Disease Continuity Index (DCI): Quantitative Evidence Supporting the "Disease as a Continuous State Space" Hypothesis in Lung Adenocarcinoma

Author: Lin Lin

Department of Respiratory Medicine, The Second Affiliated Hospital of Harbin Medical University

Email: hayidalinlin@163.com

Abstract

Background: Traditional medicine categorizes diseases into discrete labels based on organs or pathology (e.g., lung adenocarcinoma, coronary heart disease, diabetes). However, significant prognostic heterogeneity exists within the same diagnosis, and comorbidity is extremely common—suggesting that these discrete labels may obscure a deeper biological reality. We propose the "disease as a continuous state space" hypothesis: not only is a single disease continuous, but different diseases are also continuous with one another. So-called different diseases are essentially different regions or dynamical bifurcations within the same multi-dimensional continuous space, and comorbidity is a direct manifestation of this continuity. Using lung adenocarcinoma as an example, this study constructs a Disease Continuity Index (DCI) to provide quantitative evidence for this hypothesis and demonstrates a methodological shift from "applying AI to solve established problems" to "using AI to redefine the problems themselves."

Methods: Transcriptomic data from The Cancer Genome Atlas (TCGA) lung adenocarcinoma cohort (n=539) were used. A variational autoencoder (VAE) was first employed for unsupervised extraction of a three-dimensional latent space. Univariate Cox regression identified the dimension most associated with overall survival (z1), and its percentile rank was inverted to define DCI (0–1, with higher values indicating worse prognosis). The overall prognostic value of DCI and its risk stratification ability within Stage I patients were evaluated. The latent space was visualized using t-SNE, and the correlation between DCI and the principal axes was calculated to quantify its continuous gradient.

Results: DCI significantly stratified overall prognosis (log-rank p=0.0008) and identified a high-risk subgroup within Stage I patients (n=290, p=0.0176). Latent space visualization revealed a remarkably strong ordered gradient of DCI along the t-SNE2 axis (Spearman ρ = -0.923, p < 0.001), whereas TNM staging was completely intermingled, showing no natural separation. In univariate Cox analysis, DCI was significant (HR=1.67, p=0.045); after adjusting for stage in multivariate analysis, DCI showed an independent trend (HR=1.67, p=0.06).

Conclusion: DCI exhibits a remarkably strong continuous gradient in the latent space, orthogonal to the staging system, providing quantitative evidence for the "disease as a continuous state space" hypothesis. This index can identify high-risk Stage I patients missed by traditional staging. More importantly, this study offers a novel perspective for understanding comorbidity—comorbidity is

not the coincidental coexistence of multiple independent diseases, but an inevitable manifestation of multiple bifurcation points within the same continuous space. This framework represents a paradigm shift from "applying AI to solve established problems" to "using AI to redefine the problems themselves," laying the foundation for a quantitative, individualized language of disease description that transcends organ- and disease-name-based distinctions.

Keywords: Disease Continuity; Lung Adenocarcinoma; Comorbidity; Variational Autoencoder; Latent Space; TNM Staging

1. Introduction

The modern medical classification system is built upon discrete labels based on organs, pathology, or syndromes—lung adenocarcinoma, coronary heart disease, type 2 diabetes, rheumatoid arthritis, among others. While these labels have played a crucial role in clinical practice, they increasingly reveal fundamental limitations when confronted with the widespread heterogeneity within patient populations and the near-ubiquitous phenomenon of comorbidity.

The prevalence of comorbidity poses the most direct challenge to the discrete disease classification system. If a patient simultaneously has diabetes, hypertension, and coronary heart disease, the traditional view holds this as the coincidental coexistence of three independent diseases. However, clinical reality often shows these conditions occur together, with clear temporal

associations and pathophysiological connections—they are more akin to manifestations of the same metabolic-cardiovascular continuum at different stages or temporal bifurcations, rather than truly independent entities.

We propose a more fundamental hypothesis: Not only is a single disease continuous, but different diseases are also continuous with one another. Diseases are inherently a high-dimensional continuous state space, and discrete diagnostic labels based on organs or disease names are merely human projections onto this continuous space. So-called different diseases may be probability clusters in different regions of the same multi-dimensional space, or bifurcations of the same dynamical process at different critical points. Comorbidity is the direct manifestation of this continuity and these bifurcations.

