Identification of non-small cell lung cancer subgroups with distinct immuno-therapy outcomes from integrating genomics and electronic health records on a graph convolutional network

Chao Fang[1], Dong Xu[2], Jing Su[3, *], Jonathan Dry[1, *], Bolan Linghu[1, *]

[1] Oncology Research and Early Development, Oncology R&D, AstraZeneca, Boston, Massachusetts, 02451, USA

[2] Department of Electrical Engineering and Computer Science, Christopher S Bond Life Sciences Center, University of Missouri, Columbia, Missouri, 65211, USA.

[3] Department of Biostatistics and Data Science, Wake Forest University, Winston-Salem, North Carolina, 27157, USA.

Chao Fang: chao.fang4@astrazeneca.com

Dong Xu: xudong@missouri.edu

Jing Su: jsu@wakehealth.edu

Jonathan Dry: jonathan.dry@astrazeneca.com

Bolan Linghu: bolan.linghu@astrazenca.com

* To whom correspondence should be addressed.

Jing Su: Email: jsu@wakehealth.edu

Jonathan Dry: Email: jonathan.dry@astrazenca.com

Bolan Linghu: Email: bolan.linghu@astrazeneca.com

Abstract

Recently immuno-oncology (IO) therapies, especially checkpoint inhibitor therapies, have transformed the therapeutic landscape of non-small cell lung cancer (NSCLC). However, responses to IO in NSCLC are highly disparate because patients are heterogenous with a variety of genomic and clinical-phenotype complexity. Thus, there is a pressing need to discover and characterize NSCLC subgroups to advance precision immuno-oncology. However, this is a challenging task largely due to: 1) the study cohort is too small to investigate this heterogeneous disease; 2) the datasets used in subtyping studies are not comprehensive enough to incorporate both genomic data and diverse clinical-phenotype data with long-term follow-ups, and 3) the subtyping algorithms and models are ineffective in integrating high-dimensional data from both genomic and clinical domains. To address these challenges, we have developed a graph convolutional neural network (GCN) method to discover NSCLC complexity on IO treatment responses based on the high-dimensional electronic health records (EHR) and genomic data from 1,937 IO treated NSCLC patients. First, using Flatiron Health's database, we identified a IO treated NSCLC cohort (n = 1,937), with genomic data from Foundation Medicine's targeted DNA deep-sequencing, clinical data from harmonized real-world EHR from 275 US oncology practices, and survival data after IO treatment with a median follow-up time of 6.61 months (average follow-up time 9.11 months). We then developed a GCN based artificial intelligence (AI) model to build a patient-patient similarity network from integrating both genomic and EHR data to discover novel NSCLC subgroups with dramatically different responses to IO therapies. We have demonstrated the performance of the GCN is superior to commonly used machine learning methods such as autoencoder, UMAP, and tSNE, and superior to

utilizing genomic or clinical data alone. Importantly, we have successfully discovered the IO responsive (covers 20.27% of the cohort) and the IO non-responsive (45.46%) subgroups that demonstrate significant overall survival difference after IO treatments (9.42 vs. 20.35 months, $p < 0.0001$). These two subgroups demonstrate enrichments of novel clinical phenotypes and genomic traits beyond well-known IO biomarkers of tumor mutation burden and PDL1 status, such as enrichment of abnormal blood Basophils and KRAS mutations in the responsive subgroup and the enrichment of low hemoglobin, low lymphocytes, PI3KCA amplifications, etc. in the non-responsive subgroup, suggesting distinct clinical and molecular underpinnings. To the best of our knowledge, this is the first study to employ a graph-based AI approach to integrate both high-dimensional clinical and genomic features to investigate IO treatment responses in NSCLC. The new subtypes discovered in this work cast new lights on understanding the heterogeneity of IO treatment responses, and pave ways to inform clinical decision making for precision oncology of NSCLC.

Keywords: artificial intelligence, graph convolutional neural network, machine learning, immuno-oncology, precision oncology, non-small cell lung cancer, disease subtyping

Introduction

According to American Cancer Society[1], non-small cell lung cancer is the second most common cancer in both men and women. About 228,150 new cases of lung cancer were identified in 2019 and about 142,670 deaths were from lung cancer; among them, about 80%-85% of lung cancers are non-small cell lung cancer (NSCLC)[2]. The NSCLC can be adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and other subtypes. The causes of NSCLC are various: 1) about 80% of lung cancer deaths are caused by smoking or exposure to secondhand smoke; 2) lung cancer in non-smokers can be caused by radon, air pollution, asbestos, and/or diesel exhaust; 3) risk factors for lung cancer can caused by certain DNA of lung cells. Besides the complicated causing factors, NSCLC is a heterogeneous disease, which makes the prognosis prediction very challenging. The treatment strategies in NSCLC are very limited, which urgently needs developing tools to determine the survival subgroups.

Some previous works studied the NSCLC heterogeneity among patients to identify the NSCLC molecular subtypes[3-8]. This is because genomics, mRNA profiling, and proteomics have proven promising in characterizing cancer subtypes for personalized therapeutics. For instance, tumor mutational burden (TMB) as determined via genomic profiling has been shown to affect response to nivolumab in NSCLC patients. Real-world-evidence (RWE) based clinical phenotype data such as electronic health records (EHR), which include patient exposures, lab data, diagnosis, medications, and clinical outcomes represent another promising resource for precision oncology, have also been used to identify subtypes[9-14]. However, most existing cancer patient stratification methods use genomics evidence only, without consideration of real-world clinical phenotypes. Conversely, most EHR studies

model clinical phenotypes with respect to outcome (e.g. patient survival), rarely considering the integration of genomics data. The integration of genomics and real-world clinical phenotype evidence is expected to drive to reveal the full landscape of human cancer, therefore enabling novel discoveries towards cancer biology, therapeutics and patient stratification.

Recently, IO treatments, especially checkpoint inhibitor therapies, have significantly improved the overall survival in some NSCLC cases and achieve some great success[15-20]. However, responses to IO are highly disparate, because of the genetic and clinical heterogeneity across NSCLC patients. Subtyping the heterogeneous NSCLC cases for precision immune-oncology remains a challenging problem. For instance, understanding which of these many DNA-seq mutations are important in NSCLC treatment response is not easy to solve. Although studies of NSCLC genomes have implicated several genes as likely crucial mediators for tumor initiation and progression (TP53[21], EGFR[22], CDKN2A/CDKN2B[23], etc.), the experimental validation of the most important, functional genomic changes in NSCLC cells remains a challenge. Also, the available preclinical and clinical results are either too small, lacking important data domains, or follow-up time is too short[24-28], which is also another challenging factor for patient subtyping.

Existing AI work in cancer subtyping and prognostic modeling has some limitations because the subtyping algorithms and models are not effective in incorporating big data from different domains. Previous methods using a grid-based problem formulation[29-31] did not take into consideration the patient-patient interaction as what a graph formulation does. Several traditional machine-learning based methods have been applied to cancer subtyping[32-34]. However, there are some limitations such that those methods may not be

effective to detect high-level distinguishable features from the integrated heterogeneous patient EHR and genomics features for patient subtyping. Some recent applications have utilize deep learning methods such as word embedding[35,36], recurrent neural networks[30,37-39], convolutional neural networks[31] or stacked denoising autoencoders (SDA)[29,40] and demonstrating significant performance for several prediction tasks. However, those models are not trained use a sufficient large amount of patient EHR data to some degree, and hence, the performance of those model may not be sufficient to be used for patient cancer subtyping and prognostic purposes.

To address the abovementioned problems, for the first time, we integrate both genomics and real-world clinical-phenotype evidence using Flatiron Health's longitudinal database. We have built a NSCLC cohort (n = 1,973) for precision IO oncology, with genomic data from Foundation Medicine's targeted DNA deep-sequencing, real-world-evidence (RWE) from harmonized electronic health records (EHR), and survival data after IO treatment. We applied both unsupervised machine learning and data-driven graph methods[41,42]. These two types of methods are complementary and together enable more comprehensive discoveries. Graph-based convolutional neural networks were used to build autoencoder framework as the implementation for integration of both clinical and genomic features. We believe the systematic integration of genomics and real-world clinical phenotype evidence will redefine the NSCLC disease landscape and may revolutionize personalized treatment paradigms. The autoencoders framework learns high-level feature embedding through reconstructing the original input using combinations of nonlinear functions. The learned feature embedding can then be further used for spectral clustering and identify differentiable survival subgroups, i.e. responsive and non-responsive. Several clinical

features enrichment and genomic signaling pathways were also identified, which may be used for directing clinical cancer research. This provides a truly unique opportunity to improve our understanding of how cancer should be segmented for personalized therapies to achieve our goals.

We discovered two subtypes with significant differences in survival, apart from clinical characteristics. We also identified two subtypes have different signaling pathways. Hence, the survival subtype graph-based model proposed here is essential for NSCLC therapeutic intervention, and potentially for some other diseases as well.

Materials and Methods

Datasets and study design

The aim of this study is to develop a data-driven, AI graph-based approach to identify differential survival subgroups across NSCLC IO-treated patients using their clinical phenotype traits and genomic features.

Flatiron NSCLC data set

The patients had NSCLC had genomic testing from January 2010 to October 2018 were selected. The overall Flatiron Health longitudinal EHR-derived database included over 210 cancer clinics representing more than 1.2 million active patients across the United States. The database is updated 3-4 months actively with increasing number of patients added. The clinical data were collected from Flatiron Health, while the genomic data were from Foundation Medicine. The clinical data included in the Flatiron database contains several aspects of tables: demographic, clinical, and outcomes data. This study was conducted in accordance with the Declaration of Helsinki.

IRB not applicable since the analysis uses retrospective data from secondary source.

The inclusion criterium for the experiment was that NSCLC patients diagnosed advanced disease between January 2011 and August 2018. The patients were treated with immunotherapy: Perbrolizumab, Nivolumab, Atezolizumab, or Durvalumab. The end point is using overall survival time starting from the first line of IO treatment. If a patient has deceased, then the decease date was used as the end point date. If a patient has lost his follow-up information, then the last visit date was used to calculate survival months from the 1L IO starting date. For the PDL1 status, the most recent successful test is used, if available. To put another word, if there are two recent successful tests with the same result date, then the positive test is preferred and used.

In this dataset, 1,973 patients were IO treated. The patients are 7.4% African American, 73.7% European white, 2.4% Asian, and 7.4% others as reported. The data were composed of 953 (49.2%) females and 984 (50.8%), and the median age is 69.0 years for overall populations. The overall characteristics are shown in Figure 1. The individuals represented in the clinical dataset are from diverse racial, ethnic, and socioeconomic backgrounds. The EMR data are deidentified, and the study was governed by institutional review board approval and informed consent.

Clinical features and genomic features

To make accurate prediction, it is important to provide useful input features to machine learning methods. Here, a feature vector corresponding to the patient clinical and genomic features, which consists of a rich set of information derived from individual patient.

