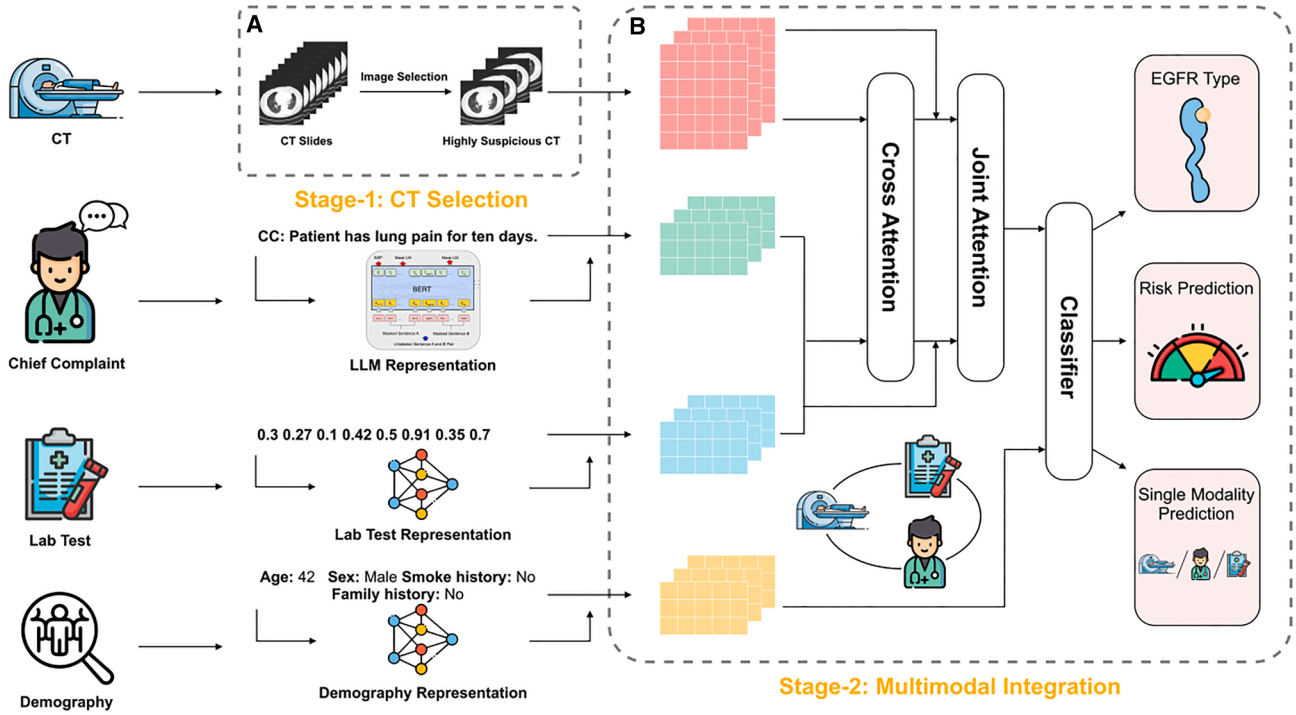


Figure 1. Pipeline of the multimodal LUCID lung cancer diagnostic model

(A and B) The model comprises two primary stages. (A) Stage 1 performs binary classification on CT images to identify those with high suspicion of malignancy. (B) Stage 2 integrates four modalities—CT images, textual complaints, lab results, and demographic data—to predict *EGFR* mutation type or survival time. LUCID is robust to missing modalities and adapts accordingly.

In clinical settings, the stage-1 model can function independently as a software tool for radiologists, enabling them to quickly isolate CT images that require detailed examination. This capability can significantly streamline the CT review process, allowing radiologists to concentrate on the most suspicious images, thereby potentially improving diagnostic efficiency and accuracy.

Multimodal integration synthesizes holistic insights

The LUCID framework effectively integrates multimodal medical data, ranging from CT images, patients' textual chief complaints, and laboratory test results to demographic information, through a hierarchical pipeline (Figure 3A). This multimodal integration capability enhances the framework's flexibility and applicability.[41,42]

During training and inference, CT images are first divided into patches and encoded using a pretrained CNN model.[38] Simultaneously, the patients' textual chief complaints are processed through BERT language model[39] to generate text embeddings, while laboratory test results are encoded to match the dimensionality of these text embeddings. Then, the text embedding and laboratory test embedding features are concatenated into a combined vector as the clinical embeddings. At the core of LUCID's architecture, we employed a cross-attention mechanism[43,44] within the first three layers. This cross-attention mechanism enables bidirectional information exchange between modalities through symmetric attention flows. When processing CT image embeddings, the clinical features act as attention keys while the CT image embeddings serve as queries and values. This process is mirrored in the opposite direction, allowing each modality to guide and enhance the representation of the other through mutual cross-attention (Figure 3B). We employ a joint-attention mechanism in the subsequent nine layers to further consolidate and integrate the multimodal features[37] (Figure 3C). Prior to final classification, patient demographic information is encoded and concatenated with the fused multimodal representations. This hierarchical combination of cross-attention followed by joint-attention enables the model to progressively fuse the heterogeneous data streams. By first establishing modality-specific interactions through cross-attention and then performing deeper integration through joint-attention, the model effectively synthesizes the complementary information from both modalities.

We designed a specialized loss function that combines multiple classification objectives (Figure 3B, see STAR Methods for details). The function incorporates three components: the primary loss from the final multimodal predictions and auxiliary classification losses from both the CT image and clinical embeddings after the cross-attention stage. This multi-level supervision ensures that each modality's representations remain discriminative even before final fusion, enabling robust performance when only partial modalities are available.

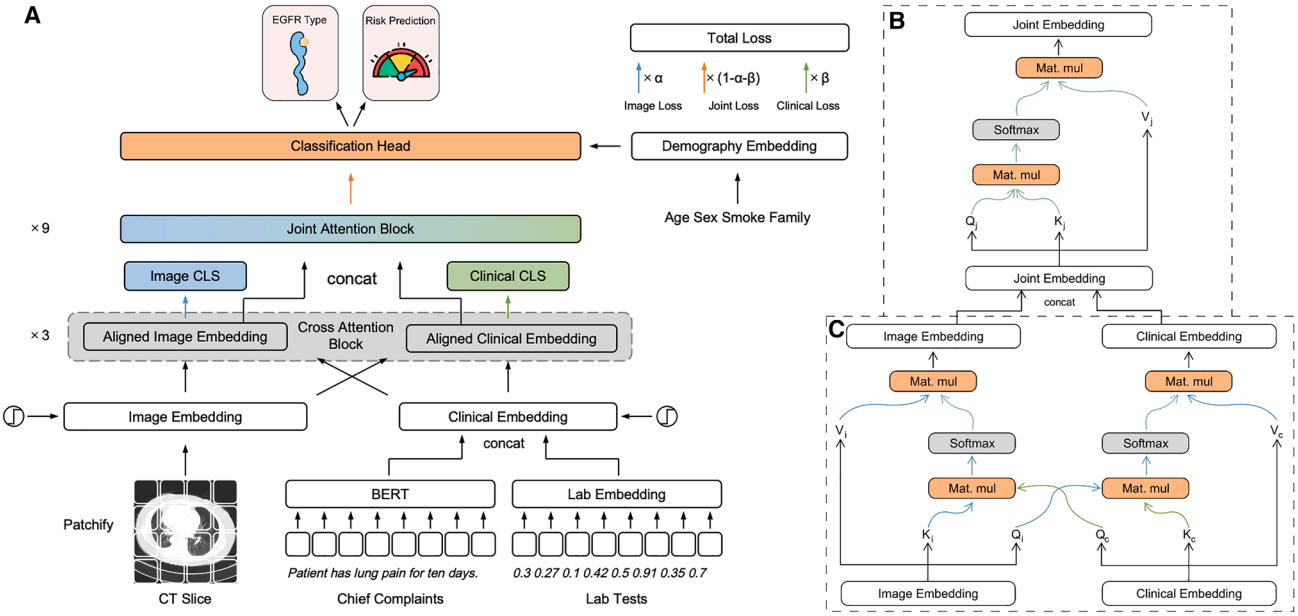


Figure 2. The architecture of the multimodal integration module of LUCID

(A) LUCID incorporates four distinct input modalities: CT images, patient's textual chief complaints, laboratory test results, and demographic information. Each modality is initially encoded into a vector representation before sending into the hierarchical multimodal integration framework. The loss computation encompasses image classification loss, clinical classification loss, and joint loss.