To test this hypothesis, it is first necessary to demonstrate the continuity of disease states within a single disease category and to construct a quantitative tool capable of capturing this continuous state from molecular data. This study uses lung adenocarcinoma as an entry point, employing a Variational Autoencoder (VAE) on TCGA transcriptomic data to extract a latent space and define a Disease Continuity Index (DCI). We aim to answer: (1) Does DCI reveal a continuous gradient within lung adenocarcinoma? (2) Is this gradient orthogonal to TNM staging? (3) How can this finding inform our understanding of cross-disease continuity and comorbidity?

More importantly, this study seeks to demonstrate a fundamental methodological shift. Most existing artificial intelligence research remains at

the level of "using more complex models to fit existing labels," which essentially reinforces, rather than challenges, potentially problematic foundational classification systems. We argue that the real breakthrough lies in leveraging AI's powerful pattern discovery capabilities, breaking free from the constraints of preset labels in supervised learning, and directly exploring the intrinsic structure and dynamic laws within multi-dimensional patient data. This represents a paradigm shift from "applying AI to solve established problems" to "using AI to redefine the problems themselves."

2. Materials and Methods

2.1 Data Source and Preprocessing

TCGA lung adenocarcinoma (LUAD) RNA-seq data and clinical annotations were obtained. The raw data contained 539 tumor samples and 59,436 features; the first 9 columns were clinical variables (status, stage, stageM, stageN, stageT, gender, age, event, time), and the subsequent 59,427 columns were gene expression values (log2(TPM+1) transformed). Expression data were z-score normalized. Genes with zero variance were removed, retaining 56,654 genes. Missing values (total 237,708, 0.74% of all data points) were imputed with column means. After excluding samples with missing survival data, 526 cases were included in the final analysis. TCGA data were obtained with institutional ethical approval and patient consent.

2.2 Dimensionality Reduction and VAE Training

To reduce computational burden, PCA was first applied to the normalized

expression matrix, retaining 500 principal components (explaining 99.2% of the variance). A VAE was constructed with an encoder and decoder each containing two hidden layers of 256 neurons, and a latent space dimension of 3. The Adam optimizer (lr=0.001) was used with a batch size of 32 for 50 epochs. The model was implemented using PyTorch.

2.3 Definition of the Disease Continuity Index (DCI)

Three-dimensional latent coordinates (z1, z2, z3) were extracted for each patient. Univariate Cox regression for overall survival was performed for each dimension. The dimension with the smallest p-value (z1, p=0.045, negative coefficient) was selected as the prognosis-related dimension. The values of z1 were converted to percentile ranks, and DCI was defined as DCI = 1 – percentile rank, such that DCI $\in$ [0,1] with higher values indicating worse prognosis.

2.4 Latent Space Visualization and Quantification

The 3D latent coordinates were reduced to 2D using t-SNE (perplexity=50, random_state=42) and colored by DCI and by TNM stage. Spearman correlation coefficients between DCI and the t-SNE axes were calculated to quantify the gradient direction.

2.5 Statistical Analysis

Patients were divided into low, medium, and high groups based on DCI tertiles. Kaplan-Meier curves and log-rank tests compared survival differences. Primary analysis focused on Stage I patients, divided into high and low-risk

groups by the median DCI. Exploratory analyses were performed for Stages II–IV. Univariate Cox regression assessed the prognostic value of DCI, age, and numerical stage (stage_num). Multivariate Cox regression included DCI and stage_num to test whether DCI was independent of stage. All analyses were performed using the Python lifelines library, with $p<0.05$ considered significant.

3. Results

3.1 Patient Characteristics

A total of 526 patients were included in the final analysis. Stage distribution: Stage I, 290 (55.1%); Stage II, 121 (23.0%); Stage III, 81 (15.4%); Stage IV, 26 (4.9%). Median follow-up time was not reported.

3.2 Latent Space Quantification: DCI Exhibits a Remarkably Strong Ordered Gradient

In the t-SNE projection, coloring by DCI revealed a smooth vertical gradient (Figure 1A). Spearman correlation analysis showed a remarkably strong negative correlation between DCI and the t-SNE2 axis ($\rho = -0.923$, $p < 0.001$), and only a weak correlation with the t-SNE1 axis ($\rho = -0.274$, $p < 0.001$). This indicates the presence of a continuous axis within the latent space that is highly aligned with DCI and explains the vast majority of its variation. In contrast, coloring by TNM stage showed complete intermingling of all stages with no natural separation (Figure 1B), demonstrating that the staging system is orthogonal to this continuous axis.