For clinical feature array contains three parts: 1) Clinical measurements: PDL1 tumor status, PDL1 TIL status, TMB, microsatellite instability (MSI), gender, race, eastern cooperative oncology group (ECOG) performance, group stage, smoking status, ALK

pathological biomarker, BRAF pathological biomarker, EGFR pathological biomarker, KRAS pathological biomarker, ROS1 pathological biomarker, PDL1 pathological tumor cell measurement and PDL1 pathological immune cell measurement and line of therapy. 2) Lab measurements: leukocytes, hemoglobin, platelets, hematocrit, erythrocytes, creatinine, urea nitrogen, alanine aminotransferase, sodium, potassium, aspartate aminotransferase, alkaline phosphatase, albumin, bilirubin, protein, lymphocytes per 100 leukocytes, calcium, lymphocytes, monocytes per 100 leukocytes, glucose, chloride, monocytes, neutrophils, basophils per 100 leukocytes, glomerular filtration rate, basophils, eosinophils per 100 leukocytes, glomerular filtration rate, eosinophils, magnesium, granulocytes per 100 leukocytes, neutrophils, lactate dehydrogenase, and ferritin. The lab features are sorted by frequency and for all IO patients 1,937, only the lab measurements above 800 counts are kept. Otherwise, some lab measurements contain many missing values, which are not included. 3) Vital features contain: body height, body weight and oxygen saturation in arterial blood by pulse oximetry. For current version of Flatiron, the blood pressure measurements contain different units, and not cleaned, and hence, for current implementation, we did not include the blood pressure. We will include it in the future work.

For the genomic features, each patient can have different gene mutation measurements, all patient genes have been measured, and the results are either mutated or not mutated. We filtered out the gene list and kept those genes with at least 50 patients have such gene mutation in the measurement. The reduced list contains following genes (sorted by frequency): “TP53”, “KRAS”, “CDKN2A”, “STK11”, “CDKN2B”, “EGFR”, “PIK3CA”, “LRP1B”, “MYC”, “KEAP1”, “NF1”, “NKX2.1”, “PTEN”, “SMARCA4”, “ARID1A”,

"RBM10", "RB1", "SOX2", "NFKBIA", "CCND1", "FGF3", "FGF4", "FGF19", "BRAF", "MLL2", "ATM", "MDM2", "ERBB2", "TERC", "MET", "SPTA1", "FGFR1", "RICTOR", "MCL1", "DNMT3A", "ARID2", "PRKCI", "FAT1", "ZNF703", "TERT", "APC", "NFE2L2", "FGF12", "MYST3", "FRS2", "TET2", "PTPRD", and "CCNE1". For each patient, the genomic features are represented as a vector of length 48. The value of it is either 1 (mutated) or 0 (not mutated).

The clinical feature vector is concatenated with the genomic feature vector to represent as the feature for each patient.

Problem formulation

The patient subtyping can be formulated as a graph community spatial clustering problem on an undirected graph encoding patient-patient relationships. We consider the patient-patient relationship is represented by a graph with node content as $G = (V, E, X)$ with N nodes (patients) $v_i \in V, i \in [0, N]$, edges connectivity $(v_i, v_j) \in E$, where the edge connectivity can be either 0 (disconnected) or 1 (connected), and $x_i \in X, i \in [0, N]$ is the attribute vector associated with vertex v_i. Each patient has an attribute vector such as clinical features (such as: age; TMB; LDH (Lactate dehydrogenase)) and mutation features (such as EGFR, KRAS, PARK2) as node features. The node features are converted into categorical feature vectors X: the mutation features are binary encoding, i.e. if a patient has that specific gene mutation, the corresponding gene feature is 1; otherwise, 0. For continuous features, we used the high- and low-bound measurement bound provided by the Flatiron and categorize the continuous features into categorical features. For example, a patient has hemoglobin measurement as 8.3, the low- and high-bound for hemoglobin is 14 and 18, respectively. It falls between two bounds, which indicate a "normal" class. The

two nodes are connected if the node feature vectors are similar. Here we applied cosine similarity to be the measurement to define similarity. If cosine similarity is less than 0.5, then there is not a link connected between two nodes; otherwise, connected. So formally, the graph can be represented by two types of information, the patient content information $X \in R^{n \times d}$ and the structure information $A \in R^{n \times n}$, where A is an adjacent matrix of G and $A_{i,j} = 1$ if $e_{i,j} \in E$ otherwise, 0.

Given a patient-patient graph G, patient subtyping is to partition the patient nodes from G into k disjoint subgroups $\{G_1, G_2, \dots, G_k\}$ so that: (1) the patient nodes within the same cluster have similar clinical outcome (survival) to each other than patient nodes in different clusters in terms of graph structure; (2) the patient nodes within the same subgroup are more likely to have similar clinical and genomic attribute values.

Deep feature representation for graph clustering

It is beneficial to formulate the patient-patient relationship into a graph since both the node content (patient clinical and genomic features) and structure interaction (patient-patient connectivity) will be integrated and used. To fully extract and have deep feature representation, we apply the marginalized graph autoencoder (MGAE) method [43] to exploit the patient network information.

The MGAE is based on graph convolutional network (GCN) [44] and to learn the convolution feature representation on the structure information with node content in the spectral domain. The MGAE extends GCN to a purely unsupervised clustering task. MGAE can exploit the interplay between node content and graph structure information by using a marginalization process, which is to encode content features of graph into the deep

learning framework. MGAE also uses stacked graph convolutional network for learning deep graph representation for clustering. [43]

Graph convolutional network

Graph convolutional network (GCN) [44] applies the convolution operation on a graph from the spectral domain. Given the adjacency matrix A and content matrix X of a graph, the spectral convolution function used to calculate layer-wise transformation is defined as:

$$Z^{(l+1)} = f(Z^{(l)}, A)$$

Here, $Z^{(l)} \in R^{n \times d}$ (n nodes and d features) defines the input for layer l. The input layer contains the patient clinical and genomic feature matrix for our problem. The feature dimension of input layer is 227, which was derived from original 100 features. Our graph model has three hidden layers and the embedding dimension is same as input layer 227. MGAE embedding method reconstructs the feature matrix of node without hidden layers.

GCN [44] applies Chebyshev polynomials [45] to approximate the convolution filter. The layer-wise propagation rule for GCN can be then defined as:

$$f(Z^{(l)}, A) = \sigma(DZ^{(l)}W)$$

Here, D is the degree matrix for A. W is the learnable weights. $\sigma(\cdot)$ is an activation function such as ReLU [46].

Marginalized graph autoencoder (MGAE)

The MGAE [43] is a content and structure augmented autoencoder. MGAE reconstructs the input $X = \{x_1, \dots, x_n\} \in R^{n \times d}$ by using a single mapping function $f(\cdot)$, that minimizes the squared reconstruction loss:

$$\|X - f(X)\|^2$$

For graph convolution networks, the loss function becomes:

$$\|X - DXW\|^2 + \lambda\|W\|_F^2$$

Here, D is the degree matrix for A. W is the parameter matrix. $\|W\|_F^2$ is a regularization term with coefficient λ being a tradeoff.

The marginalized graph autoencoder provides an effective way to integrate both content and structure information. To encourage the interplay between content and structure information, MGAE introduces some random noises into the content features during training. The corruption process can be randomly removing some features or setting them to 0. Given the corrupted version of the original input X, the corrupted version of original input X is:

$$\tilde{X} = \{\widetilde{x_1}, \dots, \widetilde{x_n}\}$$

The objective function becomes:

$$\frac{1}{m}\sum_{i=1}^{m}\left\|X - D\widetilde{X}_i W\right\|^2 + \lambda\|W\|_F^2$$

And the final graph embedded representation Z is defined as:

$$Z = \hat{A}XW$$

Patient subtype clustering with MGAE

We applied the spectral clustering algorithm [43] for patient subtyping. The symbol used and pseudo-code is defined as follows:

Given the patient graph network G with n nodes, each patient node is a d-dimension attribute vector. The patient attribute matrix $X \in R^{n \times d}$ of G, the total number of patient subtypes k, the corruption probability p, and the number of stacked autoencoder layers Γ. In our problem formulation, $\Gamma = 2$. $Z^{(0)} = X$ is the input to the first layer.

Step-1:

For $l = 1$ to Γ

Construct a single layer denoise autoencoder with input data $Z^{(l-1)}$

Learn the autoencoder output representation $Z^{(l)}$ according to

$$Z = \hat{A}XW$$

Step-2:

$$Z_0 \Leftarrow Z^{(\Gamma)}$$

$$Z_1 \Leftarrow Z_0 {Z_0}^{\mathrm{T}}$$

$$Z_2 \Leftarrow \frac{1}{2}(|Z_1| + |Z_1^T|)$$

Step-3:

Run spectral clustering on Z_2

Interpretable clinical patient similarity networks

Formulating the patient-patient relationship as a graph-based model can address many challenges in analysis and is naturally interpretable. In the network design, each node represents an individual patient and an edge between two patients represents pairwise similarity. In terms of individual patient, each pair with similar characteristics can then be tightly connected to each other. In terms of a subgroup of patients, i.e. clusters, each subgroup can enrich for different comorbidities and biological meanings, such as survival, adverse event and/or drug dose.

NSCLC-specific patient graph network overall workflow

The patient clinical features are from electronic medical records and genomic features are from Foundation Medicine. We collected and filtered out patients treated by immune therapy. There are total 1,973 patients in the cohort.

We designed and applied an unsupervised, graph based neural network that uses both patient clinical feature and genomic feature to infer a patient-patient similarity network as the computational model to represent a complex patient population. The overall method workflow is shown in Figure 1. We first applied marginalized graph autoencoder (MGAE) [43] to get graph representation encoding both patient network graph structure information and patient node clinical and genomic features. Then, we applied two-layer stacked graph convolutional network to extract high level feature representation.

After getting the latent space representation, the spectral clustering algorithm [43] was applied on the patient graph network to generate graph clusters. Those clusters are patient subgroups. For each subgroup, we performed survival analysis and detect potential biomarkers for each subtype.

Patient subtype survival analysis

We used the Kaplan-Meier (KM) estimate [47] to measure the fraction of subgroup patient drug response. In clinical trials, the effect of an intervention is assessed by measuring the number of patients survived or saved after that intervention over a period of time. The starting time in our case is the first IO treatment starting time for patients, and the patient deceased time is used to be the event occurrence date in the survival analysis. In Flatiron database, there are also patients under study that were uncooperative and refused to be

remained in the study or when some of the patients may not have experienced the event or death before the end of the study, although they would have experienced or died if observation continued, or were lost in touch midway in the study. We labeled these situations as censored observations. The statistical difference between different groups of patients of KM estimate could be used to study the drug response of different groups.

Signaling pathway analysis

We hypothesized that genetic abnormalities alter the activities of key signaling pathways and consequently lead to the differential survivals observed between the responsive and the non-responsive subtypes. We further assumed that a single functional genetic alteration of a determinative gene in such key signaling pathways is sufficient to impact patients' responses to IO. Key signaling pathways thus could be revealed by searching for the combinations of such determinative genetic alterations. Mathematically, such a combination of alterations satisfies the following criteria: 1) Differentially enriched in responsive vs. non-responsive subtypes. That is, patients in the responsive subgroup are more likely to have *at least one* of these genetic alterations, and the Non-responders are more likely to host *none* of such alterations, or vice versa. 2) necessary. If any of the genetic alterations is *removed* from the combination, the differential enrichment of these genetic alterations in the two subtypes will become weaker. 3) Complete. If any other genetic alteration is *added* into the combination, the differential enrichment will not turn stronger. Genetic alterations considered in this work included functional mutations (gain or loss of biological functions), copy number amplifications, and copy number deletions. The differential enrichment of the combination of genetic alterations in responsive or non-responsive subtypes was measured by the p-values of Fisher's exact test. The optimal

combinations were discovered by greedy search algorithm. The possible signaling pathways related with the discovered differentially enriched genetic alterations were inferred using Ingenuity Pathway Analysis[48] (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis).

