Clustering-based Pattern Discovery in Lung Cancer Treatments

Daniel Gómez-Bravo*, Aaron García*, Guillermo Vigueras*§, Belén Ríos-Sánchez*, Alejandra Pérez-García*, Vanessa Ospina†, María Torrente†, Ernestina Menasalvas*§, Mariano Provencio†, Alejandro Rodríguez-González*§
*Centro de Tecnología Biomédica
§Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid, Madrid, Spain
Email: {daniel.gomez-bravo,aaron.garcia.muniz,guillermo.vigueras,belen.rios,alejandra.perez.garcia,
ernestina.menasalvas, alejandro.rg}@upm.es
†*Hospital Universitario Puerta de Hierro Majadahonda*, Madrid, Spain
Email: avospina@hotmail.com,{mtorrente80,mprovenciop}@gmail.com

***Abstract*—Lung cancer is the leading cause of cancer death. More than 238,340 new cases of lung cancer patients are expected in 2023, with an estimation of more than 127,070 deaths. Choosing the correct treatment is an important element to enhance the probability of survival and to improve patient's quality of life. Cancer treatments might provoke secondary effects. These toxicities cause different health problems that impact the patient's quality of life. Hence, reducing treatments toxicities while maintaining or improving their effectivenes is an important goal that aims to be pursued from the clinical perspective.**

On the other hand, *clinical guidelines* include general knowledge about cancer treatment recommendations to assist clinicians. Although they provide treatment recommendations based on cancer disease aspects and individual patient features, a statistical analysis taking into account treatment outcomes is not provided here. Therefore, the comparison between clinical guidelines with treatment patterns found in clinical data, would allow to validate the patterns found, as well as discovering alternative treatment patterns.

In this work, we have analyzed a dataset containing lung cancer patients information including patients' data, prescribed treatments and their outcomes. Using a Chi-square test and K-Modes clustering algorithm in combination with Pattern Discovery metrics we identify patterns, within the clusters, based on cancer stage and treatment outcomes. Obtained results are analyzed based on statistical and clinical relevance and compared with lung cancer clinical guidelines. The comparison reveals that all patterns found coincide with clinical guidelines recommendations, assessing the validity of the proposed method for pattern discovery in a clinical dataset.

***Index Terms*—Cancer treatment, Machine learning, Clustering, Pattern Discovery**

I. Introduction

Among the different types of cancer, lung cancer (LC) has a high mortality being the leading cause of cancer death. From the estimation of cases that will be in 2023, more than 238,340 new cases are expected to be of LC patients, with an estimation of more than 127,070 deaths [1]. Although survival of patients has significantly increased ("NSCLC 2-year relative survival increased from 34% for persons diagnosed during 2009 through 2010 to 42% during 2015 through 2016") [2]), greater efforts to improve screening methods and treatments are still needed in order to improve patient's survival and quality of life.

Choosing the correct treatment is an important element to enhance the probability of survival and to improve patient's quality of life. Cancer treatments might provoke secondary effects. Variability found in patient's toxicities depend on several factors such as cancer stage and combination of treatments [3]. These toxic outcomes can endanger the homeostatic equilibrium of the body, causing different health problems that impact patient's quality of life. Hence, it is important to reduce treatments toxicities while maintaining or improving their effectiveness. Thus, clinical guidelines aim to collect knowledge and best practices from oncologists proposing recommendations in order to improve clinical practices for LC treatment. Nevertheless, although LC clinicians are expected to adopt these clinical recommendations [4], [5], the inter and intrapatients' variability of response to the different combinations of treatments, makes it is necessary to personalize different treatment-patterns on certain cases.

The identification of patterns on both the patients that have been treated following the standard of care, or outside it, would allow to identify potential treatment recommendations based on the patient characteristics. As part of the so-called "Precision Medicine", some previous works have used data-driven approaches to analyze how similarities and di-similarities in patient's profiles can affect cancer treatment outcomes [6]–[8]. Within these approaches, descriptive machine learning models can assist in result interpretation since these models describe the represented domain in a meaningful manner, both for humans and computers. Based on descriptive models, other works have used a classifier to describe cluster groups based on specific variable values [9]. On the other hand, more recent works have tried to describe clusters using Treemaps [10] or by computing the mode of clusters variables [11]. Another

This paper has been supported by Fundación AECC and Instituto de Salud Carlos III (grant AC19/00034), under the frame of ERA-NET PerMed.

work proposes the combination of subgroup discovery and hierarchical clustering to obtain groups of frequent patterns and select the most relevant patterns for describing each cluster [12]. However, all these approaches lack validation of obtained cluster descriptions. For that reason, in this paper we have used K-Modes [13], a descriptive and unsupervised machine learning method for finding patient and treatment patterns. In this way, patterns from clusters with similar properties are identified and validated against clinical treatment recommendations. Nevertheless, clustering methods require the selection of data features and identification of the adequate number of clusters while providing meaningful interpretation of data contained withing the clusters.

