DEVELOPMENT OF A SOFTWARE SUPPORT SYSTEM FOR TARGETED TREATMENT OF NON-SMALL CELL LUNG CANCER UTILIZING GENETIC MUTATION ANALYSIS DATA

Le Tu Linh[1,2], Nguyen Viet Nhung[1,2,3], Trinh Le Huy[1,4]
Le Van Quang[1,5], Dinh Van Luong[1,2] and Nguyen Thi Trang[1,6,7,]✉

[1]Hanoi Medical University
[2]Department of Oncology,National Lung Hospital
[3]Department of Pulmonology, University of Medicine and Pharmacy, Vietnam National University,
[4]Department of Oncology and Palliative Care, Hanoi Medical University Hospital
[5]Department of Head and Neck Surgery, Vietnam National Cancer Hospital
[6]Department of Biology and Medical Genetic, Hanoi Medical University
[7]Genetic Counseling Center, Hanoi Medical University Hospital

This study presents the development of a software support system aimed at improving targeted treatment outcomes in non-small cell lung cancer (NSCLC) by utilizing genetic mutation analysis data. PubMed BERT model were generated to identify tumor gene mutated features associated with gene-drug responses. Then the best classifier with highest accuracy served for the development of the support software. A multi-center retrospective cohort study was conducted to evaluate of treatment outcomes based on software support to identify and predict targeted therapeutic in NSCLC. The results demonstrated the potential of leveraging genetic mutation analysis data and AI technology to optimize treatment strategies and improve outcomes for NSCLC patients with PubMed BERT model in extracting and categorizing keywords to build the database (a Recall (sensitivity) score of 98.12%). The study revealed that among the 109 patients with EGFR mutations identified through NGS analysis, 72% showed partial responsiveness, overall response rate (ORR) was 67.0%, and the disease control rate (DCR) was 82.6%. The software support system developed in this study holds promise for enhancing personalized treatment approaches in NSCLC by leveraging genetic mutation analysis data to guide targeted therapies and improve the treatment outcome for patients.

Keywords: Supported software; Targeted therapy, EGFR, Lung cancer.

I. INTRODUCTION

Epidermal growth factor receptor (EGFR) mutations are detected in approximately 40% to 50% of East Asian patients diagnosed with lung adenocarcinoma.[1,2] The administration of EGFR tyrosine kinase inhibitors (EGFR-TKIs) such as erlotinib, gefitinib, or afatinib has demonstrated significant enhancements in survival rates when compared to standard chemotherapy among patients with EGFR mutations in advanced NSCLC.[3] Notably, patients harboring mutations in exon 19 deletion or exon 21 exhibit a median progression-free survival (PFS) of approximately 9 to 13 months, accompanied by an objective response rate ranging from 60% to 70%.[4,5] Despite the promising outcomes associated with EGFR-TKI therapy, challenges persist, particularly in the form of primary resistance, which is more prevalent in the East

Corresponding author: Nguyen Thi Trang
Hanoi Medical University
Email: trangnguyen@hmu.edu.vn
Received: 23/05/2024
Accepted: 08/07/2024

Asian population. Recent reports underscore the importance of considering concurrent genetic changes that may contribute significantly to resistance mechanisms, thereby elucidating the marked variability observed in individual patient responses. Coexistent genetic alterations, such as HER2 amplification, MET amplification, PIK3CA mutation, and KRAS mutation, have been identified as potential contributors to primary resistance for EGFR-TKIs treatment.[6]

To assist physicians, particularly in Vietnam, in choosing suitable targeted therapy drugs for patients, several databases such as OncoKB and COSMIC have been developed.[7,8] However, these databases are updated manually, potentially causing delays in reflecting the latest advancements in research. The utilization of deep learning methodologies presents a promising approach to expedite diagnostic processes and produce highly accurate outcomes, particularly in prognosis. These methodologies have gained recognition for their remarkable ability to enhance the efficiency, precision, and consistency of predictions, especially in the realm of oncology. Equipped with an accurate database, artificial intelligence (AI) can accurately forecast a patient's response to specific targeted therapies, greatly assisting clinicians in treatment decision-making[9]. Nonetheless, the challenge of acquiring large and precise data from hospital records presents a significant obstacle in the training of deep learning models. This challenge primarily arises from the stringent and limited availability of medical data (due to patient privacy protection and prevention of data misuse). The complexity of the situation is compounded by the variability in treatment choices among healthcare providers.[10]

Consequently, we develop a machine learning system aiming to create a dataset utilized for AI training derived from established databases and research on lung cancer mutation in Pubmed, Human Genetic Mutation Database (HGMD), OncoKB, and COSMIC. Finally, the evaluation of treatment outcomes based on software support is to identify and predict targeted therapeutic in NSCLC.

II. MATERIALS AND METHODS

1. For the first research objective

We aim to develop a database to draw relations between cancer gene selecting proper medications. We are building an AI model to collect PubMed, HGMD, COSMIC, ClinicalTrials.gov, and OncoKB data. Our AI can automatically search and create a database from these articles. Our study's method consists of 3 steps:

- Building a classification model to identify cancer gene mutation related articles.

- Develop keywords extraction and data mapping model to create a database on targeted therapies' response on different mutations.