Results

Cohort characteristics

The baseline demographic and pathologic characteristics is shown in Figure 2. Figure 2a shows how the cohort was identified with inclusion and exclusion criteria. Figure 2b shows the cohort characteristics. There are totally 1,937 patients in the cohort. The median age is 67 and about half of the population is female. We also report the race, histology, stage, ECOG value, smoking status, and previous treatment in the statistics. All the patients in this cohort were treated by immunotherapy (IO), which means the patients were treated by one of the IO drugs: Perbrolizumab, Nivolumab, Atezolizumab, or Durvalumab. Figure 2c shows the clinical features, such as hemoglobin, erythrocytes, hematocrit, etc. They can be classified into molecular pathological features, blood test, and demographic behavioral and vital pathologic features. Figure 2d shows the waterfall plot for gene mutation. Each row represents a gene and each column represents a patient. The mutation type can be SV (structural variant), CN (copy number variations), and RE (rearrangement). From Figure 2d, we found there are seven genes found most frequency of mutation along the NSCLC patient cohort. The potentially actionable somatic mutations found in this study is consistent with prior studies[49]. The EGFR and KRAS were the most commonly identified oncogenic drivers and were with very rare exception mutationally exclusive. CDKN2A and

CDKN2B have copy number mutation patterns. Comparing with other two TCGA lung cancer comprehensive paper [50,51], our finding is consistent with theirs. Cancer Genome Atlas Research Network [50] found statistically recurrent mutations in 11 genes, including mutation of TP53 in nearly all specimens. Significantly altered pathways included NFE2L2 and KEAP1 in 34%, squamous differentiation genes in 44%, phosphatidylinositol-3-OH kinase pathway genes in 47%, and CDKN2A and RB1 in 72% of tumors [50]. EGFR mutations were more frequent in female patients, where mutations in RBM10 were common in males [51].

Patient subgroups (clusters)

The NSCLC patients were clustered into five subgroups using the spectral algorithm. Figure 3a shows the tSNE plot for patients assigned to five clusters. The survival plots for each subgroup is shown in Figure 3b. A total of 1,937 patients NSCLC IO treatment patients were selected to build the patient cohort in this experiment. The responsive and non-responsive subtype groups were selected by our graph-based method. The responsive subtype group has better overall survival than the non-responsive subtype group. There are 400 patients in the selected responsive subtype group while 897 patients in the non-responsive subtype group. Figure 3c shows how graph based the convolutional graph network outperforms other best machine learning methods such as autoencoder, UMAP, etc. Specifically, we compared MGAE clustering methods with other clustering methods, which are: tSNE [52] + clustering, UMAP [53] + clustering, autoencoder + clustering, denoise autoencoder + clustering, and MGAE without denoising, and MGAE. We used a volcano plot and KM plot to visualize and evaluate how good the clustering algorithm can perform on patient subtyping tasks. The purpose is to find out whether the denoising process can

contribute to the feature learning and eventually for patient subtyping. Also, it is important to validate whether the problem formulation makes a difference, that is formulating the patient-patient relationship as graph vs. a plain matrix. We can observe that graph-based MGAE approach problem formulation has the best patient subtyping among other methods. The volcano plots and KM survival analysis shows that the responsive (green curve) and non-responsive (red curve) have a significant difference. The other methods do not have a good way to stratify the subgroups into responsive or non-responsive, which can be observed both from volcano plots. The reason behind this is that patient EHR contains high-dimensional, heterogeneous and incomplete features and a graph-based model can effectively describe such complex patient network system. Graph-based approach favors over grid-based approach in terms of leveraging the rich patient clinical information in graph-structure data and learn effective node or graph representation from both node/edge attributes and the graph topological structure. Figure 3d shows that both clinical and genomic features show better clustering results comparing to use just one feature alone. Also, k=5 gives good clustering results that can better stratify responsive and non-responsive subtypes comparing to choosing other parameters. Figure 3e shows the five subtypes identified by MGAE. For each cluster, the K-M estimation will get the median survival time. Same for the overall cohort. From that figure, the clinical important clusters can be selected for further analysis.

Annotation describes molecular feature and clinical feature

The subgroups detected can contain potential biomarkers. The GCN successfully detect several subgroups from IO cohort. The KM plots were generated for each subgroup. The

responsive group represents the patients with good survival prognosis, whereas the non-responsive group represents the patients with bad survival prognosis. (Figure 3b and Figure 3e). After merging the two non-responsive groups, (since their corresponding survival prognosis are very close), Figure 4 shows the clinical and genomic features enrichment in responsive and non-responsive groups. The enrichment was determined by the greater number of occurrences in responsive or non-responsive subgroups. From Figure 4, an observation is that some blood measurements abnormality may be act as biomarker for future study as they are differential between responsive and non-responsive subgroups.

Two genetic alteration combinations were identified (Supplemental Table 1), one enriched in responsive subtype (genetic abnormalities were highlighted in green shade) and the other in non-responsive subtype (in brown shade). Genes specific to responsive subtype, non-responsive subtype, or shared by both subtypes were presented in green, brown, and yellow colors. Interestingly, functional mutations (in 6 genes) and copy number amplifications (in 7 genes) dominated the 12 genes specific to the responsive subtype and copy number deletion (in 1 gene) was rare. In contrast, no obvious pattern of genetic alteration types were found in Non-responder-specific genes or shared genes.

The landscapes of the genetic alterations in responsive and non-responsive subtypes were further illustrated by the inferred underlying signaling pathways (Supplemental Figure 4). The responsive group demonstrated a classical oncogenic signaling axis from EGFR1/HER2 over-activation to the KRAS and BRAF signaling hubs to the uncontrolled proliferation featured with p15/p16, cyclins, and RB1. In contrast, the non-responsive subtype was dominated by DNA damage repair mechanisms, demonstrating the deletions of STK11 and p53 genes and the functionally active mutation of the TERT gene. Both

subtypes shared a common cMET/FGFR – PI3K/AKT pathway but with different activation mechanisms. For example, The activation of cMET and the deactivation of PTEN were achieved through copy number variations in responsive group, but through functional mutations in non-responsive patients. The activation of the FGF pathway was by the copy number amplification of the FGF12 ligand or the FGFR1 receptor in Non-responders, but by the functional mutation of FGFR1 in Responders. In summary, while both subtypes seemly were driven by the common PI3K/AKT pathway, the responsive subtype demonstrated enhanced proliferation, while the non-responsive subtype exceled in fast genetic evolution through the malfunctioned DNA damage repairing mechanisms.

Figure 5a shows when TMB[54] is in intermediate level (We applied the classification standard as what Foundation Medicine Inc used)[54], the GCN approach can further stratify patients into responsive and non-responsive subtypes. In contrast, when the PDL1 expression level is high, both responsive and non-responsive subtypes showed similar good responses to IO treatment. Our findings are that GCN subtyping can provide novel IO stratification beyond TMB and PDL1. We found GCN can stratify IO-treated patients with low or intermediate TMB into subtypes with distinct clinical outcomes. When TMB is in low or intermediate level, the GCN can further stratify patients into responsive and non-responsive subtypes. This shows that for traditional biomarker TMB, if it labels a patient not suitable for IO treatment, our approach can be based on that to further suggest subtypes for new subtypes and for precision treatment. In our dataset, above 90% patients do not have PDL1 measurements, which suggest that under such situation, our method can perform patient subtyping independent on traditional biomarkers.

Genomic alternations that are known to boost EGFR/KRAS signaling were examined for their effects on survival. These genomic variants include EGFR mutation, KRAS mutation, or KRAS copy number amplification. In the Responsive Subtype (Figure 5c), patients with any of these genomic variations demonstrated significantly worse survival (14.4 months vs 23.0 months, log-rank test p-value <= 0.01). Meanwhile, no significant survival difference was observed in the overall IO cohort (Figure 5d). Further examination on EGFR mutations (Supplemental Figure 1a) or KRAS mutations/copy number amplifications (Supplemental Figure 1c) alone also showed no correlation with patients' overall survival in the whole IO cohort.

The effects of BRAF genomic alteration types on patients' response to IO treatment in the overall IO cohort were examined. Patients with BRAF mutations and recombinations responded better than patients without BRAF genomic abnormality (Supplement Figure 3), while copy number amplification of BRAF did not affect survival.

Discussion

We carried out our study based on a large number of patients with immune therapy treatment. A computational practical pipeline was built to detect patient subtypes and findings in our data. The reason why our approach works well on patient graph-based problem is that GCN can automatically learn a low-dimensional feature representation for each node in the graph. The low-dimensional representations are learned to preserve the structural information of graphs, and thus can be used as features in building machine learning models for various downstream tasks, such as clustering (here, patient subtyping). The graph-based approach compared with grid-based approach can better utilizing the

structures within the patient clinical and genomic data. The graph convolutional operation can effectively include the entire graph information and embed the local patient information. It is a more global approach of using the entire data.

Five clusters are buried in noises in the raw data. Or, more formally, due to the intrinsic relationship between samples in the raw data, these samples, when projected to 5 clusters using different approaches, show consistent patterns across these approaches. Other approaches can recognize the existence of these clusters, but due to the noise in the raw data, these methods cannot effectively distinguish which sample belongs to which cluster. Therefore, clusters identified by other approaches are not pure enough and thus fail to show clinical importance (i.e., not good enough to pass statistic tests).

When working on real-world EMR data, since the noises are in the raw data, traditional cluster purity metrics are not suitable for evaluating the performance of an approach. Therefore, we used a pragmatic metric to compare the performance of our approach with others: whether an approach can effectively identify clinically meaningful subtypes. Meanwhile, we use concordance analysis to check whether our findings are artifacts.

In the results section, we found some interesting clinical and genomic findings that can be used to stratify patient subtypes, which can be used to describe associations between driver mutations and response to targeted therapy. These findings demonstrate the powerfulness of deep learning and can provide support for further research and discovery evaluation this approach in oncology.

There are some limitations in this study. First of all, the overall survival analysis is based on the patient data collected from routine clinical practice. Not every deceased patient's data of death was captured, some were imputed using "last day of visit". Also, the

unknown distribution of missingness is also a factor. Furthermore, the time receiving therapy end point may not account for non-progression-related reasons for discontinuing therapy.[55]

References

1 Society, A. C. *Key Statistics for Lung Cancer*, <https://www.cancer.org/content/cancer/en/cancer/lung-cancer/about/key-statistics.html> (2019).

2 Society, A. C. *What is Lung Cancer*, <https://www.cancer.org/cancer/lung-cancer/about/what-is.html> (2019).

3 Pikor, L. A., Ramnarine, V. R., Lam, S. & Lam, W. L. Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung cancer* **82**, 179-189 (2013).