In order to overcome these limitations, we propose a way of selecting data features through a statistical hypothesis test to analyze the optimal number of clusters through a variance ratio criterion, fully implemented in K-Modes Python Library [13]. Then, resulting clusters are filtered to obtain statistically and clinically relevant groups, by identifying patterns associated to each cluster and computing pattern-related metrics for each cluster. Thus, patterns identified are used to provide interpretability to the data contained on each cluster. Additionally, discovered patterns have been compared with clinical guidelines as a reference for evaluating results and for identifying new potential treatments. The clinical information about the patients was provided by Hospital Universitario Puerta de Hierro - Majadahonda and it is the result of previous work extracting clinical information from electronic health records [14], [15].

II. Materials and Methods

A. Data adquisition and preprocessing

In this study we exploit the structured information about 1,242 patients included in a dataset provided by the Medical Oncology Department at Puerta de Hierro University Hospital (HUPHM). It contains a detailed description of the diagnosis and treatment phases, as well as personal and medical data recorded during anamneses. The information comes from the Electronic Health Records (EHRs) and is structured as follows:

- Patient information: demographic information, medical history and diagnosis information. Only cancer initial stage variable was considered in this study.
- Treatments: chemotherapy, radiotherapy, surgery, immunotherapy and drugs. A deeper description is presented later on this paper (section II-B).
- Outcomes: cancer relapses and toxicities associated to the treatments were taken into account for this study. They are presented in a binary form as can be seen in the section II-B.

During the initial exploratory analysis we detected some irrelevant information, missing values and data inconsistencies that were fixed or removed. This study is focused on NSCLC, so data related to other cancer or histologic types was removed. Also, patients who did not receive any treatment and patients whose toxicity was unknown were discarded as they do not add any relevant medical information. In addition, as temporality is a key factor in our analysis, we cleaned data focusing on treatment and diagnostic dates. Particularly, we checked that a patient's treatments have the corresponding dates associated, or at least the starting date, as well as the coherency of dates (i.e. starting date must be previous to end date, and diagnostic date must be previous to treatment date). Therefore, after carrying out a cleaning process eliminating irrelevant or erroneous information, the number of patients was reduced from 1,242 to 652.

B. Data description

This study aims to characterize lung cancer patients, focussing on the initial stage of their tumors and the first treatment they received, to find treatment patterns of clinical interest based on treatment outcomes. Accordingly, a subset of variables was selected from the original dataset described below.

Cancer initial stage was selected from patient information. This variable follows the TNM staging system of the American Joint Committee on Cancer [16] where different cancer stages are identified depending on tumor size and organs affected by metastasis. TNM estadification system was updated from version 7 [17] to version 8 [16] in January 2017, but the dataset refers to the stages established in both versions.

Table I shows distribution of patients across different cancer initial stages and treatments, in the form of a cross table. It can be clearly seen that our dataset presents patients in all stages, but it is not balanced.

Regarding treatments presented in Table I, we observe five different types: chemotherapy (CT), chemotherapy and radiotherapy concurrencies (CT - RT), surgery (SUR), targeted oral therapy (also named as *Drugs*) which is intended to treat patients with driver mutations, and immunotherapy (*Imm*), which prevents PD-1 from binding to PD-L1, making the T-lymphocyte able to detect and reduce tumor cells. In the analysis, the type, date, purpose, and application information about the treatments were considered.

According to their intention, we can find curative surgeries (SUR_{Curat}), which are oriented to remove all malignant tissue, and palliative surgeries (SUR_{Pallia}), targeted to reduce symptoms. Similarly, we can distinguish between palliative chemotherapy (CT_{Pallia}), which is aimed to improve symptoms and prolong survival and neoadjuvant chemotherapy (CT_{NeoAdj}), applied before any local treatment to reduce the tumor size. Moreover, radiotherapy can be differentiated into palliative radiotherapy (RT_{Pallia}), which aimed to improve symptoms and prolong survival, and radical radiotherapy (RT_{Rad}), employed to cure the disease and/or maintain the function of the organ.