3.3 Overall Prognostic Value of DCI

Kaplan-Meier curves based on DCI tertiles showed significant differences in survival among the three groups (log-rank p=0.0008, Figure 2), with the high DCI group having the worst prognosis.

3.4 DCI Identifies a High-Risk Subgroup within Stage I Patients

Within Stage I patients, those with DCI above the median had significantly shorter survival than those below the median (log-rank p=0.0176, Figure 3). This demonstrates that DCI captures prognostic heterogeneity missed by traditional staging. In Stages II, III, and IV, grouping by median DCI did not reach statistical significance (p=0.73, 0.39, 0.30, respectively), possibly due to limited sample sizes.

3.5 Univariate and Multivariate Cox Regression

Univariate Cox analysis showed that DCI was significant (HR=1.67, 95% CI 1.01–2.76, p=0.045), as were age and stage (Figure 4A). After adjusting for stage in multivariate Cox analysis, DCI maintained an independent trend (HR=1.67, 95% CI 0.98–2.83, p=0.06), with a model C-index of 0.68 (Figure 4B).

3.6 Subgroup analysis of DCI in Stage II–IV lung adenocarcinoma patients

In subgroup analyses by stage, the Disease Continuity Index (DCI) did not significantly stratify prognosis in later stages, with log-rank p-values of 0.7316 for Stage II (n=121), 0.3865 for Stage III (n=81), and 0.2980 for Stage IV (n=26) (see Supplementary Material 1 for details).

4. Discussion

4.1 Main Finding: Quantitative Evidence for "Intra-Disease Continuity"

This study is the first to extract DCI from a VAE latent space and demonstrate its remarkably strong continuous gradient (ρ = -0.923), which is orthogonal to the TNM staging system. This finding provides the most direct quantitative evidence to date for the "intra-disease continuity" aspect of the "disease as a continuous state space" hypothesis. DCI not only significantly stratified prognosis in the overall cohort but, more importantly, identified a high-risk subgroup within traditionally "homogeneous" Stage I patients, proving its ability to capture continuous state information missed by staging.

The exceptionally high correlation (ρ = -0.923) between DCI and the t-SNE2 axis has profound theoretical implications: it demonstrates that an unsupervised VAE can spontaneously organize a continuous axis highly concordant with clinical prognosis. This axis is precisely the core dimension of our hypothesized "disease state space." The staging system can partially predict prognosis only because it coincidentally falls upon certain regions of this continuous axis, but it has never revealed the axis's complete structure, nor can it explain the nature of its continuous variation.

4.2 From Intra-Disease Continuity to Inter-Disease Continuity: A Novel Perspective on Comorbidity

The implications of this finding extend beyond lung adenocarcinoma. Expanding our view to encompass multiple diseases allows us to propose a

more ambitious hypothesis: Different diseases are also continuous with one another. So-called different diseases may be distinct regions within the same multi-dimensional continuous space, or bifurcations of the same dynamical process at different critical points. Comorbidity is the direct manifestation of this continuity.

Comorbidity is not the coincidental coexistence of multiple independent diseases, but an inevitable manifestation of multiple bifurcation points within the same continuous state space. Consider metabolic diseases—an individual progressing along a metabolic axis might, at certain critical points, bifurcate into a cardiovascular trajectory (manifesting as coronary heart disease), a renal trajectory (diabetic nephropathy), or a neurological trajectory (cerebrovascular disease). These "different diseases" share the same underlying pathophysiological processes, differing only in their organ-level manifestations. Similarly, in oncology, lung, colorectal, and breast cancers may share core oncogenic pathways (e.g., p53, MYC), and their positions in a latent space might be far closer than we imagine.

The DCI and VAE latent space methodology developed here can be readily extended to cross-disease analysis. By inputting data from patients with multiple diseases (e.g., lung, breast, and colorectal adenocarcinoma) into a VAE, we might discover that these "different diseases" are not discretely separated in the latent space, but rather form a continuous transition, potentially with overlapping regions. This would constitute a direct test of the

"inter-disease continuity" hypothesis.