4 Thomas, A., Liu, S. V., Subramaniam, D. S. & Giaccone, G. Refining the treatment of NSCLC according to histological and molecular subtypes. *Nature reviews Clinical oncology* **12**, 511 (2015).

5 Wislez, M. *et al.* Non-mucinous and mucinous subtypes of adenocarcinoma with bronchioloalveolar carcinoma features differ by biomarker expression and in the response to gefitinib. *Lung cancer* **68**, 185-191 (2010).

6 Kim, H. S. *et al.* Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer. *Cell* **155**, 552-566 (2013).

7 Timms, K. M. *et al.* Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. *Breast Cancer Research* **16**, 475 (2014).

8 Bergamaschi, A. *et al.* Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer* **45**, 1033-1040 (2006).

9 Miller, V. A. *et al.* Bronchioloalveolar pathologic subtype and smoking history predict sensitivity to gefitinib in advanced non-small-cell lung cancer. *Journal of Clinical Oncology* **22**, 1103-1109 (2004).

10 Shim, H. S., Lee, D. H., Park, E. J. & Kim, S. H. Histopathologic characteristics of lung adenocarcinomas with epidermal growth factor receptor mutations in the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society lung adenocarcinoma classification. *Archives of pathology & laboratory medicine* **135**, 1329-1334 (2011).

11 Beaulieu-Jones, B. K. & Greene, C. S. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of biomedical informatics* **64**, 168-178 (2016).

12 Shinagare, A. B. *et al.* Unsuspected pulmonary embolism in lung cancer patients: comparison of clinical characteristics and outcome with suspected pulmonary embolism. *Lung cancer* **78**, 161-166 (2012).

13 Bepler, G., Neumann, K., Holle, R., Havemann, K. & Kalbfleisch, H. Clinical relevance of histologic subtyping in small cell lung cancer. *Cancer* **64**, 74-79 (1989).

14 Dai, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research* **5**, 2929 (2015).

15 Reck, M. What future opportunities may immuno-oncology provide for improving the treatment of patients with lung cancer? *Annals of oncology* **23**, viii28-viii34 (2012).

16 Kim, E. S. *et al.* The BATTLE trial: personalizing therapy for lung cancer. *Cancer discovery* **1**, 44-53 (2011).

17 Jeremic, B., Shibamoto, Y., Acimovic, L. & Milisavljevic, S. Initial versus delayed accelerated hyperfractionated radiation therapy and concurrent chemotherapy in

limited small-cell lung cancer: a randomized study. *Journal of clinical oncology* **15**, 893-900 (1997).
18 Gamerith, G., Kocher, F., Rudzki, J. & Pircher, A. ASCO 2018 NSCLC highlights—combination therapy is key. *memo-Magazine of European Medical Oncology* **11**, 266-271 (2018).
19 Elias, A. *et al.* Neoadjuvant therapy for surgically staged IIIA N2 non-small cell lung cancer (NSCLC). *Lung cancer* **17**, 147-161 (1997).
20 Jotte, R. M. *et al.* IMpower131: Primary PFS and safety analysis of a randomized phase III study of atezolizumab+ carboplatin+ paclitaxel or nab-paclitaxel vs carboplatin+ nab-paclitaxel as 1L therapy in advanced squamous NSCLC. *J Clin Oncol* **36**, LBA9000 (2018).
21 Mogi, A. & Kuwano, H. TP53 mutations in nonsmall cell lung cancer. *BioMed Research International* **2011** (2011).
22 Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497-1500 (2004).
23 Suzuki, H. *et al.* Intragenic mutations of CDKN2B and CDKN2A in primary human esophageal cancers. *Human molecular genetics* **4**, 1883-1887 (1995).
24 Camidge, D. *et al.* Progression-free survival (PFS) from a phase I study of crizotinib (PF-02341066) in patients with ALK-positive non-small cell lung cancer (NSCLC). *J Clin Oncol* **29**, 2501 (2011).
25 Tagalakis, V. *et al.* High risk of deep vein thrombosis in patients with non-small cell lung cancer: a cohort study of 493 patients. *Journal of Thoracic Oncology* **2**, 729-734 (2007).
26 Boch, C. *et al.* The frequency of EGFR and KRAS mutations in non-small cell lung cancer (NSCLC): routine screening data for central Europe from a cohort study. *BMJ open* **3**, e002560 (2013).
27 Velcheti, V. *et al.* Programmed death ligand-1 expression in non-small cell lung cancer. *Laboratory investigation* **94**, 107 (2014).
28 Spoerke, J. M. *et al.* Phosphoinositide 3-kinase (PI3K) pathway alterations are associated with histologic subtypes and are predictive of sensitivity to PI3K inhibitors in lung cancer preclinical models. *Clinical cancer research* **18**, 6771-6783 (2012).
29 Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports* **6**, 26094 (2016).
30 Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. in *Machine Learning for Healthcare Conference.* 301-318.
31 Nguyen, P., Tran, T., Wickramasinghe, N. & Venkatesh, S. $\mathtt {Deepr} $: a convolutional net for medical records. *IEEE journal of biomedical and health informatics* **21**, 22-30 (2016).
32 Lu, C.-F. *et al.* Machine learning–based radiomics for molecular subtyping of gliomas. *Clinical Cancer Research* **24**, 4429-4436 (2018).
33 Sherafatian, M. Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene* **677**, 111-118 (2018).
34 Huang, S. *et al.* Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics* **15**, 41-51 (2018).
35 Choi, E. *et al.* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1495-1504 (ACM).
36 Choi, Y., Chiu, C. Y.-I. & Sontag, D. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings* **2016**, 41 (2016).

37 Winslow, M. M. *et al.* Suppression of lung adenocarcinoma progression by Nkx2-1. *Nature* **473**, 101 (2011).
38 Choi, E. *et al.* in *Advances in Neural Information Processing Systems.* 3504-3512.
39 Lipton, Z. C., Kale, D. C. & Wetzel, R. Modeling missing data in clinical time series with rnns. *arXiv preprint arXiv:1606.04130* (2016).
40 Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* **8**, 6085 (2018).
41 Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine* **7**, 311ra174-311ra174 (2015).
42 Kaplan, A. & Lock, E. F. Prediction with dimension reduction of multiple molecular data sources for patient survival. *Cancer informatics* **16**, 1176935117718517 (2017).
43 Wang, C., Pan, S., Long, G., Zhu, X. & Jiang, J. in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 889-898 (ACM).
44 Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
45 Hammond, D. K., Vandergheynst, P. & Gribonval, R. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis* **30**, 129-150 (2011).
46 Nair, V. & Hinton, G. E. in *Proceedings of the 27th international conference on machine learning (ICML-10).* 807-814.
47 Bland, J. M. & Altman, D. G. Survival probabilities (the Kaplan-Meier method). *BMJ* **317**, 1572, doi:10.1136/bmj.317.7172.1572 (1998).
48 Kramer, A., Green, J., Pollard, J., Jr. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523-530, doi:10.1093/bioinformatics/btt703 (2014).
49 Jordan, E. J. *et al.* Prospective comprehensive molecular characterization of lung adenocarcinomas for efficient patient matching to approved and emerging therapies. *Cancer discovery* **7**, 596-609 (2017).
50 Network, C. G. A. R. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519 (2012).
51 Network, C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543 (2014).
52 Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
53 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
54 Schrock, A. *et al.* 1170PAnalysis of POLE mutation and tumor mutational burden (TMB) across 80,853 tumors: Implications for immune checkpoint inhibitors (ICPIs). *Annals of Oncology* **28** (2017).
55 Singal, G. *et al.* Association of patient characteristics and tumor genomics with clinical outcomes among patients with non–small cell lung cancer using a clinicogenomic database. *Jama* **321**, 1391-1399 (2019).

Acknowledgements

This research was supported by AstraZeneca (AZ) Postdoc Funding. The high-performance computing resource was supported by Scientific Computing Platform (SCP) at AZ. D.X.'s work was partially supported by the National Institutes of Health (R35-GM126985). We would like to thank our colleagues Kris Sachsenmeier, Zhongwu Lai, Melinda Merchant, Mingchao Xie and Steven Criscione for their helpful discussion and constructive suggestions.

Authors' contributions

J.D., J.S. and B.L. designed and directed this study. D.X. provided scientific suggestions and helpful discussion. C.F. collected and processed the data, built the model, trained the model, carried out experiments. All people analyzed and validated the experimental results of data analysis. All people wrote, reviewed and revised the manuscript.

Additional information

Supplementary Information accompanies this paper available online.

Conflict of interests

The authors declare no competing interests.

Figures

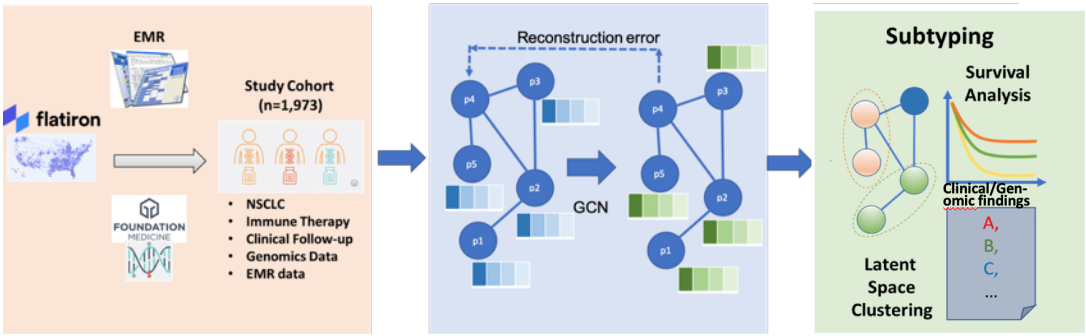
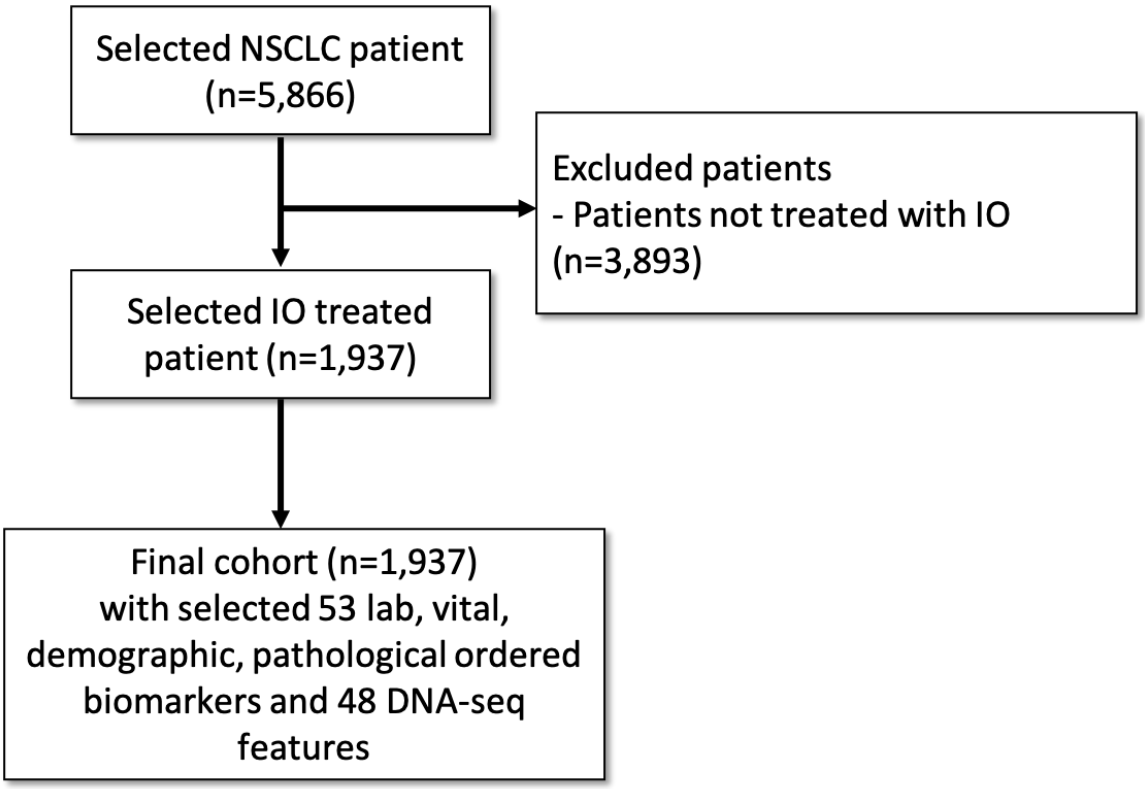