Chemotherapy and radiotherapy concurrencies (CT-RT) are classified as: Neoadjuvant chemotherapy - radiotherapy ($CT-RT_{Neoadj}$) or Radical chemotherapy - radiotherapy ($CT-RT_{Radical}$). $CT-RT_{Neoadj}$ refers to the combination of treatments aimed to reduce the tumor size before surgery or a radical treatment while $CT-RT_{Radical}$ denotes the combination of treatments aimed to reduce the tumor size and

TABLE I: Cross table for first treatment received per cancer stage

	IA	IB	IIA	IIB	IIIA	IIIB	IIIC	IV	IVA	IVB	Total
SUR_Curat	13,65%	10,74%	5,21%	6,13%	8,74%	0,77%	0,00%	1,38%	0,31%	0,31%	47,24%
CT_Pallia	0,00%	0,00%	0,00%	0,46%	0,77%	3,22%	1,07%	5,83%	5,37%	6,29%	23,01%
CT_NeoAdj	0,00%	0,15%	0,31%	0,31%	8,13%	2,15%	0,00%	0,31%	0,00%	0,00%	11,35%
Drugs	0,00%	0,15%	0,00%	0,00%	0,31%	0,15%	0,00%	2,30%	0,77%	2,76%	6,44%
CT-RT_Rad	0,00%	0,00%	0,00%	0,00%	1,53%	1,99%	0,15%	0,15%	0,00%	0,00%	3,83%
Imm	0,00%	0,00%	0,00%	0,15%	0,00%	0,00%	0,00%	1,07%	0,77%	1,84%	3,83%
RT_Rad	0,15%	0,31%	0,00%	0,31%	0,31%	0,00%	0,00%	0,15%	0,15%	0,00%	1,38%
RT_Pallia	0,00%	0,00%	0,00%	0,00%	0,31%	0,15%	0,15%	0,00%	0,46%	0,15%	1,23%
CT-RT_NeoAdj	0,00%	0,00%	0,00%	0,00%	0,46%	0,00%	0,00%	0,00%	0,15%	0,00%	0,61%
SUR_Pallia	0,15%	0,15%	0,00%	0,00%	0,15%	0,00%	0,00%	0,15%	0,00%	0,00%	0,61%
CT-RT_Adj	0,00%	0,00%	0,00%	0,00%	0,00%	0,31%	0,00%	0,00%	0,00%	0,00%	0,31%
CT-Refract_NeoAdj	0,00%	0,00%	0,00%	0,00%	0,00%	0,15%	0,00%	0,00%	0,00%	0,00%	0,15%

curate it. Furthermore, we can find Drugs, which are aimed to attack a specific mutation, and immunotherapy (IMM), which aims to target PD-L1 mutation.

TABLE II: Treatments outcome description

Outcome	Percentage
Yes Prog-Rel, No Tox	40,49%
No Prog-Rel, No Tox	38,03%
Yes Prog-Rel, Yes Tox	14,88%
No Prog-Rel, Yes Tox	6,59%

In order to describe the outcomes, two binary variables (have/not have) related to the first treatment were selected: toxicities and progression/relapse. As can be seen in Table II a large proportion ($\sim$79%) of the population do not have toxicities associated to initial treatments considered, where almost half ($\sim$47%) of these treatments consist of curative surgery with reduced secondary effects (see Table I). On the other hand, progression-relapses are quite balanced.

C. Clinical guidelines

Clinical guidelines include recommendations intended to optimize patient care and therefore guidelines are essential to assist clinicians in lung cancer treatment. These guidelines define a set of treatment paths, where recommendations depend on cancer disease aspects and individual features for a concrete patient. Table III describes the aspects present in the guidelines we have considered [18]–[25].

TABLE III: Aspects checked in clinical guidelines

Aspect	Description	Answers
Surgery (SUR)	Can surgery be performed?	Yes No
Resectable (RES)	Is the tumor resectable?	Yes No Potentially
Mutations (MUT)	Patient with driver mutations?	Yes No

We have defined a rule representation of clinical guidelines paths to ease the comparison with the patterns obtained in this paper. Thus, we propose a representation using the following general notation:

$$\underbrace{STAGE\ (ASPECT)}_{\text{antecedent}} \rightarrow \underbrace{[TREATMENT\ LIST]}_{\text{consequent}} \quad (1)$$

The notation in (1) states that the list of treatments in the consequent is advised when the antecedent is met, in terms of cancer stage and cancer aspects (see Table III). Thus for example, the following rule:

$$I\ (SUR?\ :\ YES) \rightarrow [CT,\ SUR] \quad (2)$$

Advises that, when tumor is in stage 1 (2) and is possible to perform surgery (SUR? : YES), then apply two treatments, one after another: first apply chemotherapy (CT), then perform surgery (SUR).