(B) Joint-attention mechanism: image and clinical embeddings are merged to create a joint embedding. Multihead attention calculations are then performed to discern the inner-attention across varied features.

(C) Cross-attention mechanism: embeddings from one modality serve as the key vector in multihead attention computations, while embeddings from the other modality function as value and query vectors. This facilitates the attention calculations between distinct modalities.

LUCID can accurately distinguish *EGFR* type

We evaluated LUCID on two binary *EGFR* classification tasks: (1) distinguishing *EGFR*-sensitive cases ($N = 591$) from *EGFR*-resistant cases ($N = 444$) and (2) distinguishing *EGFR*-wild-types cases ($N = 599$) from other cases ($N = 436$). LUCID demonstrated strong performance on both *EGFR* classification tasks. For *EGFR*-sensitive versus *EGFR*-resistant classification, LUCID achieved an AUC of 0.881 (95% CI, 0.860–0.901) (Figure 4A). In the *EGFR*-wild-types versus *EGFR*-mutant-type classification, LUCID achieved an AUC of 0.851 (95% CI, 0.836–0.866) (Figure 4B).

To better demonstrate the effectiveness of the LUCID framework, we compare it against a variety of unimodal and bimodal models (see STAR Methods for details). These include the *Basic Info Only model*, an support vector machine classification model[45] that relies solely on basic demographic information such as age, sex, family disease history, and smoke history; the *Text Only model,* which utilizes a pretrained BERT language model[39] to classify text derived from patients' chief complaints; the *Lab Test Only model,* a random forest model[46] categorizing patients based on their laboratory test results; the *Image Only model*, a pretrained ViT model[34] for binary classification on CT images; and the *Text + Lab Test model*, a model that is based on the transformer architecture,[47] integrating both patient complaints and laboratory test data through cross-attention for classification.

Performance comparisons across AUC, accuracy, recall, precision, and F1 metrics between LUCID (*All Modalities*) and other baseline models are shown in Figures 4A and 4B. LUCID consistently outperformed unimodal and bimodal baselines on all evaluation metrics. In the *EGFR*-sensitive classification task, the Image Only model achieved the highest performance among other baseline models with an AUC of 0.822 (95% CI, 0.801–0.843), followed by the Text + Lab Test model (AUC of 0.787 [95% CI, 0.732–0.841]), Basic Info Only model (AUC of 0.762 [95% CI, 0.743–0.782]), Lab Test Only model (AUC of 0.725 [95% CI, 0.715–0.736]), and Text Only model (AUC of 0.669 [95% CI, 0.645–0.694]). For *EGFR*-mutant-type classification, a similar performance pattern emerged. The Image Only model led with an AUC of 0.807 (95% CI, 0.800–0.814), followed by the Text + Lab model (AUC of 0.794 [95% CI, 0.726–0.862]), Basic Information model (AUC of 0.764 [95% CI, 0.746–0.781]), Laboratory Test model (AUC of 0.730 [95% CI, 0.715–0.744]), and Text Only model (AUC of 0.650 [95% CI, 0.620–0.680]). All confidence intervals were computed across five random seeds.

Multimodal *EGFR* mutation prediction offers several clinical advantages. By leveraging readily avaliable data such as CT images, laboratory test results, and information from electronic medical records (EMR) systems, this approach could reduce reliance on invasive tissue biopsies, minimizing patient risk and discomfort. The rapid, non-invasive nature of this prediction method could potentially accelerate treatment decisions by

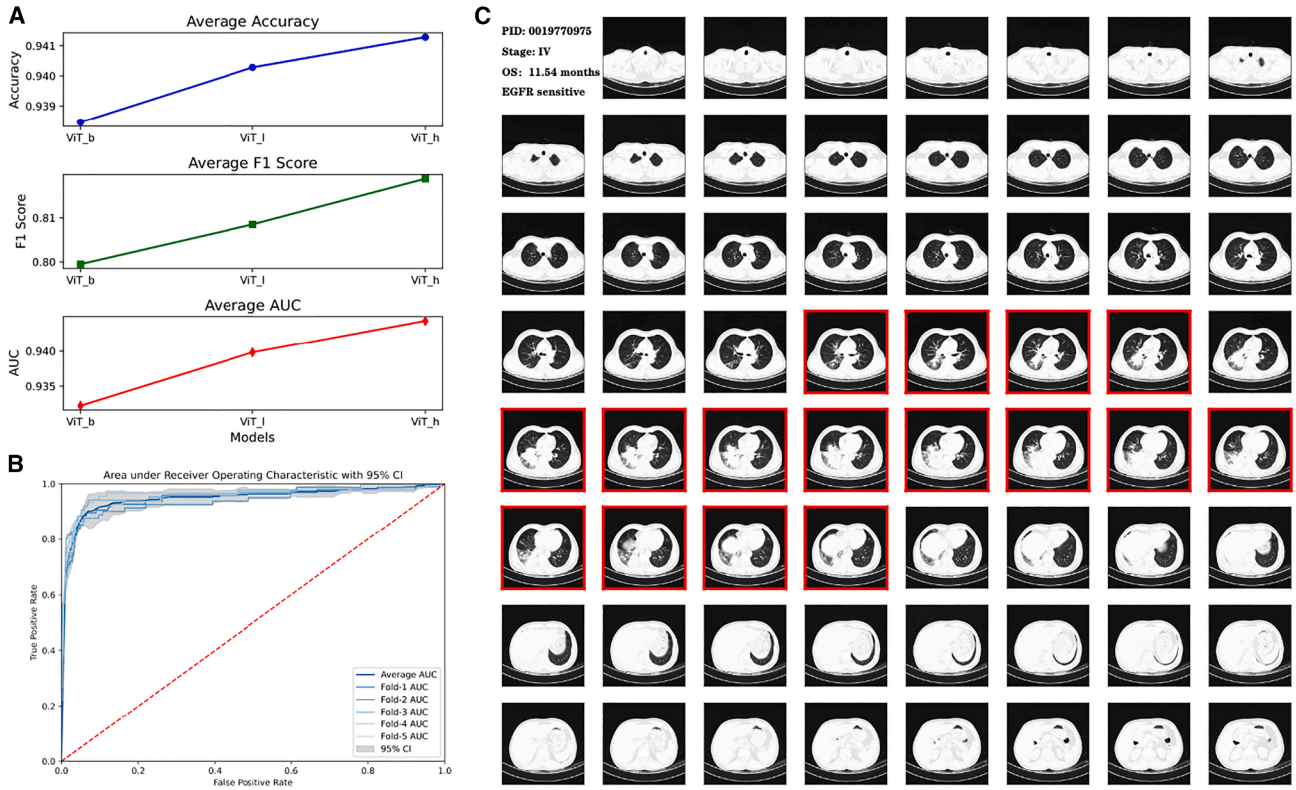


Figure 3. Performance and empirical examples of the stage-1 classification model

(A) We evaluated ViT models at varying parameter scales and observed a scaling law in classification capability with the increase of the models' parameters.