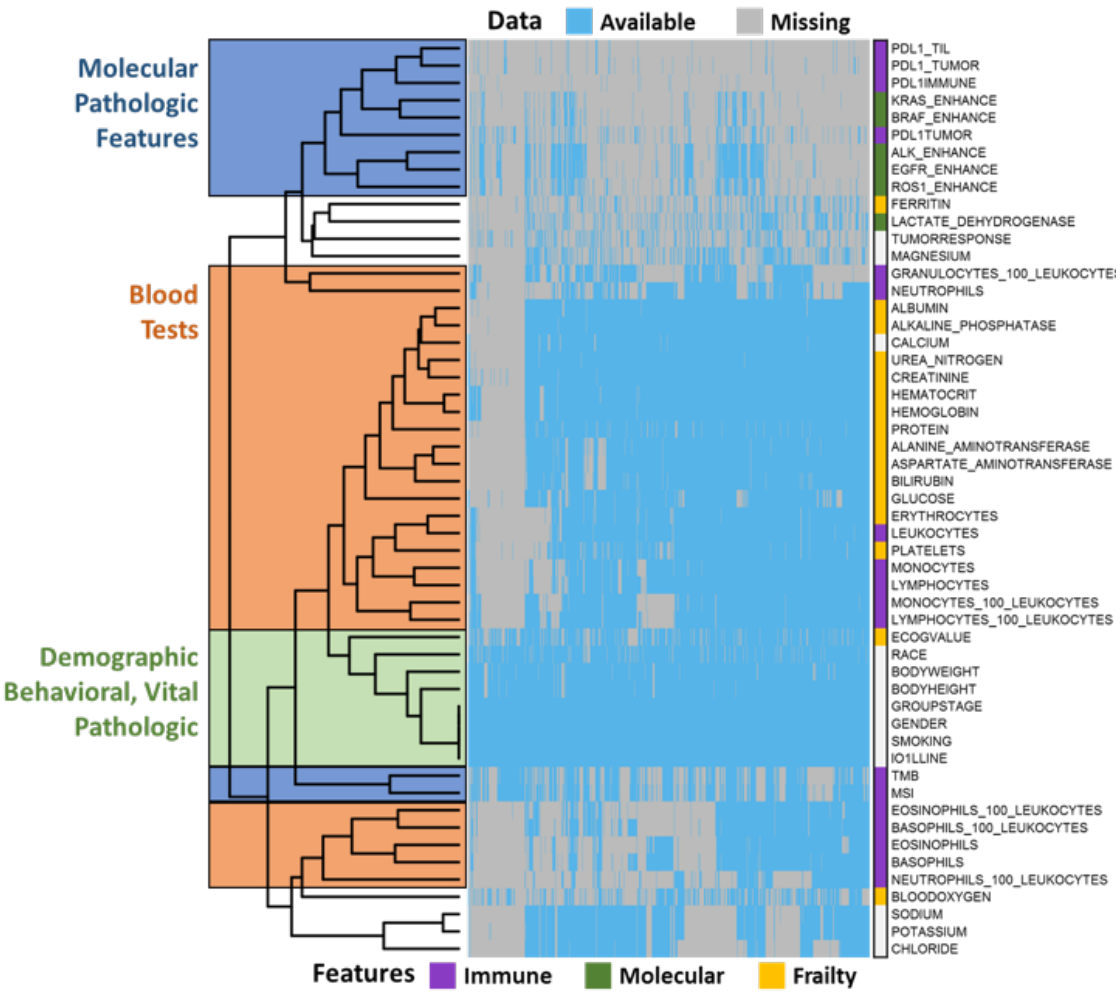
Figure 1. The overall workflow: The dataset is from real world evidences in Flatiron. Each patient has his clinical features and genomic features measured by Foundation Medicine. The features are preprocessed (See Material and Method part for details) and then concatenated. We applied marginalized graph autoencoder (mGAE) to learn each patient latent representation. Then, we applied spectral clustering algorithm to identify patient subtypes (clusters). Last but not least, we performed data analysis such as survival plot to compare between different subgroups and identified the clinical and genomic findings.


Selected NSCLC patient
(n=5,866)
Excluded patients
- Patients not treated with IO
(n=3,893)
Selected IO treated
patient (n=1,937)
Final cohort (n=1,937)
with selected 53 lab, vital,
demographic, pathological ordered
biomarkers and 48 DNA-seq
features

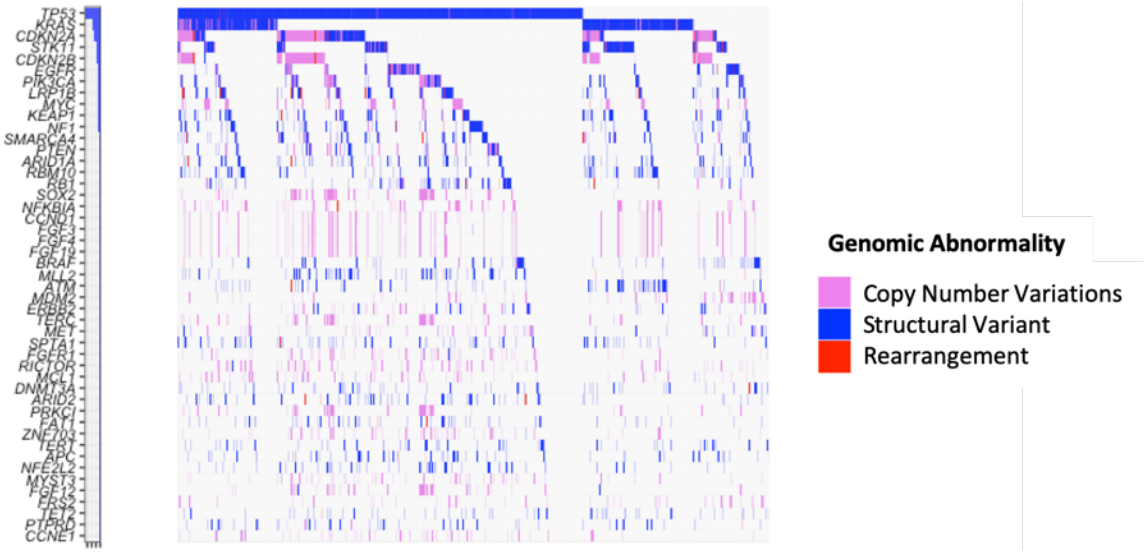
(a)

Baseline Demographic and Pathologic Characteristics			
Characteristics	All	Responsive Subgroup	Non-responsive Subgroups
number of patients	1,937	400	897
Age (year)			
Median, MAD	67.0, 10.4	67.0, 10.4	67.0, 10.4
Range	26.0-85.0	26.0-85.0	28.0-85.0
Sex: no., %			
Female	953 (49.2)	252 (63.0)	375 (41.8)
Race			
African American	144 (7.4)	31 (7.8)	80 (8.9)
White	1,428 (73.7)	289 (72.3)	648 (72.2)
Asian	46 (2.4)	10 (2.5)	19 (2.1)
Other Race	143 (7.4)	41 (10.3)	68 (7.6)
Histology			
Non-squamous cell carcinoma	1,433 (73.9)	329 (82.2)	601 (67.0)
Squamous cell carcinoma	419 (21.6)	62 (15.5)	259 (28.9)
NSCLC histology NOS	75 (3.8)	9 (2.3)	37 (4.1)
Stage: no., %			
Stage I	164 (8.5)	45 (11.3)	69 (7.7)
Stage II	122 (6.3)	30 (7.5)	55 (6.1)
Stage III	372 (19.2)	95 (23.8)	176 (19.6)
Stage IV	1,241 (64.1)	225 (56.3)	571 (63.7)
ECOG Score: no., %			
0	375 (19.4)	104 (26.0)	129 (14.4)
1	856 (44.2)	176 (44.0)	430 (47.9)
2	273 (14.1)	27 (6.8)	149 (16.6)
3	50 (2.6)	6 (1.5)	27 (3.0)
4	2 (0.1)	0 (0.0)	2 (0.2)
Smoking Status: no., %			
History of smoking	1,657 (85.5)	342 (85.5)	775 (86.4)
No history of smoking	276 (14.2)	57 (14.3)	120 (13.4)
Previous Treatment: no., %			
No	718 (37.1)	170 (42.5)	228 (25.4)
Yes	1,219 (62.9)	230 (57.5)	669 (74.6)
Eastern Cooperative Oncology Group (ECOG)			
MAD: Median Absolute Deviation.			