4.3 Methodological Framework: From "Applying AI to Solve Established Problems" to "Using AI to Redefine the Problems Themselves"

The DCI validated in this study represents the first empirical module of a larger methodological framework. We envision a complete framework consisting of three interconnected discovery modules:

Module 1: Deep Generative Models to Reveal Phenotypic Continuous Spectra (achieved in this study). By compressing high-dimensional molecular data into a low-dimensional latent space using VAEs and validating the continuity of this space using continuous biological indices (like DCI), this approach repositions deep generative models from dimensionality reduction tools to "empirical testers of existing disease classification systems" and "discoverers of fundamental phenotypic dimensions."

Module 2: Temporal Models to Define Disease Evolution "Trajectotypes." Using Neural Ordinary Differential Equations (Neural ODEs) or sequential Transformers on longitudinal follow-up data, each patient's sequence of observations can be modeled as a continuous, interpretable "personal disease trajectory." Unsupervised clustering or manifold learning on these high-dimensional trajectories can identify common dynamic patterns, or "trajectotypes." This shifts the core of phenotypic analysis from cross-sectional states to longitudinal dynamical processes.

Module 3: Representation Learning Integrated with Causal Discovery.

Combining deep representation learning with constraint-based (e.g., FCI) or score-based causal discovery algorithms, we can learn potential causal graph structures among variables on the "purified" feature representations, allowing for unobserved confounders. This module aims to discover relationships not fully described by existing knowledge and identify common latent causes driving multiple phenotypic features, creating a "data-driven hypothesis generation engine" to provide clear targets for subsequent mechanistic biological studies.

This framework represents a paradigm shift from "applying AI to solve established problems" to "using AI to redefine the problems themselves." It provides a systematic toolkit for challenging existing disease paradigms. We are fully aware of the challenges: data quality and heterogeneity, inherent uncertainty in causal discovery algorithms, the translational gap from computational findings to biological mechanisms, and the lengthy process of clinical acceptance. However, by emphasizing a cycle of unsupervised discovery, iterative validation, and deep engagement with domain experts, we aim to gradually build the credibility of this new paradigm.

4.4 Vision: Towards a Quantitative Language of Disease Description Without Organ- or Disease-Name-Based Distinctions

The long-term vision of this framework is not to immediately replace existing classifications, but to accumulate transformative evidence for constructing a completely new, quantitative language for describing disease. We envision that

future disease definitions may evolve into a set of universal, quantitative phenotypic coordinates, for example:

Axis A: Tissue inflammation-repair/fibrosis axis

Axis B: Immune response polarization axis (adaptive vs. innate immunity bias)

Axis C: Organ functional reserve/decline axis

Axis D: Metabolic dysregulation axis (insulin resistance, lipid abnormalities, etc.)

A patient would occupy a dynamic position within this multi-dimensional space and move along a trajectory over time. Current discrete diagnostic categories (e.g., lung adenocarcinoma, coronary heart disease, type 2 diabetes) could be reinterpreted as regions of higher probability density within this continuous space, or as bifurcation points along specific trajectories. Organs and disease names are merely manifestations at the symptomatic level, not the essence of the disease.

If realized, this vision would fundamentally transform medical practice:

Diagnosis: No longer asking "What disease do you have?", but "What are your coordinates in the disease space?"

Treatment: No longer prescribing based on disease name alone, but selecting interventions based on the patient's position and trajectory in the space.

Prevention: Intervening before bifurcations occur by identifying critical points, potentially preventing comorbidity at its root.

4.5 Clinical Significance: DCI Enables Personalized Management of

Early-Stage Lung Adenocarcinoma

Within the limited scope of this study, DCI has already demonstrated direct clinical value:

(1) Risk Stratification and Treatment Decisions

Currently, Stage I lung adenocarcinoma patients typically undergo only observation after surgery, yet approximately 20-30% experience recurrence or metastasis. DCI can help clinicians identify these "high-risk Stage I patients" who might benefit from adjuvant chemotherapy or closer surveillance, while low-DCI patients can safely avoid overtreatment.

(2) Dynamic Monitoring and Recurrence Warning

If gene expression could be monitored post-operatively (e.g., via liquid biopsy) and DCI calculated dynamically, trends in DCI change could provide early warning of recurrence risk. A sustained increase in DCI might indicate active minimal residual disease, prompting earlier intervention rather than waiting for CT-detected new nodules.

(3) Precision Navigation for Drug Development

DCI can precisely select truly high-risk patients for clinical trial enrollment, making it easier to observe drug efficacy and potentially accelerating drug approval and reducing development costs.