(B) A 5-fold validation was conducted during the training of the stage-1 model, and we plotted the AUC along with the 95% CI region.

(C) All CT images from a patient (PID:0019770975) diagnosed with stage-4 lung cancer were processed through our stage-1 model. CT images outlined in red indicate the model's predictions with a high suspicion of lung cancer presence. Data are represented as mean ± 95% CI.

providing preliminary *EGFR* status information while awaiting confirmatory molecular testing results.

LUCID can effectively predict patient outcomes

We extended LUCID's evaluation to survival prediction tasks across multiple time horizons: 1-year (835 [87.7%] survivors, 117 [12.3%] deceased), 3-year (469 [49.3%] survivors, 483 [50.7%] deceased), and 5-year (199 [20.9%] survivors, 753 [79.1%] deceased). Following with the same comparative framework used in *EGFR* mutation prediction, we benchmarked LUCID against baseline models for each survival endpoint. The results are presented in Figures 4C–4E. LUCID demonstrated strong predictive performance across all survival prediction task. For 1-year survival prediction, the model achieved an AUC of 0.821 (95% CI, 0.806–0.836) and accuracy of 0.764 (95% CI, 0.754–0.774). Similar performance was observed for 3-year survival prediction, with an AUC of 0.884 (95% CI: 0.844–0.924) and accuracy of 0.843 (95% CI: 0.822–0.864). The 5-year prediction task yielded an AUC of 0.912 (95% CI: 0.886–0.937) and accuracy of 0.888 (95% CI: 0.835–0.941). These results consistently exceeded the performance of unimodal and bimodal baselines across all evaluation metrics.

In addition, we visualized Kaplan-Meier (KM) survival curves for LUCID-identified risk groups across all three survival times in Figures 4C–4E. For 1-year survival prediction, our model correctly identified almost all low-risk patients, and the survival status of the high-risk patients, as indicated by the model, aligned with the survival curve. The 3-year and 5-year predictions also demonstrated clear separation between risk groups, with survival patterns closely matching the model's risk stratification. Statistical significance of these risk stratifications was confirmed through log rank tests (p-value of 3.39×10^{-6} for 1 year predictions, 8.27×10^{-20} for 3 year predictions, and 1.63×10^{-30} for 5 year predictions) of the KM curves.

LUCID's ability to stratify patients by survival risk across multiple time horizons has practical clinical relevance. The model's risk predictions could assist clinicians in treatment planning and resource allocation by providing objective prognostic information.

Multimodal integration improves single modal performance

Although multimodal integration using LUCID offer high predictive performance, they require comprehensive data collection across multiple modalities for training and validation, which can be resource intensive in clinical settings. This limitation motivated us to investigate whether LUCID's multimodal architecture could benefit single-modality scenarios through knowledge

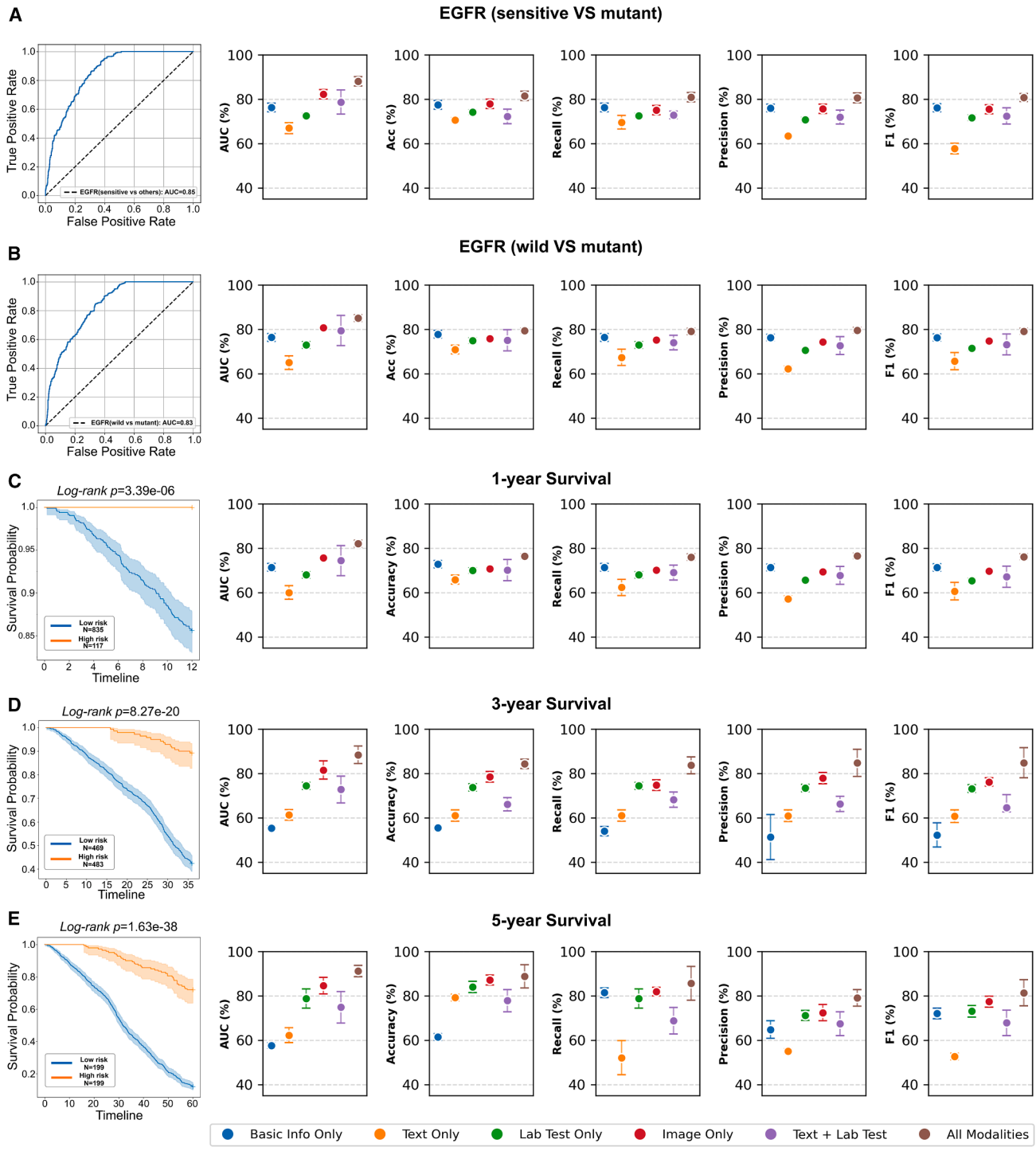


Figure 4. Performance evaluation of LUCID in EGFR type and survival prediction

Comparison of LUCID's accuracy with other methods in two EGFR mutation prediction tasks (N = 1,035) and three survival prediction tasks (N = 952).