On the other hand, clinical guidelines also specify treatments within the different options, that must be applied jointly but in two different ways: concurrently or sequentially. Examples of these treatment applications can be seen in the following stage III rules:

$$III\ (RES?\ :\ NO) \rightarrow [conc(CT + RT),\ DRUGS] \quad (3)$$

$$III\ (RES?\ :\ NO) \rightarrow [seq(CT + RT)] \quad (4)$$

The first rule (3) advises that, if tumor is in stage III and tumor is not resectable (RES? : NO), apply two treatments, one after another: the first one a concurrent application (conc(...)) of chemotherapy (CT) and radiotherapy (RT), then apply targeted oral therapy (DRUGS). The second rule (4) advises that, if tumor is in stage III and tumor is not resectable, then use one treatment consisting on a sequential application (seq(...)) of chemotherapy (CT) and radiotherapy (RT). Based on this notation, the full set of clinical guidelines paths we have obtained, is listed on the right most column in Table V, grouped by cancer stage.

D. Machine Learning methods

Different types of cluster analysis methods are found in literature [26]. We can distinguish partitioning methods (denoted as flat) which optimize assignment of objects into a certain number of clusters, and methods for hierarchical cluster analysis with graphical outputs which make assignment of objects into different numbers of clusters. In the first group, k-centroids and k-medoids methods are used for disjunctive clustering. The former is based on initial assignment of the objects into k clusters. For this purpose, k initial centroids are selected which are the geometric centers of the k clusters. After that, the distances of each object from all centers are calculated to assign each cluster object to the closest centroid. Later, centroids are updated based on assignment of objects

to clusters. In a next step, the distances from each object to all centroids are re-calculated again, and if an object is found to be closer to the centroid of a different cluster, the object is re-assigned to that cluster. This process is repeated a number of given iterations or while objects can be re-assigned.

On the other hand, hierarchical clustering methods can be classified as agglomerative (step-by-step clustering of objects and groups to larger groups) or divisive (step-by-step splitting of the whole set of objects into the smaller subsets and individual objects). Furthermore, we can distinguish monothetic (only one variable is considered in individual steps) and polythetic (all variables are considered in individual steps) clustering.

Among the different methods described, we have used partitioning-based clustering adapted to categorical data due to the nature of the variables present in the clinical dataset used (see Section II-B). Concretely, the selected method is K-Modes, where distance between categories is measured by means of special coefficients based on the chi-square dissimilarity measure [26].

We have combined K-Modes with the use of WRAcc (Weighted Relative Accuracy), a quality measure from Pattern Discovery methods [27]. According to this metric, statistically significant clusters that fulfill a given target property are found. Thus, we defined the clustering target based on two binary variables of the input dataset (i.e. disease progression-relapse and toxicity). In this way, our target contains the information of both binary variables, i.e. $Progression - Relapse$=[YES/NO] & $Toxicity$=[YES/NO], obtaining four different targets as a result of the combination of variables values. A sub-population fulfilling each target is defined and clustering using K-Modes is perform for each one. For the clusters obtained, WRAcc is described as the balance between the group coverage and its accuracy gain, and can be computed for each subgroup as follows:

$$WRAcc(G,T) = p(G) \times (p(T \ \& \ G) - p(T)) \qquad (5)$$

where term *p(G)* represents the subgroup coverage computed as the probability of belonging to a particular subgroup *G*. On the other hand, term *p(T&G) - p(T)* represents the accuracy gain, where *p(T)* is the probability of fulfilling the cluster input target *T* and *p(T&G)* is the probability of jointly being part of a particular subgroup *G* and fulfilling the input target *T*.

Finally, we have relied on p-value statistical measure as a clustering filtering criterion which is calculated using the Chi Square statistical test [28]. Thus, the combination of p-value along with the *WRAcc* metric, described before, enables the K-Modes method to discover statistically and clinically relevant clusters in the medical field [29].

Clustering analysis has been implemented using K-modes Python library [13]. This library allows to handle the majority of K-Modes parameters found in literature. Furthermore, WRAcc and p-value were calculated following general formulation described in literature.