(4) Revealing the Biological Essence of Disease

The gene network underlying DCI (the z1 dimension) can be explored. Subsequent analysis identifying genes contributing most to z1 could reveal key

molecular pathways driving lung adenocarcinoma progression, providing novel biomarkers for targeted therapy.

4.6 Challenges and Limitations

As the first empirical module of this framework, this study has several limitations:

Single-Disease Design: Includes only lung adenocarcinoma data, preventing direct testing of the "inter-disease continuity" hypothesis.

Single-Cohort Design: Lacks independent external validation; subsequent validation in datasets like GSE31210 is needed.

Limited Late-Stage Samples: Small sample sizes for Stages III/IV reduced power for subgroup analyses.

Absence of Temporal and Causal Modeling: Modules 2 and 3 require longitudinal data and larger sample sizes for future work.

Lack of Biological Interpretation: The gene network underlying DCI remains to be elucidated.

4.7 Future Directions

Subsequent work will:

Validate DCI in independent datasets such as GSE31210.

Extend the analysis to multiple cancer types (lung, breast, colorectal adenocarcinoma) to test "inter-disease continuity."

Introduce Neural ODEs to model disease trajectories.

Integrate causal discovery algorithms to elucidate driving networks.

Perform enrichment analysis to reveal the biological basis of DCI.

We believe this framework will lay the computational foundation for redefining disease taxonomy, replacing discrete labels with a continuous state space, and ultimately driving a paradigm shift in medicine from a "disease-name-centered" to a "continuous-state-space-centered" approach.

5. Conclusion

We constructed and validated the Disease Continuity Index (DCI), which exhibits a remarkably strong continuous gradient in the latent space ($\rho = -0.923$) orthogonal to the TNM staging system. This provides quantitative evidence for the "intra-disease continuity" aspect of the "disease as a continuous state space" hypothesis. DCI identifies high-risk Stage I lung adenocarcinoma patients missed by traditional staging, potentially enabling personalized management of early-stage disease.

More importantly, this study offers a novel perspective for understanding comorbidity—comorbidity is not the coincidental coexistence of multiple independent diseases, but an inevitable manifestation of multiple bifurcation points within the same continuous space. This framework represents a paradigm shift from "applying AI to solve established problems" to "using AI to redefine the problems themselves," laying the foundation for a quantitative, individualized language of disease description that transcends organ- and disease-name-based distinctions. We invite colleagues to jointly validate and develop this framework, driving medicine towards a new paradigm of

continuous state space.

Acknowledgements

The author thanks The Cancer Genome Atlas (TCGA) Research Network for providing data (https://www.cancer.gov/tcga). The concept of the Disease Continuity Index (DCI) and the "inter-disease continuity" hypothesis were first proposed by the author.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares no conflict of interest.

Data Availability Statement

The TCGA lung adenocarcinoma data used in this study are publicly available from the UCSC Xena platform (https://xenabrowser.net/). Analysis code is available from the corresponding author upon reasonable request.

Ethics Statement

TCGA data were obtained with institutional ethical approval and patient consent. This study complies with relevant ethical requirements.

Figure Legends

Figure 1: t-SNE visualization of the VAE latent space. (A) Colored by DCI, showing a smooth vertical gradient. (B) Colored by TNM stage, showing complete intermingling with no natural separation.