(A) AUC and method comparisons for distinguishing EGFR-sensitive cases from others (591:444).

(B) AUC and method comparisons for distinguishing EGFR wild type from mutant-type cases (599:436), including unimodal and bimodal approaches.

(C–E) KM curves and method comparisons for predicting 1-year (835:117), 3-year (469:483), and 5-year (199:753) survival. Data are represented as mean ± 95% CI. p values for KM curves are calculated.

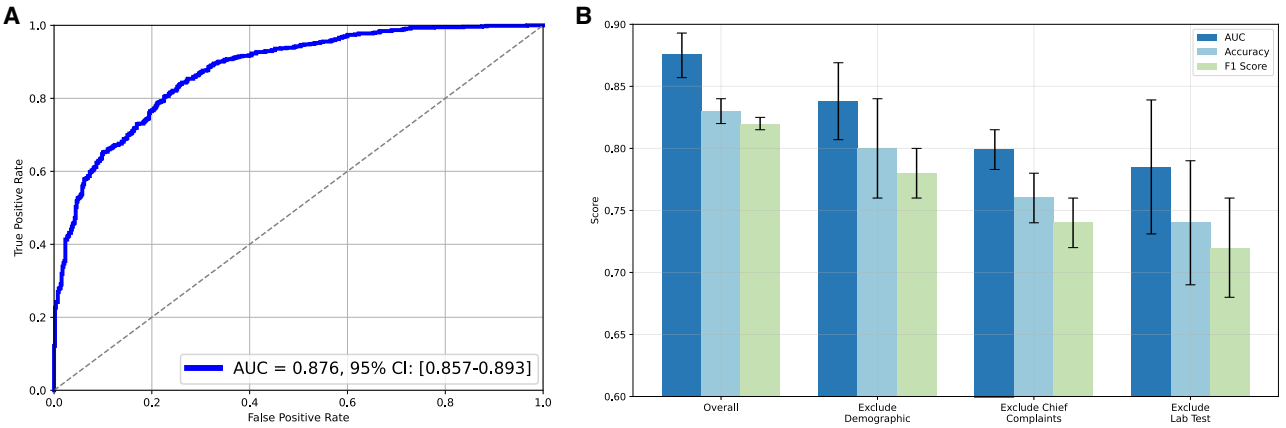


Figure 5. LUCID's performance on external validation dataset and evaluations of different modalities

(A) LUCID demonstrated strong predictive performance with an AUC of 0.876 (95% CI: 0.857–0.893) on the external validation dataset (N = 1,285).

(B) Overall performance and three feature exclusion experiments suggest that different feature modalities contribute differently to the model's predictive capability. Data are represented as mean ± 95% CI.

transfer. Recent advances in computer vision have shown that models trained on paired image-text data can develop robust zero-shot capabilities,[48] suggesting that knowledge from multimodal training might transfer effectively to unimodal scenarios. Building on these insights, we evaluated whether LUCID's multimodal training could enhance performance when operating with limited modalities at inference time.

We evaluated LUCID's performance with unimodal inputs by randomly initializing the missing modalities during inference. For *EGFR*-sensitive mutation prediction, our multimodal method yielded superior results: with image-only input, LUCID achieved an AUC of 0.822 and F1 score of 0.754, surpassing the image-only baseline model (AUC: 0.803, F1: 0.718). Similarly, with text-only input, LUCID attained an AUC of 0.669 and F1 score of 0.677, exceeding the text-only baseline performance (AUC: 0.657, F1: 0.606). These improvements over modality-specific baselines (Table S1) suggest that LUCID's multimodal training enables effective knowledge transfer even when operating with limited modality access.

External validation demonstrates strong generalizability

We evaluated LUCID on an external validation dataset, using patients' multimodal medical information to predict *EGFR* mutation status (*EGFR*-mutant-type cases [N = 734] versus *EGFR*-wild-type cases [N = 551]). As shown in Figure 5A, LUCID achieved an AUC of 0.876 (95% CI, 0.857–0.893) for *EGFR* mutation binary classification across the total 1,285 patients. This robust performance on an independent cohort demonstrates the model's strong generalizability beyond the training dataset.

Additionally, we conducted an ablation study to assess the importance of different input modalities. Specifically, we removed one modality among demographic information, textual chief complaints, or laboratory test results each time and calculated the model's performance using the remaining three modalities; results are shown in Figure 5B. The ablation analysis revealed that removing laboratory test results led to the largest performance drop (ΔAUC = 0.094, AUC = 0.785, 95% CI: 0.731–0.839), followed by textual chief complaints (ΔAUC = 0.080, AUC = 0.799, 95% CI: 0.783–0.815), while removing demographic information had a relatively smaller impact (ΔAUC = 0.041, AUC = 0.838, 95% CI: 0.807–0.869). This quantitative analysis confirms that, besides CT images, laboratory markers also contain particularly valuable signals for *EGFR* mutation prediction, though the integration of all modalities yields optimal performance.

DISCUSSION

This paper introduces LUCID, a multimodal integration model based on transformer architecture utilizing cross-attention and joint-attention mechanisms. By integrating diverse data sources, including CT images, textual chief complaints, laboratory test results, and patient demographic information, LUCID optimizes the diagnostic process for lung cancer and enhances the prediction of *EGFR* mutations and prognostic outcomes.

In comparative analyses with existing unimodal models, LUCID consistently demonstrates superior performance and robustness in predicting *EGFR* mutation types and prognosis, both on test datasets and external validation datasets. Its multimodal framework outperforms other unimodal models even in scenarios with missing data, further validating its applicability and resilience.

As a two-stage model, LUCID's first phase effectively identifies potential lesions from CT image sequences; this can be applied to assist radiologists and clinicians in their diagnostic processes. Additionally, as a non-invasive method, LUCID's multimodal predictive capabilities can be integrated into clinical decision support systems, providing clinicians with an analytical framework that enhances diagnostic accuracy and reduces cognitive load. Its adaptability to incomplete data ensures decision-making, while patients can benefit from personalized treatment plans and improved therapeutic outcomes. This also facilitates clearer prognostic understanding, promoting informed discussions with healthcare providers.

LUCID serves as a general training framework applicable to various multimodal medical data, such as X-ray and gene sequence, and is capable of predicting different cancer-related gene mutations and prognostic outcomes. However, the LUCID approach has certain limitations. The current model requires a substantial amount of high-quality, matched multimodal data for training, which is challenging to obtain in real-world settings. Nevertheless, advancements in vision-language foundation models and synthetic data technologies may help overcome this challenge. Additionally, incorporating a new gene mutation type necessitates retraining the model, which is computationally intensive. Future efforts could explore using multimodal data to predict comprehensive genomic information.

Our survival prediction model offers significant clinical utility by equipping oncologists with a data-driven tool to assess patient prognosis with greater accuracy. In particular, the model is designed for patients in the early stages of clinical evaluation—those who have not yet been definitively diagnosed or had treatment plans established. In practical terms, the model can be seamlessly integrated into routine clinical workflows, leveraging widely available data such as CT images, laboratory results, and demographic information. By providing survival predictions at multiple time points—specifically 1, 3, and 5 years—it empowers clinicians to develop personalized treatment plans tailored to each patient's unique risk profile. This capability optimizes resource allocation, enhances follow-up care, and enables more informed decision-making. For patients presenting with initial suspicious findings, such as abnormal CT scans or symptoms, the model provides early prognostic insights that can guide diagnostic prioritization, biopsy decisions, and preliminary treatment discussions. For patients, the model fosters a clearer understanding of their prognosis and available care options, promoting shared decision-making between patients and healthcare providers. Furthermore, the model's ability to stratify patients based on survival risk facilitates timely interventions or adjustments to treatment strategies, particularly for those identified as having poorer outcomes. This approach is non-invasive, scalable, and cost effective, making it a valuable complement to traditional clinical assessments. Ultimately, our model enhances lung cancer prognosis by enabling personalized care, which we believe can lead to improved patient outcomes.