(b)

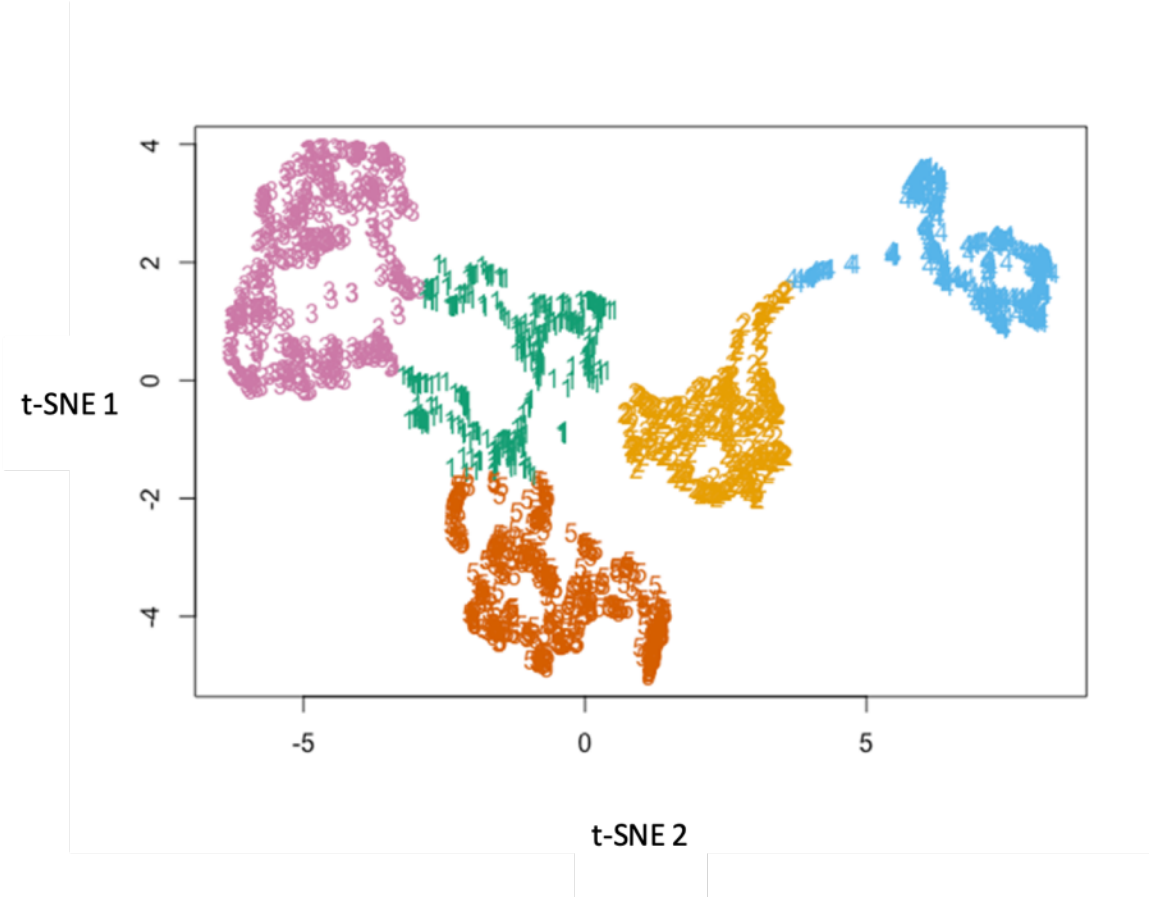

Data
Available
Missing
Molecular Pathologic Features
Blood Tests
Demographic Behavioral, Vital Pathologic
PDL1_TIL
PDL1_TUMOR
PDL1IMMUNE
KRAS_ENHANCE
BRAF_ENHANCE
PDL1TUMOR
ALK_ENHANCE
EGFR_ENHANCE
ROS1_ENHANCE
FERRITIN
LACTATE_DEHYDROGENASE
TUMORRESPONSE
MAGNESIUM
GRANULOCYTES_100_LEUKOCYTES
NEUTROPHILS
ALBUMIN
ALKALINE_PHOSPHATASE
CALCIUM
UREA_NITROGEN
CREATININE
HEMATOCRIT
HEMOGLOBIN
PROTEIN
ALANINE_AMINOTRANSFERASE
ASPARTATE_AMINOTRANSFERASE
BILIRUBIN
GLUCOSE
ERYTHROCYTES
LEUKOCYTES
PLATELETS
MONOCYTES
LYMPHOCYTES
MONOCYTES_100_LEUKOCYTES
LYMPHOCYTES_100_LEUKOCYTES
ECOGVALUE
RACE
BODYWEIGHT
BODYHEIGHT
GROUPSTAGE
GENDER
SMOKING
IO1LLINE
TMB
MSI
EOSINOPHILS_100_LEUKOCYTES
BASOPHILS_100_LEUKOCYTES
EOSINOPHILS
BASOPHILS
NEUTROPHILS_100_LEUKOCYTES
BLOODOXYGEN
SODIUM
POTASSIUM
CHLORIDE
Features
Immune
Molecular
Frailty

(c)

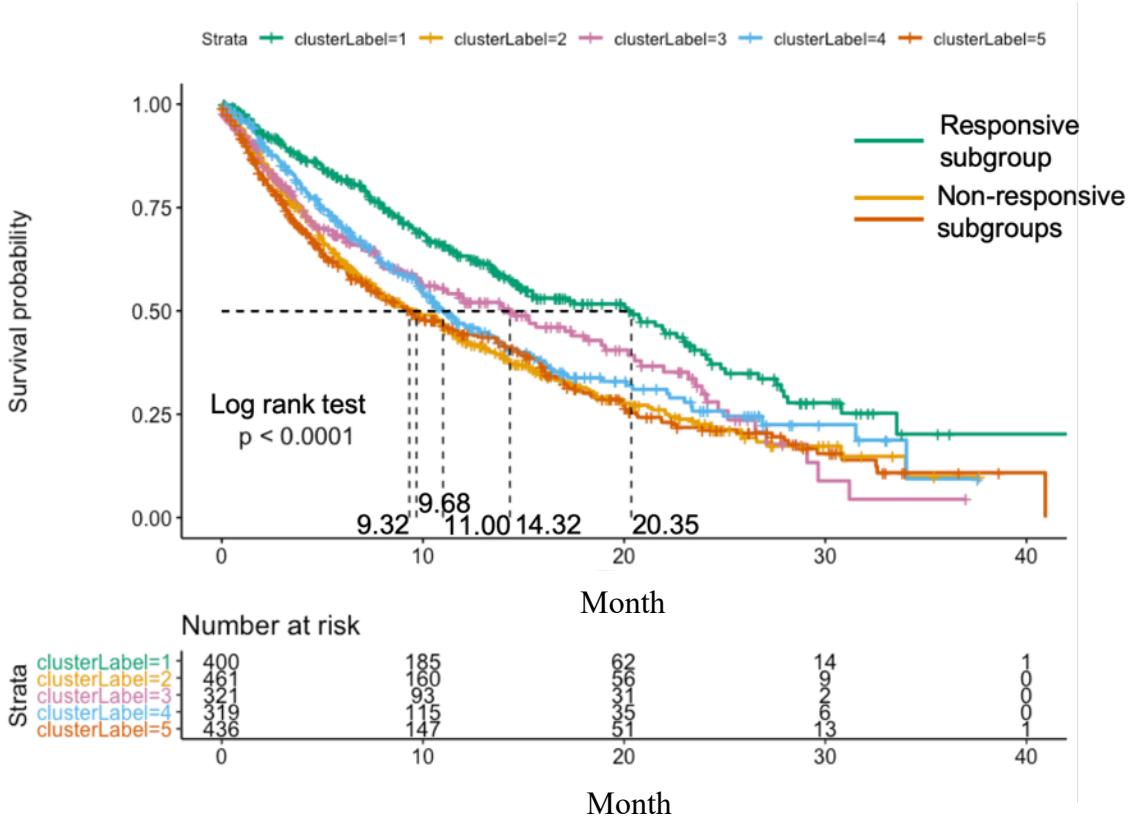


(d)

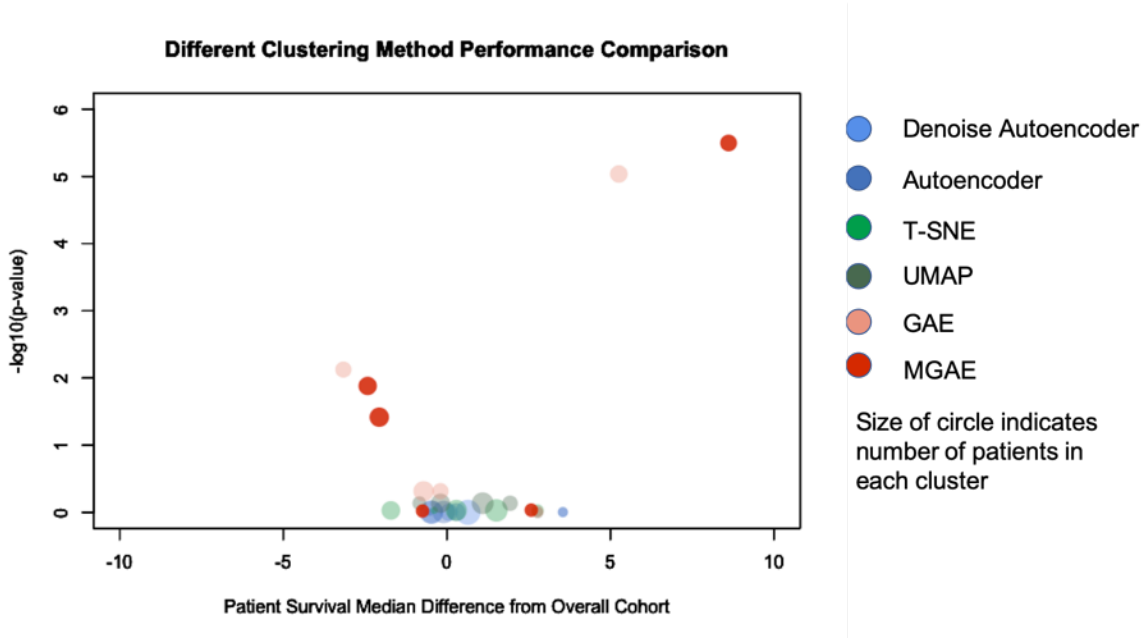
Figure 2. Identification of IO treated NSCLC cohort from Flatiron database and cohort characteristics a) Cohort Identification: define the cohort with inclusion and exclusion criteria. b) baseline demographic and pathologic characteristics c) an overview of all clinical features classified into molecular pathology features, blood test features. It also shows the incompleteness in RWE. d) genomic features are shown in waterfall plot.



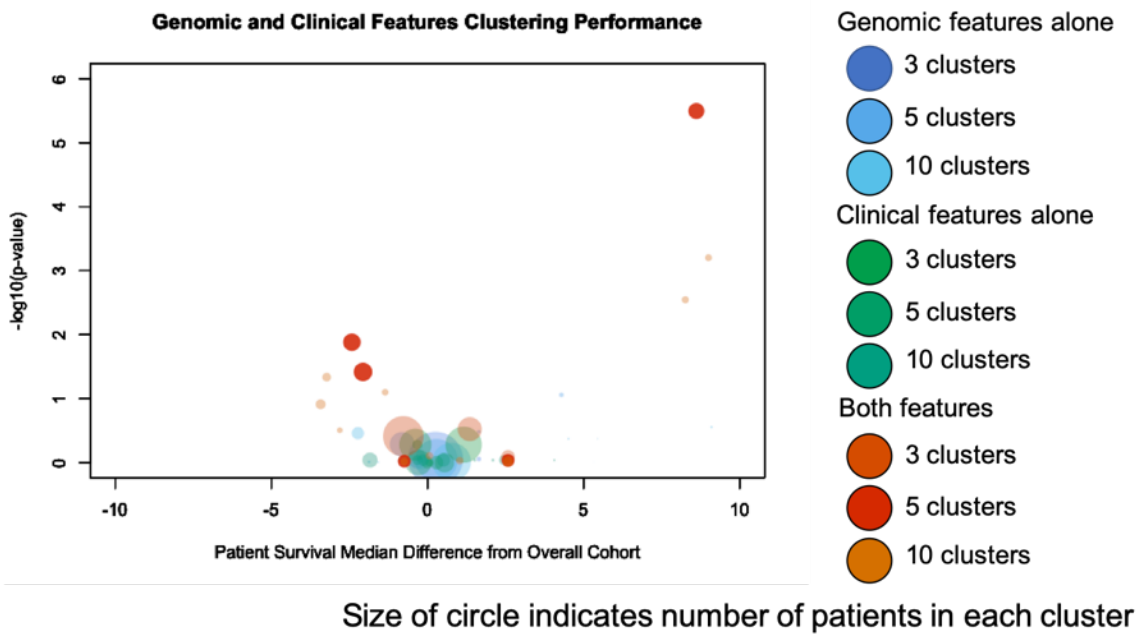
(a)