In order to find most promising patterns using K-Modes, we have define the following process. Initially, we perform feature selection using chi-square test of independence in order to verify that our variables are related with each sub-population target, and select only those having p-value below 0.05. Then, we make combinations of the selected variables in order to evaluate K-Modes to find the optimal number of clusters. This is assessed using Calinski-Harabasz (CH) score, since it has been proposed as a well suited metric to identify the best cluster configuration as the one with the highest CH score [30]. Each cluster centroid can be considered as a pattern that fulfills the target of the sub-population analyzed. In order to compare cluster patterns with clinical guidelines, cluster centroids are extended with the mode, within the cluster, of Initial Cancer stage or First Treatment variables. In this way, obtained patterns are ensured to contain the main driver variables used for treatment recommendations in clinical guidelines. Then, obtained patterns are filtered using a WRAcc threshold greater than 0.0 in order to avoid patterns with small coverage or insufficient accuracy gain. The described analysis is performed for each of the four targets mentioned before, obtaining four sets of patterns. Finally, the four pattern sets are statistically filtered using a p-value measure lower or equals than 0.05. In this way and as mentioned before, we combine the WRAcc pattern metric and the p-value statistical measure to filter clusters, and obtain the most reliable patterns of each set [25].

III. Results

This section describes the results obtained after applying K-Modes to the clinical dataset considered (Section II-B). Table IV shows clusters and CH measures obtained for all combinations of target variables, as described in Section II-D. According to Table IV, First Treatment, Toxicity and Prog/Rel variables are chosen for Prog=NO_Tox=NO and Prog=YES_Tox=NO targets, whereas Initial Stage, Toxicity and Prog/Rel variables are selected for Prog=NO_Tox=YES and Prog=YES_Tox=YES targets. Thus, the selected combinations of variables are used to perform clustering for each specific target.

Table V shows discovered patterns stating the relation between LC stage and initial prescribed treatments along with disease progression-relapse and treatment toxicity. Thus, Table V gathers the patterns found (second and third columns), progression-relapse and toxicity information (fourth and fifth columns), WRAcc and p-value measures (sixth and seventh columns) and clinical guidelines recommendations (eighth column) according to notation described in Section II-C. Additionally, subgroups and clinical guidelines paths are grouped by cancer stage.

Looking at patterns found for stage I, the association presented SUR_{Curat} as first treatment and the target no progression/relapse - no toxicities. In addition, the association has reported a p-value significantly below $5.0e-02$, which can be statistically considered as highly significant subgroup. Moreover, the high WRAcc value obtained indicates a strong dependence between the subgroup pattern and the associated target. According to clinical guidelines, if the surgery is possible at stages I or II, the application of CT_{NeoAdj} before surgery will be optional depending on tumor characteristics.

TABLE IV: Variables combination and Calinski-Harabasz score

		Target							
		Prog=No_Tox=No		Prog=No_Tox=Yes		Prog=Yes_Tox=No		Prog=Yes_Tox=Yes	
Idx	**Variables**	Nº Clust	CH Score	Nº Clust	CH Score	Nº Clust	CH Score	Nº Clust	CH Score
0	[First Treatment, Toxicity, Prog/Rel]	5	**591.42**	6	1.0	8	**818.83**	6	282.37
1	[Cancer Stage, Toxicity, Prog/Rel]	7	553.18	6	**119.91**	8	569.51	5	**357.8**
2	[First Treatment, Cancer Stage, Toxicity, Prog/Rel]	3	97.69	6	18.18	5	59.19	3	24.35