In summary, LUCID represents a significant advancement in the diagnosis and prognostic prediction for lung cancer patients through multimodal data integration. By combining CT imaging with multimodal information, it provides clinicians and researchers with a reliable tool that enhances decision-making and patient management. LUCID also models the associations between different modalities and gene mutations, laying the groundwork for future research.

Limitations of the study

Although this study has achieved promising results, there are still several limitations to be addressed. First, LUCID's performance depends heavily on the availability of high-quality, well-aligned multimodal data, including CT images, textual chief complaints, laboratory test results, and demographic information. In real-world clinical environments, such comprehensive datasets are often incomplete, heterogeneous, or inconsistently annotated,[49] which poses challenges to generalizability and model robustness across institutions. Second, while LUCID demonstrates adaptability to missing data during inference, training the model still requires a substantial volume of fully matched multimodal inputs. This limitation can hinder scalability and limit its applicability in resource-constrained settings. Advances in vision-language foundation models and synthetic data generation may help alleviate this barrier by augmenting training data or improving model pretraining.[7] Another limitation lies in LUCID's gene mutation prediction capability. The model currently requires retraining to accommodate new mutation types, which is computationally intensive and may not be practical for dynamic clinical environments where genetic markers of interest evolve over time.[50] Future extensions could explore continual learning strategies or modular architectures to enable more flexible updates.

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Xiaoying Huang (huangxiaoying@wzhospital.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Restrictions apply to the availability of the datasets, which were used with permission of the participants for the current study. De-identified data may be available for research purposes from the corresponding authors on reasonable request. The code for LUCID is deposited in GitHub at https://github.com/YuxingLu613/LUCID, which is publicly accessible.

Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This research was supported by Wenzhou Eye Hospital, Wenzhou Medical University Eye Health and Disease Advanced Institute; the Macau Science and Technology Development Fund, Macao (0007/2020/AFJ, 0070/2020/A2, and 0003/2021/AKP); Guangzhou National Laboratory (YW-SLJC0201); National Natural Science Foundation of China (32100631, 32470964, W2431057, and 6220071694); National Key Research and Development Program of China (2024YFF0507404); Young Elite Scientist Sponsorship Program by Beijing Association for Science and Technology (BYESS2023026); Beijing Nova Program (20240484627); Capital's Funds for Health Improvement and Research (2024-4-40215); China Postdoctoral Science Foundation (2023T160061); Macao Young Scholars Program (AM2023018 and AM2023024); Science and Technology Foundation of Sichuan Province, Outstanding Young Scientist Fund (2024NSFJQ0051); and four major chronic disease projects of the Ministry of Science and Technology (2023ZD0506105). We thank the physicians and patients for providing clinical data.

AUTHOR CONTRIBUTIONS

Y.L., F.L., Y.Y., C.B., W.Y., Z.Z., K.L., M.M., C.O., C.W., J.W., and K.Z. collected and analyzed the data. C.W., X.H., J.W., and K.Z. conceived, designed, and supervised the project. Y.L., F.L., Y.Y., C.W., J.W., and K.Z. wrote the manuscript. All authors discussed the results and reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

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
  - *EGFR* mutation detection
  - Data preprocessing
  - Image preprocessing
  - Text preprocessing
  - Laboratory test data preprocessing
  - Demographic data preprocessing
  - Stage-1 binary classification model
  - Stage-2 multimodal integration model
  - Baseline models for comparison
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2025.102216.

Received: June 10, 2024
Revised: February 10, 2025
Accepted: June 6, 2025
Published: July 2, 2025

REFERENCES

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *74*, 229–263.
2. Siegel, R.L., Giaquinto, A.N., and Jemal, A. (2024). Cancer statistics, 2024. CA Cancer J. Clin. *74*, 12–49.
3. 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.
4. Chaunzwa, T.L., Hosny, A., Xu, Y., Shafer, A., Diao, N., Lanuti, M., Christiani, D.C., Mak, R.H., and Aerts, H.J.W.L. (2021). Deep learning classification of lung cancer histology using CT images. Sci. Rep. *11*, 5471. https://doi.org/10.1038/s41598-021-84630-x.
5. Ladbury, C., Amini, A., Govindarajan, A., Mambetsariev, I., Raz, D.J., Massarelli, E., Williams, T., Rodin, A., and Salgia, R. (2023). Integration of artificial intelligence in lung cancer: Rise of the machine. Cell Rep. Med. *4*, 100933.
6. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature *620*, 172–180.
7. Wang, J., Wang, K., Yu, Y., Lu, Y., Xiao, W., Sun, Z., Liu, F., Zou, Z., Gao, Y., Yang, L., et al. (2025). Self-improving generative foundation model for synthetic medical image generation and clinical applications. Nat. Med. *31*, 609–617.
8. Wang, F., Li, S., Gao, Y., and Li, S. (2024). Computed tomography-based artificial intelligence in lung disease—Chronic obstructive pulmonary disease. MedComm–Future Medicine *3*, e73.
9. Kann, B.H., Hosny, A., and Aerts, H.J.W.L. (2021). Artificial intelligence for clinical oncology. Cancer Cell *39*, 916–927.
10. Swanson, K., Wu, E., Zhang, A., Alizadeh, A.A., and Zou, J. (2023). From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell *186*, 1772–1791.
11. Elemento, O., Leslie, C., Lundin, J., and Tourassi, G. (2021). Artificial intelligence in cancer research, diagnosis and therapy. Nat. Rev. Cancer *21*, 747–752.
12. Bitterman, D.S., Downing, A., Maués, J., and Lustberg, M. (2024). Promise and Perils of Large Language Models for Cancer Survivorship and Supportive Care. J. Clin. Oncol. *42*, 1607–1611. https://doi.org/10.1200/jco.23.02439.
13. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al. (2025). A vision–language foundation model for precision oncology. Nature *638*, 769–778.
14. Chen, C.-L., Chen, C.-C., Yu, W.-H., Chen, S.-H., Chang, Y.-C., Hsu, T.-I., Hsiao, M., Yeh, C.-Y., and Chen, C.-Y. (2021). An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nat. Commun. *12*, 1193.
15. Zhou, H.-Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., and Li, W. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat. Biomed. Eng. *7*, 743–755.
16. Xiao, W., Yuxing, L., Miao, H., Zhao, W., and Schanzlin, D. (2024). Deep-learning-based diagnosis of myopia in children using optical coherence tomography angiography. MedComm–Future Medicine *3*, e72.
17. Xia, K., and Wang, J. (2023). Recent advances of transformers in medical image analysis: a comprehensive review. MedComm–Future Medicine *2*, e38.
18. Abul-Husn, N.S., and Kenny, E.E. (2019). Personalized medicine and the power of electronic health records. Cell *177*, 58–69.
19. Lu Y., Zhao X., and Wang J. (2023). Medical Knowledge-Enhanced Prompt Learning for Diagnosis Classification from Clinical Text. The 61st Annual Meeting Of The Association For Computational Linguistics. pp. 278-288.
20. Shah, N.H., Entwistle, D., and Pfeffer, M.A. (2023). Creation and adoption of large language models in medicine. JAMA *330*, 866–869.
21. Osorio, D. (2022). Interpretable multi-modal data integration. Nat. Comput. Sci. *2*, 8–9.
22. Boehm, K.M., Aherne, E.A., Ellenson, L., Nikolovski, I., Alghamdi, M., Vázquez-García, I., Zamarin, D., Long Roche, K., Liu, Y., Patel, D., et al. (2022). Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nat. Cancer *3*, 723–733.
23. Wang, J., Gao, Y., Wang, F., Zeng, S., Li, J., Miao, H., Wang, T., Zeng, J., Baptista-Hon, D., Monteiro, O., et al. (2024). Accurate estimation of biological age and its application in disease prediction using a multimodal image Transformer system. Proc. Natl. Acad. Sci. USA *121*, e2308812120.
24. Sun, Y.L., and Liu, F. (2025). Real-World Implementation of an AI Learning Tool-MetaGP-Edu in Medical Education: A Multi-Center Cohort Study. Computers & Education, 105388. https://doi.org/10.1016/j.compedu.2025.105388.
25. Wan, P., Huang, Z., Tang, W., Nie, Y., Pei, D., Deng, S., Chen, J., Zhou, Y., Duan, H., Chen, Q., and Long, E. (2024). Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. Nat. Med. *30*, 2878–2885.
26. Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J.A., Wornow, M., Swaminathan, A., and Lehmann, L.S. (2024). Testing and evaluation of health care applications of large language models: a systematic review. JAMA *333*, 319–328.
27. Sharma, S.V., Bell, D.W., Settleman, J., and Haber, D.A. (2007). Epidermal growth factor receptor mutations in lung cancer. Nat. Rev. Cancer *7*, 169–181.
28. Inukai, M., Toyooka, S., Ito, S., Asano, H., Ichihara, S., Soh, J., Suehisa, H., Ouchida, M., Aoe, K., Aoe, M., et al. (2006). Presence of epidermal growth factor receptor gene T790M mutation as a minor clone in non–small cell lung cancer. Cancer Res. *66*, 7854–7858.
29. Mitsudomi, T., and Yatabe, Y. (2007). Mutations of the epidermal growth factor receptor gene and related genes as determinants of epidermal