Strata
clusterLabel=1
clusterLabel=2
clusterLabel=3
clusterLabel=4
clusterLabel=5
Survival probability
1.00
0.75
0.50
0.25
0.00
Responsive subgroup
Non-responsive subgroups
Log rank test
p < 0.0001
9.32
9.68
11.00
14.32
20.35
0
10
20
30
40
Month
Number at risk
Strata
clusterLabel=1 400 185 62 14 1
clusterLabel=2 461 160 56 9 0
clusterLabel=3 321 93 31 2 0
clusterLabel=4 319 115 35 6 0
clusterLabel=5 436 147 51 13 1
0
10
20
30
40
Month

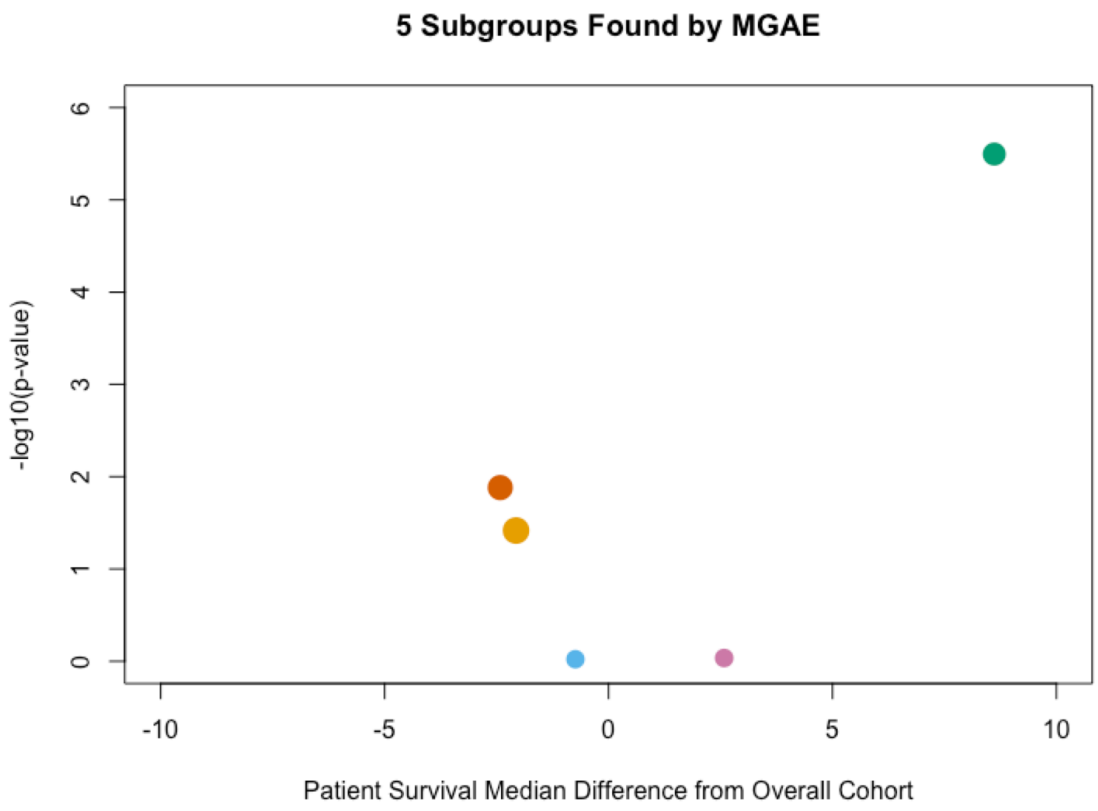
(b)


Different Clustering Method Performance Comparison
-log10(p-value)
0
1
2
3
4
5
6
-10
-5
0
5
10
Patient Survival Median Difference from Overall Cohort
Denoise Autoencoder
Autoencoder
T-SNE
UMAP
GAE
MGAE
Size of circle indicates number of patients in each cluster

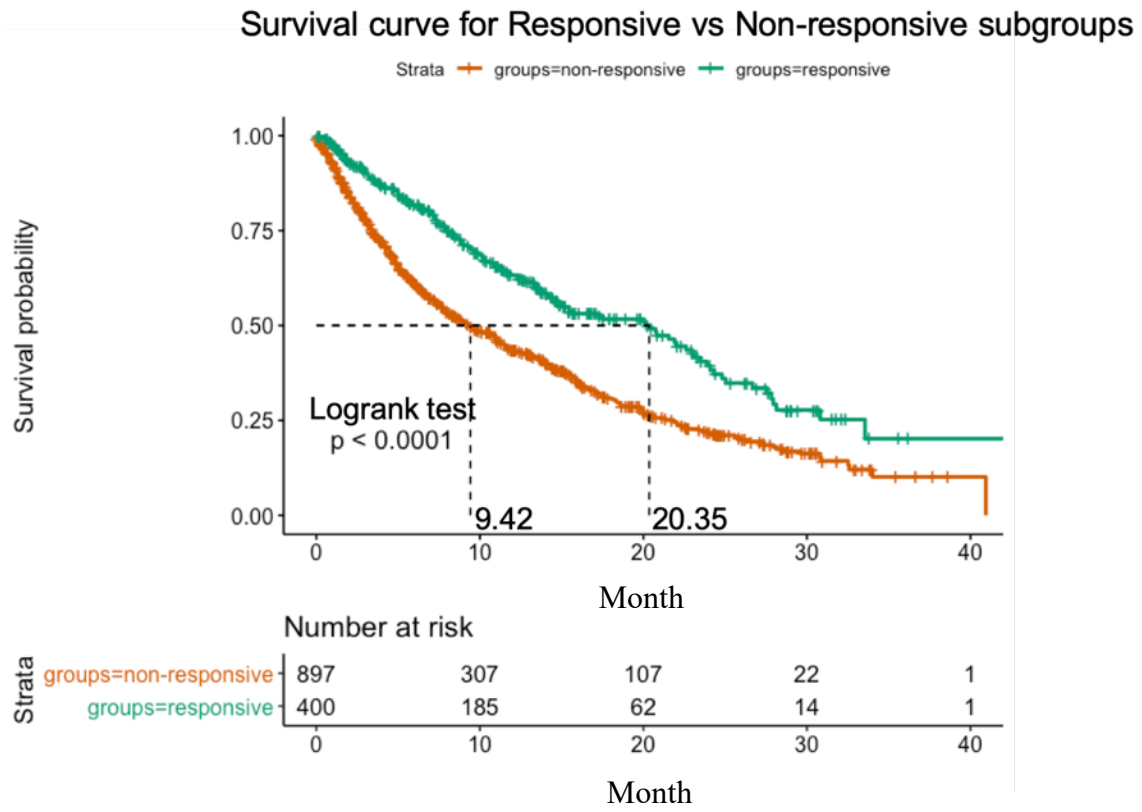
(c)


Genomic and Clinical Features Clustering Performance
-log10(p-value)
0
1
2
3
4
5
6
-10
-5
0
5
10
Patient Survival Median Difference from Overall Cohort
Genomic features alone
3 clusters
5 clusters
10 clusters
Clinical features alone
3 clusters
5 clusters
10 clusters
Both features
3 clusters
5 clusters
10 clusters
Size of circle indicates number of patients in each cluster

(d)


5 Subgroups Found by MGAE
-log10(p-value)
0
1
2
3
4
5
6
-10
-5
0
5
10
Patient Survival Median Difference from Overall Cohort

(e)



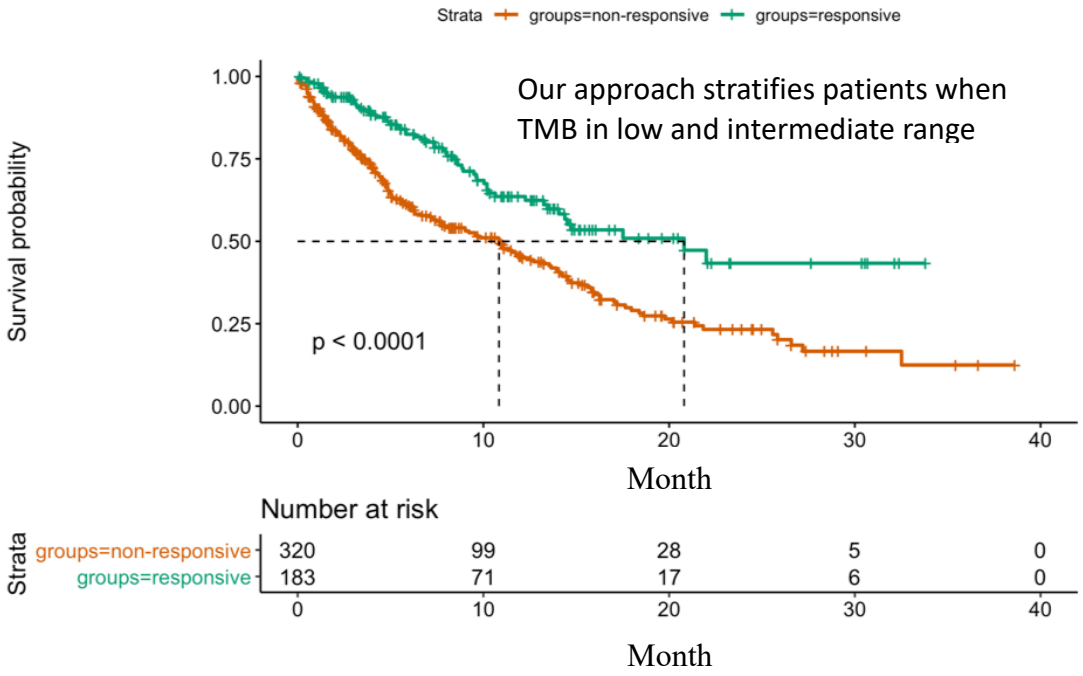
(f)

Figure 3. GCN discovered NSCLC subgroups with dramatically different responses to IO therapies from integrating both genomic data and EHR and GCN showed superior performances to commonly used machine learning methods. a) The GCN effectively learns embedding from patient clinical and genomic features. Here, a tSNE plot is to show the high level embedding in 2-d. The embeddings are formed five clusters which indicates five subgroups. 3b) The survival plots of five subgroups corresponding to the five clusters shown in Figure 3a. 3c) GCN approach compared with other machine learning approaches. Here, we compared our approach with other grid-based method and other well-established machine learning methods. In the volcano plot, each dot represents a cluster, the X axis will be the difference of the estimated median survival times between a cluster and the overall cohort, and the Y will be –log10(p-value), where the p-value will