TABLE V: Cluster patterns and Clinical Guidelines recommendations

Stage I							
Patterns			**Target**		**Measures**		
Index	**Cancer stage**	**First Treatment**	**Progression/Relapse**	**Toxicity**	**WRAcc**	**p-value**	**Clinical guidelines**
1	IA	SUR$_{Curat}$	NO	NO	4.9 %	$1.62e-03$	I (SUR? : YES) → [CT, SUR] I (SUR? : NO) → [RT, CT]
Stage III							
Patterns			**Target**		**Measures**		
Index	**Cancer stage**	**First Treatment**	**Progression/Relapse**	**Toxicity**	**WRAcc**	**p-value**	**Clinical guidelines**
2	IIIA	SUR$_{Curat}$	YES	NO	1.86 %	$6.68e-04$	III (RES? : YES) → [CT] IIIA (RES? : POT) → [CT, SUR] III (RES? : POT) → [conc(CT+RT), SUR] III (RES? : NO) → [conc(CT+RT), DRUGS] III (RES? : NO) → [seq(CT+RT)]
3	IIIA	CT$_{NeoAdj}$	NO	YES	1.0 %	$4.86e-04$	
4	IIIB	CT - RT$_{Radical}$	YES	YES	0.63 %	$3.63e-03$	
Stage IV							
Patterns			**Target**		**Measures**		
Index	**Cancer stage**	**First Treatment**	**Progression/Relapse**	**Toxicity**	**WRAcc**	**p-value**	**Clinical guidelines**
5	IV	CT$_{Pallia}$	YES	NO	1.64 %	$3.03e-04$	IV (MUT? : YES) → [DRUGS] IV (MUT? : YES) → [conc(DRUGS+CT)] IV (MUT? : YES) → [seq(DRUGS+CT)] IV (MUT? : NO) → [conc(DRUGS+CT)] IV (MUT? : NO) → [conc(CT+IMM)] IV (MUT? : NO) → [IMM]
6	IV	CT$_{Pallia}$	YES	YES	0.85 %	$1.6e-02$	
7	IVA	CT$_{Pallia}$	YES	YES	0.76 %	$2.6e-02$	
8	IVB	Drugs	YES	YES	0.83 %	$7.9e-04$	

Thus, the subgroup obtained for stage I agrees with clinical guidelines recommendations.

For stage III, three patterns were obtained using K-modes. Pattern 2 and Pattern 3 have reported SUR$_{Curat}$ and CT$_{NeoAdj}$ as first treatment respectively in IIIA cancer stage. However, Pattern 2 developed progression-relapse on the cancer status without toxicities as a side effect, since SUR$_{Curat}$ can not produce any toxicity; meanwhile, Pattern 3 did not develop progression-relapse on the cancer status but develop toxicities as a side effect. On the other hand, Pattern 4 has reported CT - RT$_{Radical}$ as first treatment in IIIB cancer stage, as well as developing progression-relapse on the cancer status and toxicities as a side effect. All these patterns have reported a p-value below $1.0e-02$, which can be considered statistically relevant, being Pattern 2 and Pattern 3 the most significant ones since they provided the lowest p-values. Moreover, Pattern 2 has the highest WRAcc value indicating a stronger Pattern-target dependence than the others. All Pattern found for stage III match with clinical guidelines since Pattern treatments are also suggested as treatment recommendations.

For stage IV, four Patterns are obtained from Cluster analysis. Pattern 5, Pattern 6, and Pattern 7 have reported CT$_{Palliat}$ as the first treatment and developed progression-relapse on the cancer status. Toxicity information is not shared among them, but as Pattern 5 has a higher WRAcc value, this outcome is more reliable as it shows higher subgroup-target dependence than the others. Pattern 8 has reported Drugs as first treatment, which according to clinical guidelines, is prescribed when patients have driver mutations. This subgroup has developed a progression-relapse in the cancer status and toxicities as a side effect. In addition, all these patterns have reported a p-value below $1.0e-02$, which can be considered statistically relevant, being Pattern 5 and Pattern 8 the most significant ones since they provided the lowest p-values.

IV. Conclusions

In this work we have proposed a method based on K-Modes clustering to find clinically relevant patterns considering patient profiles, treatments prescribed and their outcomes. Clustering has proven to be useful for identifying, within our cohort of patients, treatment patterns of high clinical interest, taking treatment results as a reference. Even with an unbalanced toxicity distribution, most of the patterns show toxicity as an outcome. In addition, discovered patterns have been compared with clinical guidelines as a reference for evaluating results and for identifying new potential treatments.

The comparison reveals that for stage I, surgery is the first prescribed treatment in agreement with clinical recommendations. This group did not developed progression-relapses on the cancer status nor toxicities as a side effect. Several studies have demonstrated that curative-intent surgery, when coupled with regional lymph node examination, is generally associated with the best long-term overall survival in patients with early-stage NSCLC [31].

Stage III NSCLC is a heterogeneous and complex disease that could be classified into subgroups: resectable, potentially resectable and unresectable locally advanced NSCLC [4]. In patients with potentially resectable disease, the optimal treatment strategy remains unclear. Several phase III trials and a meta-analysis showed that induction therapy followed by surgery might be better than surgery alone [32]. Alos, obtained results reveal that chemotherapy and surgery are prescribed as first treatments for stage III. Although the patterns are in line with guidelines recommendations, different progression-relapse and toxicity outcomes were found for the recommended treatments.