growth factor receptor tyrosine kinase inhibitors sensitivity in lung cancer. Cancer Sci. *98*, 1817–1824.

30. Rosell, R., Perez-Roca, L., Sanchez, J.J., Cobo, M., Moran, T., Chaib, I., Provencio, M., Domine, M., Sala, M.A., Jimenez, U., et al. (2009). Customized treatment in non-small-cell lung cancer based on EGFR mutations and BRCA1 mRNA expression. PLoS One *4*, e5133.
31. Nesbitt, J.C., Putnam, J.B., Jr., Walsh, G.L., Roth, J.A., and Mountain, C.F. (1995). Survival in early-stage non-small cell lung cancer. Ann. Thorac. Surg. *60*, 466–472.
32. Stanley, K.E. (1980). Prognostic factors for survival in patients with inoperable lung cancer. J. Natl. Cancer Inst. *65*, 25–32.
33. Quer, G., and Topol, E.J. (2024). The potential for large language models to transform cardiovascular medicine. Lancet Digit. Health *6*, e767–e771.
34. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11929.
35. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A Large-Scale Hierarchical Image Database (Ieee), pp. 248–255.
36. Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. science *313*, 504–507.
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*, 5998-6008.
38. LeCun Y. and Bengio Y. (1995). Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks 3361. MIT Press. 1995.
39. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.
40. Zhai X., Kolesnikov A., Houlsby N., and Beyer L. (2022). Scaling Vision Transformers. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12104-12113.
41. Vanguri, R.S., Luo, J., Aukerman, A.T., Egger, J.V., Fong, C.J., Horvat, N., Pagano, A., Araujo-Filho, J.d.A.B., Geneslaw, L., Rizvi, H., et al. (2022). Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L) 1 blockade in patients with non-small cell lung cancer. Nat. Cancer *3*, 1151–1164.
42. Cheerla, A., and Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics *35*, i446–i454.
43. Wei X., Zhang T., Li Y., Zhang Y., and Wu F. (2020). Multi-modality Cross Attention Network for Image and Sentence Matching. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10941-10950.
44. Chen C.-F.R., Fan Q., and Panda R. (2021). Crossvit: Cross-Attention Multi-Scale Vision Transformer for Image Classification. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 357-366.
45. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.
46. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.
47. Soyalp, G., Alar, A., Ozkanli, K., and Yildiz, B. (2021). Improving Text Classification with Transformer. (IEEE), pp. 707-712.
48. Meyer, J., Eitel, A., Brox, T., and Burgard, W. (2020). Improving Unimodal Object Recognition with Multimodal Contrastive Learning. IEEE ASME Trans. Mechatron., 5656–5663.
49. Yu, Y., Li, M., Liu, L., Li, Y., and Wang, J. (2019). Clinical big data and deep learning: Applications, challenges, and future outlooks. Big Data Min. Anal. *2*, 288–305.
50. Schlötterer, C. (2004). The evolution of molecular markers—just a matter of fashion? Nat. Rev. Genet. *5*, 63–69.
51. Slatko, B.E., Gardner, A.F., and Ausubel, F.M. (2018). Overview of next-generation sequencing technologies. Curr. Protoc. Mol. Biol. *122*, e59.
52. Yohe, S., and Thyagarajan, B. (2017). Review of clinical next-generation sequencing. Arch. Pathol. Lab Med. *141*, 1544–1557.
53. Hedegaard, J., Thorsen, K., Lund, M.K., Hein, A.-M.K., Hamilton-Dutoit, S. J., Vang, S., Nordentoft, I., Birkenkamp-Demtröder, K., Kruhøffer, M., Hager, H., et al. (2014). Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PLoS One *9*, e98187.
54. Meldrum, C., Doyle, M.A., and Tothill, R.W. (2011). Next-generation sequencing for cancer diagnostics: a practical perspective. Clin. Biochem. Rev. *32*, 177–195.
55. Normanno, N., Denis, M.G., Thress, K.S., Ratcliffe, M., and Reck, M. (2016). Guide to detecting epidermal growth factor receptor (EGFR) mutations in ctDNA of patients with advanced non-small-cell lung cancer. Oncotarget *8*, 12501–12516.
56. Riess, J.W., Gandara, D.R., Frampton, G.M., Madison, R., Peled, N., Bufill, J.A., Dy, G.K., Ou, S.-H.I., Stephens, P.J., McPherson, J.D., et al. (2018). Diverse EGFR exon 20 insertions and co-occurring molecular alterations identified by comprehensive genomic profiling of NSCLC. J. Thorac. Oncol. *13*, 1560–1568.
57. Sheikine, Y., Rangachari, D., McDonald, D.C., Huberman, M.S., Folch, E. S., VanderLaan, P.A., and Costa, D.B. (2016). EGFR testing in advanced non–small-cell lung cancer, a mini-review. Clin. Lung Cancer *17*, 483–492.
58. Ellison, G., Zhu, G., Moulis, A., Dearden, S., Speake, G., and McCormack, R. (2013). EGFR mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples. J. Clin. Pathol. *66*, 79–89.
59. Khosla, C., and Saini, B.S. (2020). Enhancing Performance of Deep Learning Models with Different Data Augmentation Techniques: A Survey. IEEE ASME Trans. Mechatron., 79–85.
60. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. (2024). Towards a general-purpose foundation model for computational pathology. Nat. Med. *30*, 850–862. https://doi.org/10.1038/s41591-024-02857-3.
61. Beretta, L., and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. BMC Med. Inform. Decis. Mak. *16*, 74–208.
62. Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.05101.
63. Loshchilov, I., and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. Preprint at arXiv. https://doi.org/10.48550/arXiv.1608.03983.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
West China Hospital dataset	Retrospective cohort (5,175 patients)	Available upon request (see data and code availability)
Sun Yat-sen Memorial Hospital dataset	External validation cohort (1,285 patients)	Available upon request (see data and code availability)
Software and algorithms		
PyTorch	PyTorch Foundation	https://pytorch.org/
NVIDIA Apex	NVIDIA	https://developer.nvidia.com/apex
Vision Transformer (ViT)	Torchvision	https://pytorch.org/vision/main/models.html
BERT	Hugging Face	https://huggingface.co/google-bert/bert-base-chinese
LUCID	GitHub	https://doi.org/10.5281/zenodo.15259609