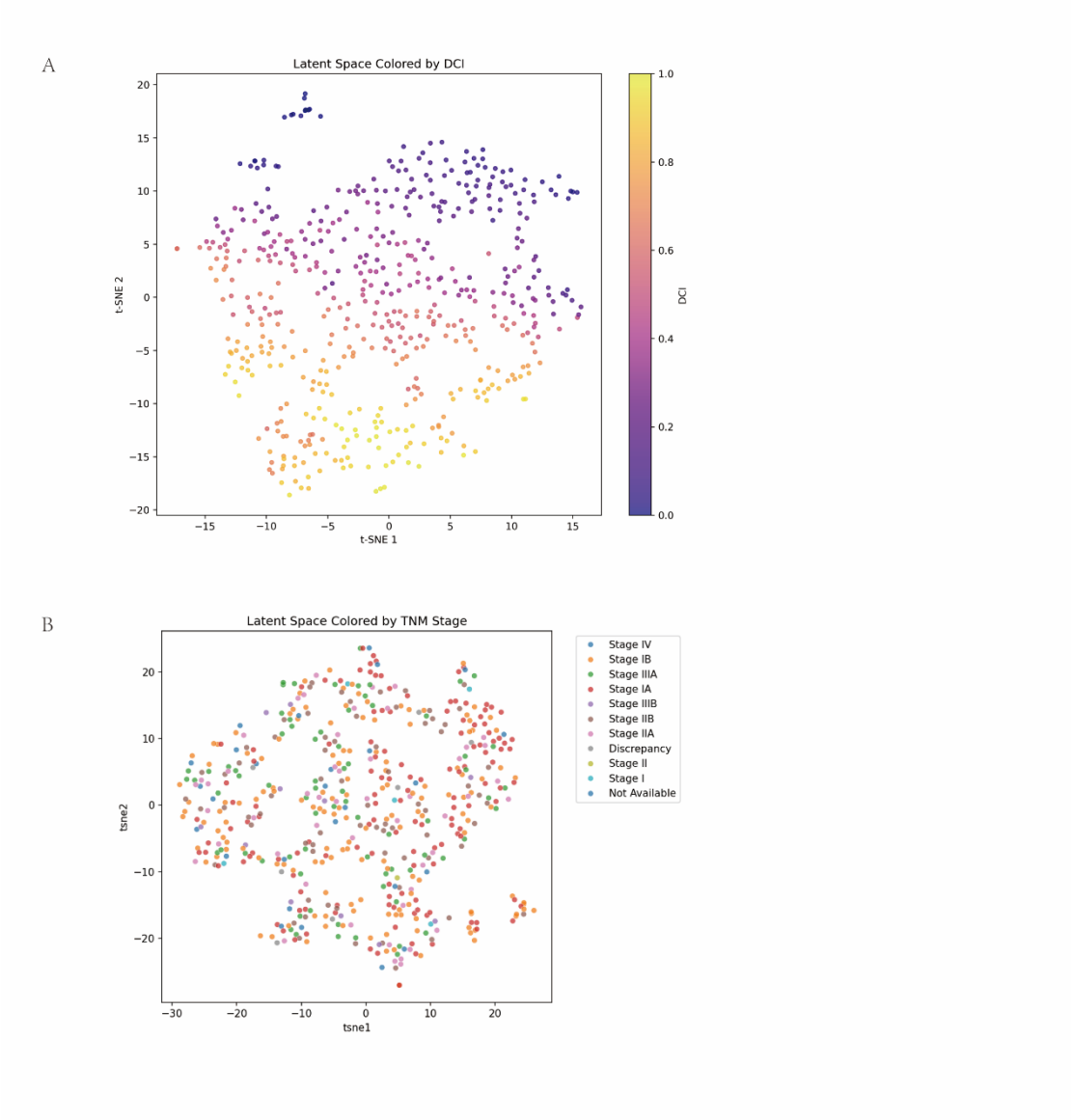


Figure 2: Kaplan-Meier curves for the overall cohort stratified by DCI tertiles (log-rank p=0.0008).