Regarding stage IV, agreement with clinical guidelines is observed in most of the cases since chemotherapy and targeted oral therapy are prescribed as initial treatments. Also, different progression-relapse and toxicity targets were found in these subgroups, but in most of the cases progression-relapse and toxicities were obtained.

REFERENCES

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023.

[2] R. Siegel, K. Miller, H. Fuchs, and A. Jemal, "Cancer Statistics, 2021," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 1, pp. 7–33, 2021.

[3] M. Or, B. Liu, J. Lam, W.-K. Chan, and F. C. Wong, "A systematic review and meta-analysis of treatment-related toxicities of curative and palliative radiation therapy in non-small cell lung cancer," *Scientific Reports*, vol. 11, no. 1, p. 5939, 2021.

[4] M. Majem, O. Juan, A. Insa, N. Reguart, J. M. Trigo, E. Carcereny, R. García-Campelo, Y. García, M. Guirado, and M. Provencio, "SEOM clinical guidelines for the treatment of non-small cell lung cancer (2018)," *Clinical & Translational Oncology*, vol. 21, no. 1, pp. 3–17, 2019.

[5] F. R. Hirsch, G. V. Scagliotti, J. L. Mulshine, R. Kwon, W. J. Curran, Y.-L. Wu, and L. Paz-Ares, "Lung cancer: current therapies and new targeted treatments," *Lancet (London, England)*, vol. 389, no. 10066, pp. 299–311, Jan. 2017.

[6] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics," *AMIA Summits on Translational Science Proceedings*, vol. 2014, pp. 132–136, Apr. 2014.

[7] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang, "Deep Patient Similarity Learning for Personalized Healthcare," *IEEE transactions on nanobioscience*, vol. 17, no. 3, pp. 219–227, Jul. 2018.

[8] S.-A. Brown, "Patient Similarity: Emerging Concepts in Systems and Precision Medicine," *Frontiers in Physiology*, vol. 7, p. 561, Nov. 2016.

[9] S. Petrovic, "A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters," in *Proceedings of the 11th Nordic workshop of secure IT systems*, vol. 2006. Citeseer, 2006, pp. 53–64.

[10] C. Díez-Sanmartín, A. S. Cabezuelo, and A. A. Belmonte, "A new approach to predicting mortality in dialysis patients using sociodemographic features based on artificial intelligence," *Artificial Intelligence in Medicine*, vol. 136, p. 102478, Feb. 2023.

[11] D. Ferreira-Santos and P. Pereira Rodrigues, "Phenotyping obstructive sleep apnea patients: a first approach to cluster visualization," *Stud Health Technol Inform*, vol. 255, no. 255, pp. 75–79, 2018.

[12] U. Niemann, M. Spiliopoulou, B. Preim, T. Ittermann, and H. Völzke, "Combining subgroup discovery and clustering to identify diverse subpopulations in cohort study data," in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 2017, pp. 582–587.

[13] N. J. de Vos, "kmodes categorical clustering library," https://github.com/nicodv/kmodes, 2015–2021.

[14] M. Najafabadipour, M. Zanin, A. Rodríguez-González, M. Torrente, B. Nuñez García, J. L. Cruz Bermudez, M. Provencio, and E. Menasalvas, "Reconstructing the patient's natural history from electronic health records," *Artificial Intelligence in Medicine*, vol. 105, p. 101860, May 2020.

[15] E. Menasalvas Ruiz, J. M. Tuñas, G. Bermejo, C. Gonzalo Martín, A. Rodríguez-González, M. Zanin, C. González de Pedro, M. Méndez, O. Zaretskaia, J. Rey, C. Parejo, J. L. Cruz Bermudez, and M. Provencio, "Profiling Lung Cancer Patients Using Electronic Health Records," *Journal of Medical Systems*, vol. 42, no. 7, p. 126, May 2018. [Online]. Available: https://doi.org/10.1007/s10916-018-0975-9

[16] M. Amin, S. Edge, F. Greene, D. Byrd, R. Brookland, M. Washington, J. Gershenwald, C. Compton, K. Hess, and et al. (Eds.), *AJCC Cancer Staging Manual (8th edition)*. American Joint Committee on Cancer, 2017.