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

In our work, we used a large dataset from West China Hospital for training and test, which contains 8,162 patients in total and an external validation dataset from Sun Yat-sen Memorial Hospital containing 1285 patients. Each patient has his/her medical records, including demographic information, laboratory test results, and clinical notes. Since all the patients had lung diseases, most of them had at least one CT scan. We aligned all patients who had both medical records and CT scans, resulting in paired data for 5,175 patients for West China Hospital dataset and all the patients for Sun Yat-sen Memorial Hospital dataset. The details of the patients' charateristics can be found in Tables 1 and 2, and a comparison between the two datasets can be found in Table S2. Additionally, we asked specialists to annotate 4,099 lung CT images of 56 independent patients and to identify those CT images with possible lesions. The annotated images were used to train the Stage-1 binary classification model to distinguish the normal and abnormal CT images.

By assuring each patient has at least one CT scan, we align the medical records according to their patient IDs. We selected patients' textual chief complaints, laboratory test results, and demographic information as multimodal inputs. In terms of chief complaints, we extract them from the patient's visit record and limit the length of the text to 40, which is about twice the median length of all chief complaints. The texts beyond the threshold are truncated. For laboratory results, we selected all the lab test results from complete blood counts consisting 92 features and normalized their values. For demographic information, we manually selected four features: age, sex, smoking history, and family lung cancer history. For patients have no chief complaints, we use "the patient reports chest pain for several days." for placeholder. For patients have no lab test results and demographic information, we use average feature value for continuous features and 0 for categorical features imputation.

For the West China Hospital dataset, we extracted 20% of the patients (1,035 patients) as the test dataset and divided the remaining 80% of the patients (4,140 patients) into a training dataset. All the 1,285 patients from the Sun Yat-sen Memorial Hospital dataset are taken as the independent external validation dataset, followed by the same preprocessing procedures.

METHOD DETAILS

EGFR mutation detection

EGFR mutation detection was performed using next-generation sequencing (NGS).[51,52] Tumor DNA was extracted from formalin-fixed paraffin-embedded (FFPE) tissue samples or fresh biopsy specimens using a commercially available DNA extraction kit.[53,54] The DNA concentration and purity were assessed using a NanoDrop spectrophotometer. The NGS panel targeted critical regions of the *EGFR* gene, including exons 18, 19, 20, and 21, which are commonly associated with lung cancer.[55–57] Sequencing was conducted on a high-throughput platform, ensuring a depth of coverage sufficient for detecting mutations with high sensitivity and specificity. Variant calling and annotation were performed using validated bioinformatics pipelines, which included quality control measures to minimize sequencing artifacts. Identified mutations, such as exon 19 deletions and *L858R* point mutations, were validated against known standards to ensure accuracy.[58]

Data preprocessing

Patient data are available in multiple modalities, including CT images, textual chief complaints, laboratory test results and demographical information. Before integrating these modalities, it is essential to preprocess and represent each type of data using a standardized pipeline.

Image preprocessing

CT images are typically high-resolution and vary in size. Before sent to the model, all CT images are standardized to ensure consistency and enhancing the model's ability to learn from the data. In the training phase, each image in the CT image set is resized to a standard dimension of 256 $\times$ 256 pixels. Following this, a random portion of each resized image is cropped. The size of this cropped patch is determined by an area ratio, which is randomly selected between 0.09 and 1.0 relative to the original image size. This introduces variability in the training data, helping the model generalize better. The randomly cropped section is then resized to 224 $\times$ 224 pixels, which is often required for models pre-trained on datasets like ImageNet.[35] To further enhance the diversity of the training data, a random horizontal flip and rotation are applied to the cropped patch.[59]

During the validation and testing phases, each image is resized to 256 $\times$ 256 pixels, similar to the training phase. However, instead of random cropping, a central square patch of 224 $\times$ 224 pixels is extracted. This ensures that the evaluation is consistent and not influenced by random transformations, providing a stable benchmark for model performance.[60]

Text preprocessing

LUCID takes textual data related to patients, specifically focusing on their chief complaints. Each text entry provides information for diagnosis and prognosis tasks. The text data varies in length and content, necessitating a preprocessing strategy to standardize the input for machine learning models.

First, each patient's chief complaint text is adjusted to a fixed length of 40 words. For texts that exceed this length, truncation is applied to retain only the first 40 words. Conversely, for texts with fewer than 40 words, padding is employed using a special < pad > token to reach the desired length. Next, we utilize an autoencoder-based LLM, specifically BERT[39] in our experiments, to convert each word in the text into a 768-dimensional context-aware vector. The combined vector for each text segment is used as the final representation. For patients without recorded chief complaints, we default to inputting "the patient reports chest pain for several days."

Laboratory test data preprocessing

The laboratory data comprise 92 test indicators derived from complete blood counts, each represented as a numerical value. To prepare these data for integration with other modalities, we begin by applying min–max normalization to each data point. This step scales the values to a uniform range, ensuring consistency across features. Subsequently, each normalized value is transformed into a 768-dimensional vector using a learnable linear layer. This transformation aligns the laboratory data with the dimensionality of other data modalities, facilitating seamless integration. To handle missing values, we utilize k-nearest neighbors (KNN) imputation techniques.[61] This method estimates and fills in missing values by considering the similarity to other data points, thereby maintaining the integrity and completeness of the dataset.

Demographic data preprocessing

The basic demographic information of the patient includes four elements: age, sex, smoking history, and family lung cancer history. In our experiment, each element is encoded using a process similar to that used for the laboratory test data. Initially, each demographic element is normalized or standardized to ensure consistency across features. For instance, age is scaled to a specific range, while categorical variables like sex, smoking history, and family history are converted into numerical representations using label encoding. Once standardized, each element is transformed into a 768-dimensional vector using a learnable linear layer. This transformation aligns the demographic data with the dimensionality of other data modalities, facilitating multimodal integration. The resulting vectors are then combined into a composite encoded vector.