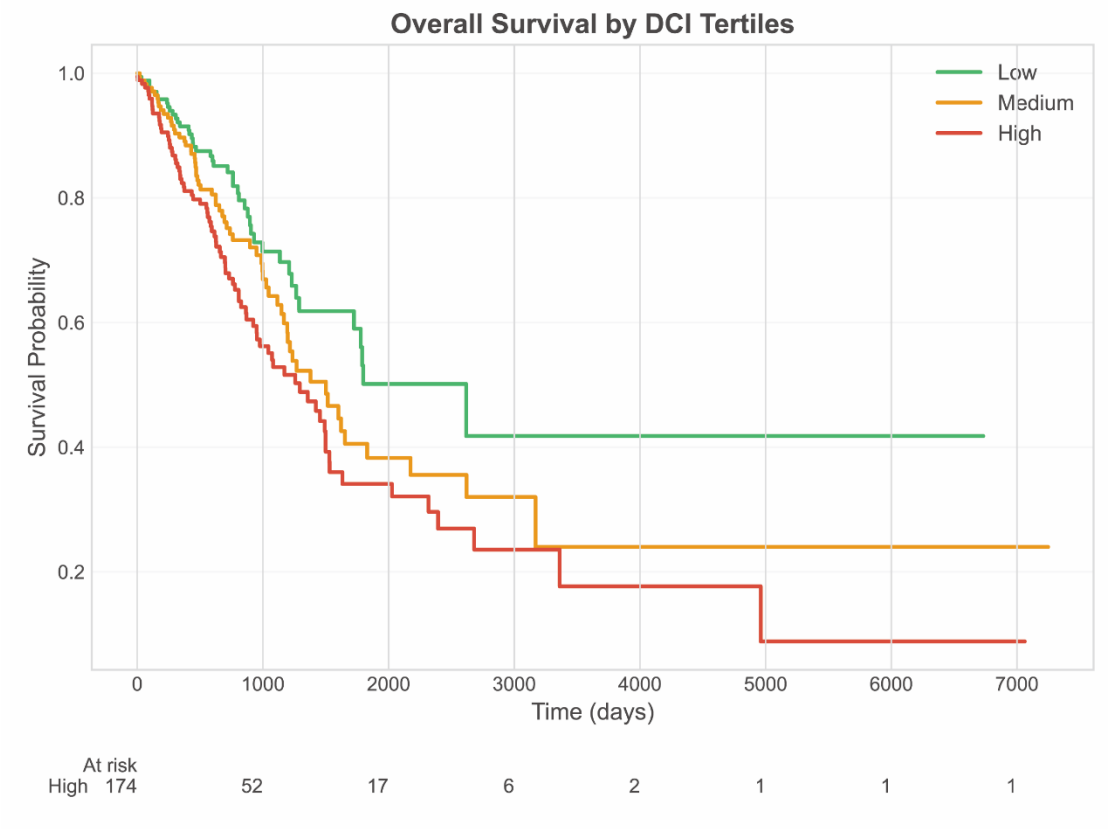


Figure 3: Kaplan-Meier curves for Stage I patients stratified by median DCI (log-rank p=0.0176).