[17] S. Mirsadraee, D. Oswal, Y. Alizadeh, A. Caulo, and E. V. van Beek, "The 7th lung cancer tnm classification and staging system: Review of the changes and implications." *World journal of radiology*, vol. 4, no. 4, pp. 128–34, 2012.

[18] P. Villalobos and I. I. Wistuba, "Lung cancer biomarkers," *Hematology/Oncology Clinics of North America*, vol. 31, no. 1, pp. 13–29, feb 2017.

[19] M. Majem, O. Juan, A. Insa, N. Reguart, J. M. Trigo, E. Carcereny, R. García-Campelo, Y. García, M. Guirado, and M. Provencio, "SEOM clinical guidelines for the treatment of non-small cell lung cancer (2018)," *Clinical and Translational Oncology*, vol. 21, no. 1, pp. 3–17, nov 2018. [Online]. Available: https://doi.org/10.1007\%2Fs12094-018-1978-1

[20] N. Horita, T. Woo, N. Miyazawa, and T. Kaneko, "Pre-operative chemotherapy for non-small cell lung carcinoma," *Translational Lung Cancer Research*, vol. 4, no. 1, 2014.

[21] F. R. Hirsch, G. V. Scagliotti, J. L. Mulshine, R. Kwon, W. J. Curran, Y.-L. Wu, and L. Paz-Ares, "Lung cancer: current therapies and new targeted treatments," *The Lancet*, vol. 389, no. 10066, pp. 299–311, jan 2017.

[22] B. Ferrell, M. Koczywas, F. Grannis, and A. Harrington, "Palliative care in lung cancer," *Surgical Clinics of North America*, vol. 91, no. 2, pp. 403–417, Apr. 2011.

[23] D. R. Gomez, G. R. Blumenschein, J. J. Lee, M. Hernandez, R. Ye, D. R. Camidge, R. C. Doebele, F. Skoulidis, L. E. Gaspar, D. L. Gibbons, J. A. Karam, B. D. Kavanagh, C. Tang, R. Komaki, A. V. Louie, D. A. Palma, A. S. Tsao, B. Sepesi, W. N. William, J. Zhang, Q. Shi, X. S. Wang, S. G. Swisher, and J. V. Heymach, "Local consolidative therapy versus maintenance therapy or observation for patients with oligometastatic non-small-cell lung cancer without progression after first-line systemic therapy: a multicentre, randomised, controlled, phase 2 study," *The Lancet Oncology*, vol. 17, no. 12, pp. 1672–1682, Dec. 2016.

[24] O. Juan and S. Popat, "Ablative therapy for oligometastatic non–small cell lung cancer," *Clinical Lung Cancer*, vol. 18, no. 6, pp. 595–606, Nov. 2017.

[25] P. Iyengar, Z. Wardak, D. E. Gerber, V. Tumati, C. Ahn, R. S. Hughes, J. E. Dowell, N. Cheedella, L. Nedzi, K. D. Westover, S. Pulipparacharuvil, H. Choy, and R. D. Timmerman, "Consolidative radiotherapy for limited metastatic non–small-cell lung cancer," *JAMA Oncology*, vol. 4, no. 1, p. e173501, Jan. 2018.

[26] H. Řezanková, V. Praze, and Praha, "Cluster analysis and categorical data," vol. 89.

[27] F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, "An overview on subgroup discovery: foundations and applications," *Knowledge and Information Systems*, vol. 29, no. 3, pp. 495–525, Dec. 2011.

[28] M. L. McHugh, "The Chi-square test of independence," *Biochemia Medica*, vol. 23, no. 2, pp. 143–149, Jun. 2013.

[29] C. Esnault, M.-L. Gadonna, M. Queyrel, A. Templier, and J.-D. Zucker, "Q-Finder: An Algorithm for Credible Subgroup Discovery in Clinical Data Analysis — An Application to the International Diabetes Management Practice Study," *Frontiers in Artificial Intelligence*, vol. 3, 2020.

[30] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, Jan. 1974.

[31] A. Chi, W. Fang, Y. Sun, and S. Wen, "Comparison of Long-term Survival of Patients With Early-Stage Non-Small Cell Lung Cancer After Surgery vs Stereotactic Body Radiotherapy," *JAMA network open*, vol. 2, no. 11, p. e1915724, Nov. 2019.

[32] S. S. Burdett, L. A. Stewart, and L. Rydzewska, "Chemotherapy and surgery versus surgery alone in non-small cell lung cancer," *The Cochrane Database of Systematic Reviews*, no. 3, p. CD006157, Jul. 2007.