Stage-1 binary classification model

Given a series of CT images represented as $I = [i_1, i_2, ..., i_n]$ and their corresponding labels $C = [c_1, c_2, ..., c_n]$, we employ the Vision Transformer (ViT) architecture for classification.[34] Each CT image i_j is first segmented into fixed-size nonoverlapping patches of size P $\times$ P. Subsequently, these patches are linearly transformed into vectors of dimension D. The embedding for a given image i_j is mathematically described as:

$$E_{i_j} = W_{emb} \cdot \text{flatten}(i_j) + b_{emb} + p_{emb},$$

where W_{emb} is an embedding weight matrix, b_{emb} is an associated bias, and p_{emb} provides the model with spatial context. Furthermore, a learned class token, denoted as [CLS], is introduced at the beginning of this sequence. This class token is instrumental for the classification procedure. The resulting sequence undergoes processing across several transformer encoder layers. Within these

layers, the relationship between patches is analyzed using a multihead self-attention mechanism,[37] which is mathematically articulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where Q, K and V are the query, key, and value matrices, respectively. Upon progressing through the transformer layers, the output linked to the [CLS] token is subjected to a linear transformation to produce the class logits. This relationship is expressed as:

$$\text{logits}_j = W_{cls} \cdot CLS_j + b_{cls},$$

Finally, these logits are passed through a softmax function to yield class probabilities. During the training process, a cross-entropy loss is computed between logits_j and the label c_j to guide the model's optimization.

Stage-2 multimodal integration model

The LUCID model contains 4 inputs from different modalities, including CT images $I = [i_1, i_2, ..., i_n]$, textual data $T = [t_1, t_2, ..., t_n]$, laboratory test data $L = [l_1, l_2, ..., l_n]$ and demographic information $B = [b_1, b_2, ..., b_n]$. Once we obtain the vector representation of each of the multimodal inputs, the LUCID model integrates them hierarchically (Figure 2).

First, LUCID combines the embeddings of the chief complaints E_{t_j} and the laboratory test data E_{l_j}. This is done by concatenating the two representation vectors to obtain a complete clinical embedding $E_{c_j} = [E_{t_j} : E_{l_j}]$ of the patient's clinical status. Then, the second step is the deep integration of image embedding E_{i_j} and clinical embedding by using the transformer model with a cross-attention mechanism.[44] In this cross-attention process, embeddings from one modality serve as the key within the MultiHead Attention (MHA) mechanism, interacting with the value and query, both of which are sourced from the other modality's embedding.

$$Q_i, K_i, V_i = \text{LP}(\text{Norm}(E_{i_j})),$$

$$Q_c, K_c, V_c = \text{LP}(\text{Norm}(E_{c_j})),$$

where $\text{LP}(\cdot)$ and $\text{Norm}(\cdot)$ represent linear projection and layer normalization, respectively. The calculation of the cross-attention block can be summarized as:

$$E'_{i_j} = \text{Attention}(Q_c, K_i, V_i),$$

$$E'_{c_j} = \text{Attention}(Q_i, K_c, V_c),$$

where $\text{Attention}(\cdot)$ capture the intramodal connections in the CT image and clinical information, respectively, and the cross-attention transformer here is mainly for learning fused mid-level representations by capturing interconnections among tokens from the same modality and across different modalities.

After the three layers of the cross-attention transformer, we concatenate the image embedding E'_{i_j} and clinical embedding E'_{c_j} together and pass the concatenated vector through nine layers of the joint-attention transformer block to continue learning the associations between the internal elements. Finally, we connect the transformer's output vector to the patient's basic demographic information vector E_{b_j} and obtain the model's prediction p through a multilayer perceptron classification layer.

$$p = \text{MLP}\left(\text{MHA}\left(\left[E'_{i_j}, E'_{c_j}\right]\right) + E_{b_j}\right)$$

In the multimodal integration stage, we integrated the four inputs three times hierarchically. This is in fact because the amount of information contained in the different inputs is different so that the model can better learn the features of each modality through this stepwise integration. When computing the loss function, our model not only computes the classification loss L_{joint} after multimodal integration but also computes the classification loss for the image embedding L_{image} and clinical embedding $L_{clinical}$ generated after the cross-attention stage so that the classification effect of each module alone can also achieve a better performance.

$$L = (1 - \alpha - \beta)L_{joint} + \alpha L_{image} + \beta L_{clinical}$$

Baseline models for comparison

Basic Info Only model

The Basic Info Only Model uses a Support Vector Machine (SVM) classifier[45] specifically designed to leverage the demographic details of patients for predictive analytics. Given the demographic vector $E_{b_j} = [E_{b_{j\_age}}, E_{b_{j\_sex}}, E_{b_{j\_smoking}}, E_{b_{j\_family}}]$ of a patient j, the Basic Info Only Model operates by determining the optimal hyperplane defined by the relation

$$w \cdot E_{b_j} + b = 0$$

where w denotes the weight vector and b denotes the bias. The subsequent decision function $f(x) = \text{sign}(w \cdot E_{b_j} + b)$ ascertains the class label, where $f(x) = 1$ may represent the positive results and $f(x) = -1$ the negative results.

Text-only model

The Text Only Model uses a pretrained BERT language model[39] to classify the text derived from patients' chief complaints. Given the textual data vector $E_{t_j} = [E_{t_{j1}}, E_{t_{j2}}, ..., E_{t_{j40}}]$ of a patient j, the Text Only Model applies a multilayer perceptron network[36] to obtain the classification results of the textual data. This network structure enables the efficient classification of the textual data, allowing for the extraction of meaningful insights from the patients' chief complaints, and thereby assists in the formulation of potential diagnosis or treatment strategies.

Lab test-only model

The Lab Test Only Model employs a random forest model[46] to categorize patients based on their laboratory test result vectors $E_{l_j} = [E_{l_{j1}}, E_{l_{j2}}, ..., E_{l_{j92}}]$, which can efficiently process and interpret the vast array of laboratory test result vectors.

Image-only model

We utilized the ViT model structure identical to that in the Stage-I classification task for the Image Only Model. In this approach, the model initially divides the input image $I = [i_1, i_2, ..., i_n]$ into fixed-size, nonoverlapping patches. These patches are then linearly embedded and subsequently processed through a series of transformer layers.[34] This process yields a representation of the image suitable for classification.

Text + lab test model

The Text+Lab Test Model combines two different modalities. Specifically, this model first concatenates data from both modalities to obtain $[E_{tj} : E_{lj}]$. The model then processes the concatenated vectors using the multihead attention mechanism[37] found in the transformer model for computation and classification.

QUANTIFICATION AND STATISTICAL ANALYSIS

The LUCID model was implemented using PyTorch, and the training stage was accelerated using NVIDIA Apex with the mixed-precision strategy. In practice, we can run the LUCID model with one NVIDIA Tesla V100 GPU, and the training process of either task can be performed within 1 day.

When training the Stage-1 classification model, we used 5-fold cross-validation to ensure the robustness of the reported results. The average result over the 5-folds along with the 95% CI will be reported. When training and validating the Stage-2 multimodal integration model, we randomly selected five different random seeds and reported the average result and their 95% CI based on these 5 random seeds. We chose AdamW[62] as our default optimizer after empirical observations indicated superior performance on baseline models and LUCID. The initial learning rate was set at $1 \times 10-5$, with a weight decay of $1 \times 10-2$. Each model underwent training for 50 epochs, during which we employed the CosineAnnealingLR scheduler.[63] For every task in the training phase, we used a batch size of 64.

Survival analysis was performed using Kaplan-Meier estimators, with statistical significance of survival curve separation assessed via log rank tests. Continuous variables are presented as mean ± standard deviation unless otherwise specified.