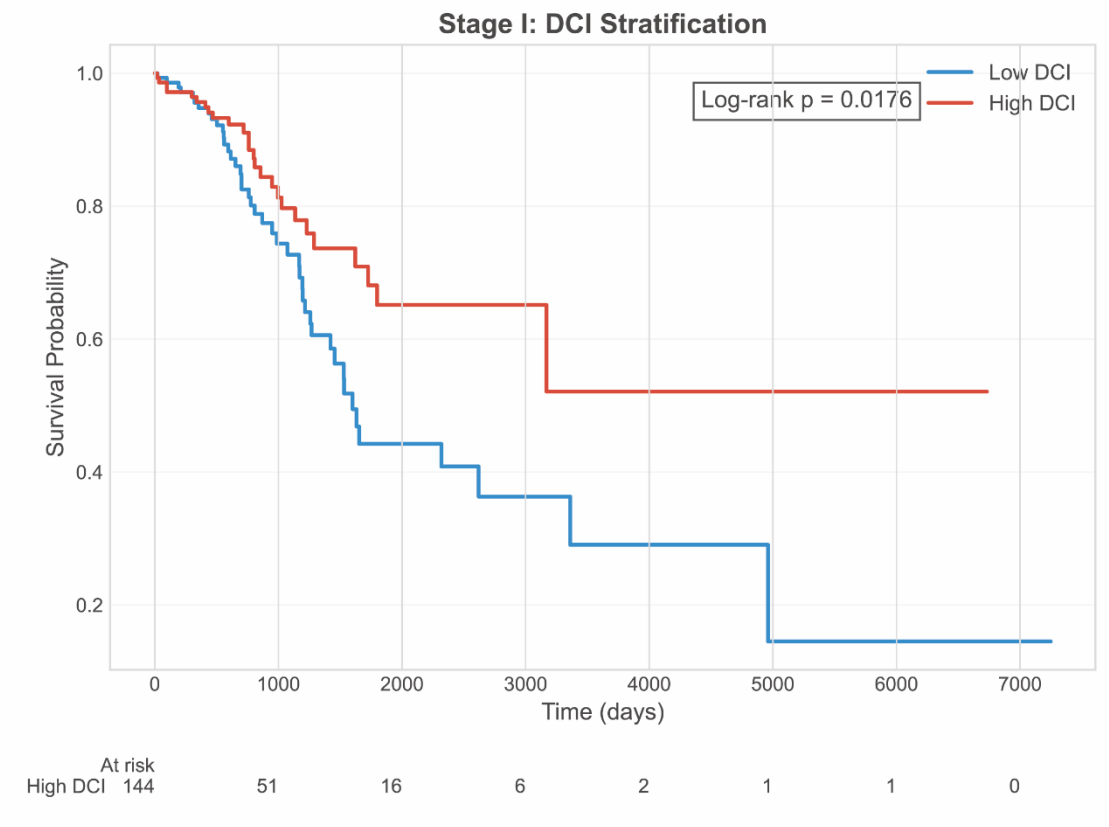
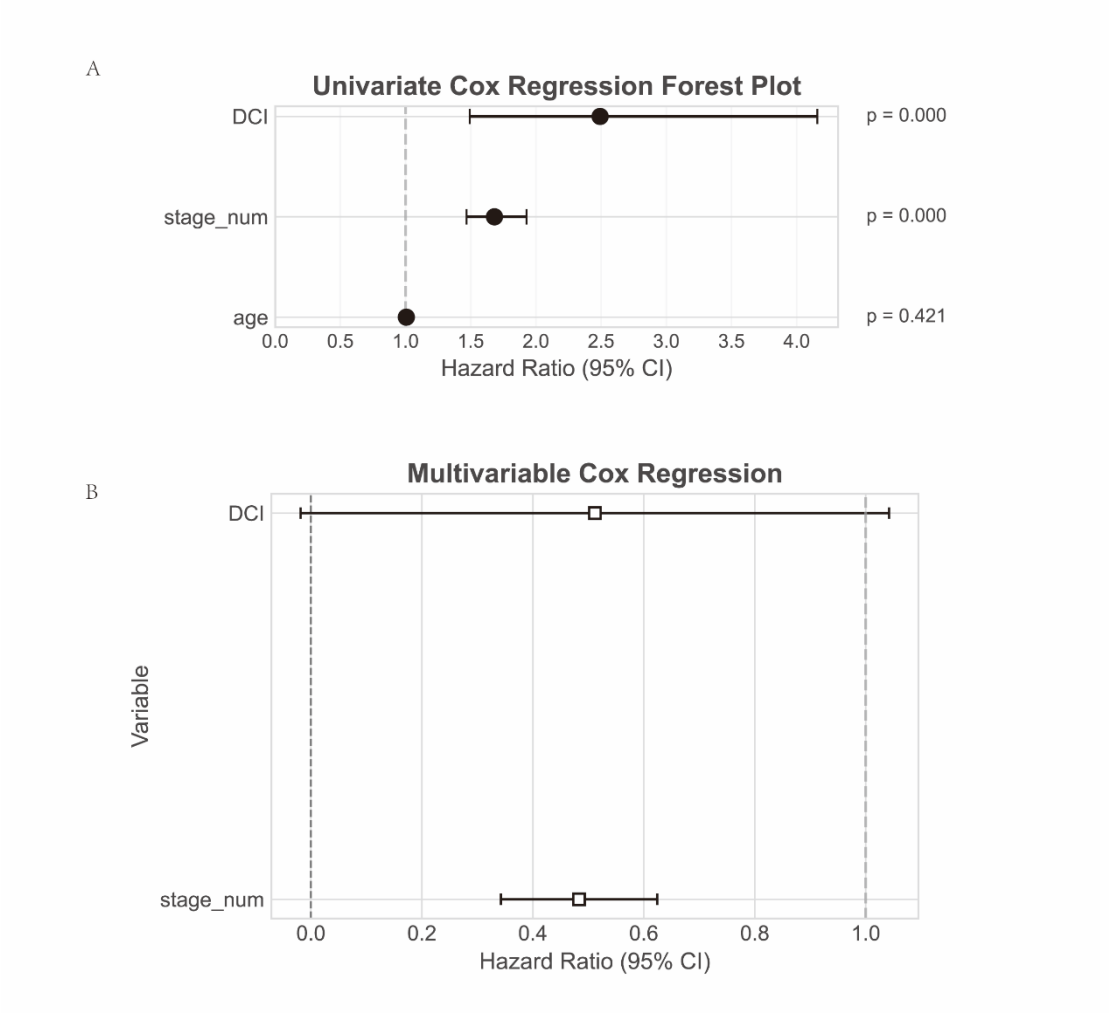


Figure 4: Forest plots from univariate and multivariate Cox regression analyses. (A) Univariate analysis showing hazard ratios for DCI, age, and stage_num. (B) Multivariate analysis showing hazard ratios for DCI and stage_num after mutual adjustment.



Supplementary Table S1: Subgroup analysis of DCI in Stage II–IV lung adenocarcinoma patients.

Stage	Sample Size (N)	Log-rank p-value
Stage II	121	0.7316
Stage III	81	0.3865
Stage IV	26	0.2980