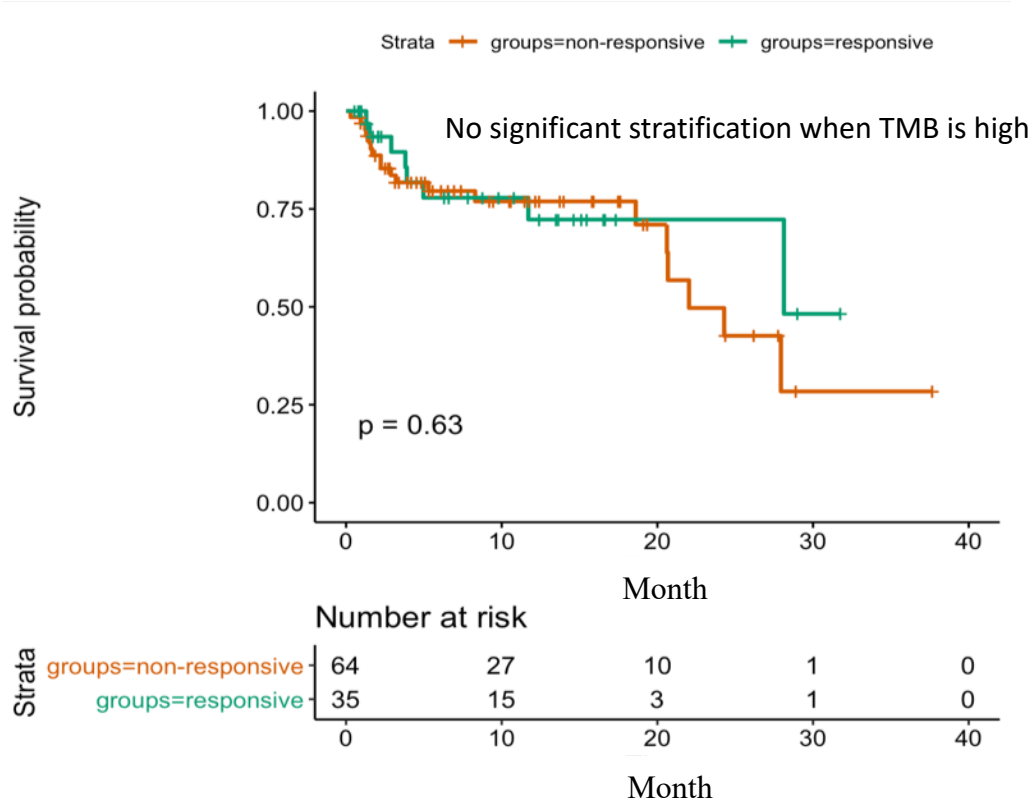
be from the log-rank test of the KM estimation of survival functions between a cluster and the overall cohort. It shows that our methods give better subgrouping results compared with other methods in terms of differentiable survival analysis and statistical significance. 4d) It shows that integrating both genomic and clinical features give the best subgroup results. Five clusters give best survival difference. We also explore other cluster parameters, but they fail to give statistically significant differentiable survival subgroups. 4e) responsive and non-responsive subgroups identified by GCN approach. Here for two non-responsive subgroups, we merge them together for following analysis. 4f) Survival plots for responsive and non-responsive subtypes after merging the two non-responsive subgroups; The survival analysis plot shows statistically significant difference between two subtypes as the responsive subgroup can have more then 10 months than nonresponsive one.

Feature	Responsive Enrichment	Nonresponsive Enrichment
Hemoglobin/ Hematocrit/ Erythrocytes		Abnormal (low)
Albumin		Abnormal (low)
Ferritin		Abnormal (high)
Lymphocytes		Abnormal (low)
Neutrophils		Abnormal (high)
Basophils	Abnormal (high, low)	
Protein		Abnormal (high,low)
Monocytes		Abnormal (high)
Lymphocytes %		Abnormal (low)
Platelets		Abnormal (high)
Granulocytes %		Abnormal (high)
Leukocytes		Abnormal (high)
Eosinophils		Abnormal (high, low)
Calcium		abnormal (high, low)
Glucose		Abnormal (high)
TMB	High	
Gender	Female	

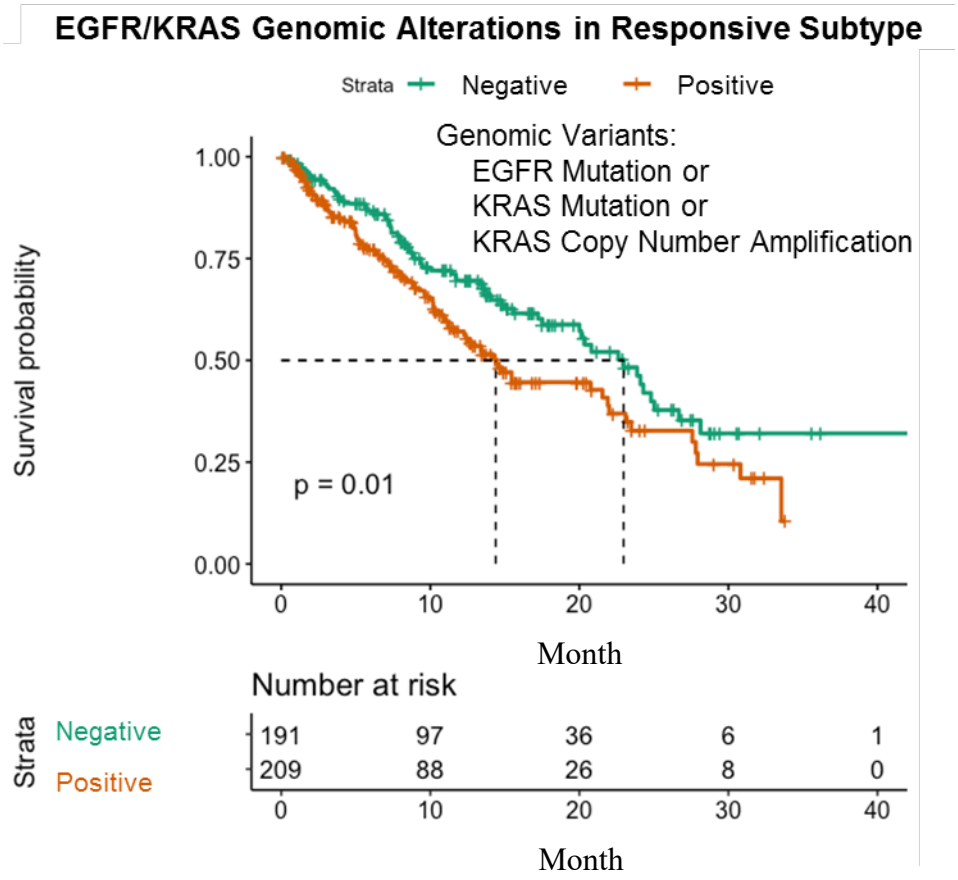
Figure 4: Enrichment of distinct clinical-phenotype differentiating the responsive and nonresponsive IO-treated NSCLC subgroups identified by GCN.


Strata
groups=non-responsive
groups=responsive
Our approach stratifies patients when TMB in low and intermediate range
Survival probability
1.00
0.75
0.50
0.25
0.00
p < 0.0001
0
10
20
30
40
Month
Number at risk
Strata
groups=non-responsive
groups=responsive
320
99
28
5
0
183
71
17
6
0
0
10
20
30
40
Month

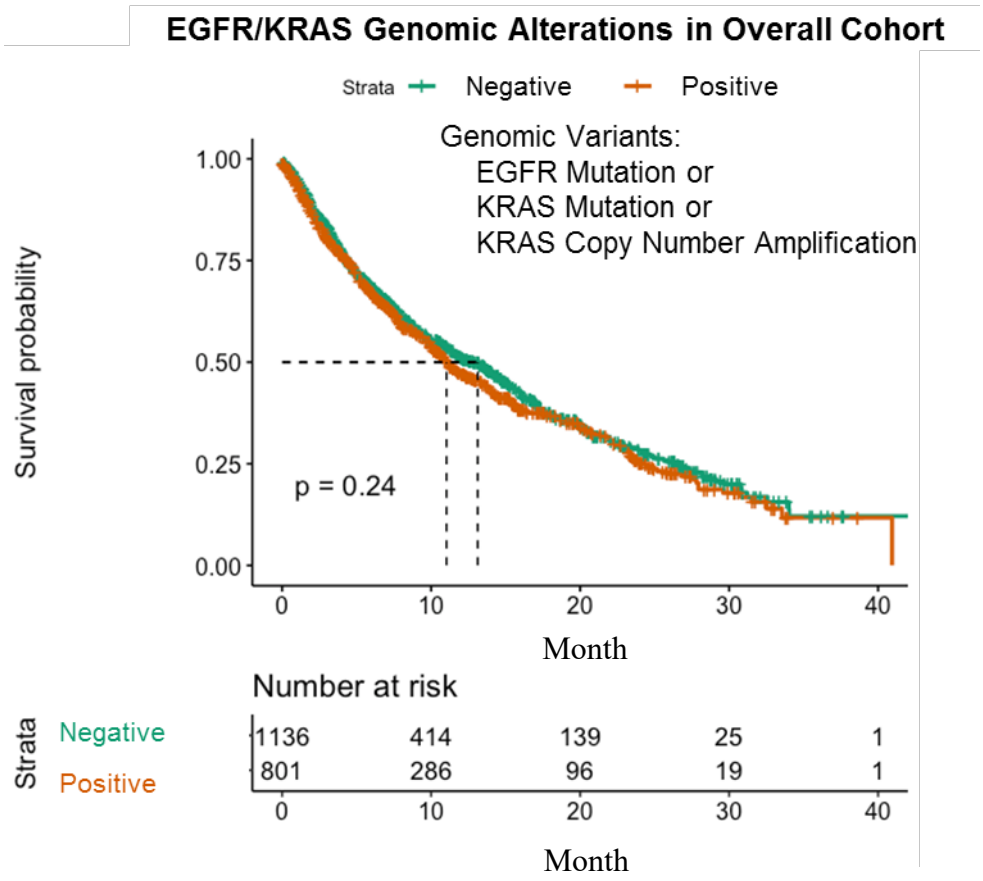
(a)


Strata
groups=non-responsive
groups=responsive
No significant stratification when TMB is high
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.63
0
10
20
30
40
Month
Number at risk
Strata
groups=non-responsive
64
27
10
1
0
groups=responsive
35
15
3
1
0
0
10
20
30
40
Month

(b)


EGFR/KRAS Genomic Alterations in Responsive Subtype
Strata
Negative
Positive
Genomic Variants:
EGFR Mutation or
KRAS Mutation or
KRAS Copy Number Amplification
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.01
0
10
20
30
40
Month
Number at risk
Strata
Negative
Positive
191
97
36
6
1
209
88
26
8
0
0
10
20
30
40
Month

(c)



(d)

Figure 5: GCN based subgrouping enabled IO based patient stratifications beyond utilizing TMB

Figure 5a) When TMB in low and intermediate range, GCN further stratifies patients into responsive and non-responsive subtypes. 5b) When TMB in high range, GCN cannot further stratify patients. From 5a and 5b it shows that although TMB is a commonly use biomarker to suggest patient IO treatment response, when TMB is in low and intermediate range, our approach can still stratify patient into differential survival subgroups with statistically significance. The effects of genomic alterations that may upregulate the EGFR/KRAS signaling on overall survivals are demonstrated in the Responsive subtype (5c) and the overall cohort (5d).