- Building an automated software supporting the diagnosis and treatment of lung cancer.

Building a classification model to identify cancer gene mutation related articles.

Data preprocessing involved concatenating titles and abstracts, removing special characters, stop words, lemmatization, and tokenization. The **TF-IDF (Term Frequency - Inverse Document Frequency) method** is used to represent the feature vectors input into the classification model.

To develop the binary classification module to identify cancer genetic mutation- related articles, we divided our training dataset into type 1 (positive) and type 2 (negative) class. The positive class are articles related to genetic mutation, while negative class are articles not related to genetic mutation.

For the positive class, we collected data from the HGMD database, which provides related scientific articles directly associated with gene mutations, including full texts and citations from PubMed. By using this database, positively labeled articles were obtained from HGMD, and their PubMed Identifiers (PMIDs) were used to gather information from PubMed.

For the negative class, we randomly collected articles from PubMed to ensure equal data points in both classes.

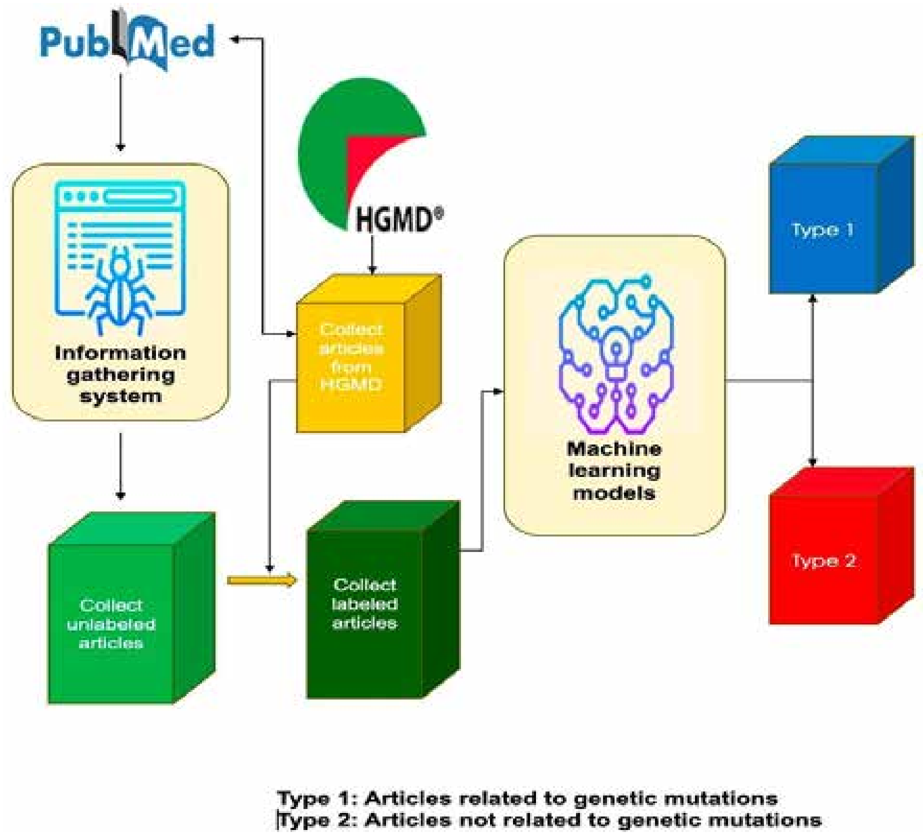
PubMed
HGMD®
Information gathering system
Collect articles from HGMD
Machine learning models
Type 1
Type 2
Collect unlabeled articles
Collect labeled articles
Type 1: Articles related to genetic mutations
Type 2: Articles not related to genetic mutations

Fig. 1. Flowchart of Binary Text Classification Model Operation

Model performance was evaluated using **K-fold cross-validation** on a dataset of 225,149 articles, with 124,850 in the positive class (including 4,079 articles from HGMD and others cited from them), and 100,299 articles belong to the negative class. The articles are downloaded and their titles and summaries are recorded. Data preprocessing involved concatenating titles and abstracts, removing special characters, stop words, lemmatization, and tokenization.

A traditional machine learning classification module is built based on 6 algorithms:

- K-Nearest Neighbours
- Bernoulli Naive Bayes
- Multinomial Naive Bayes
- Complement Naive Bayes
- Logistic Regression
- Linear Support Vector Machine.

By evaluating these algorithms, we aim to find a text classification model that is accurate and reliable.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

The primary evaluation metric for related articles and main indexes will be called Recall index (Sensitivity)The same system is also being used for further classification of the articles into lung-cancer-related articles and targeted-therapy-for-lung-cancer-related articles.

Develop keywords extraction and data mapping model to create a database on targeted therapies' response on different mutation.

Experiments with **BioBERT and PubMedBERT models** were conducted to extract relationships between entities in medical texts. Various methods were compared to select the optimal model. A Mapping Module Architecture was proposed for Variants Related to Genes/Target Drugs in Lung Cancer, comprising six main components:

- Named Entity Recognition Module (NER)

- Cancer Relation Extraction Module (RE)

- Ethnicity and Nationality Extraction Module (ENE)

- Data Collection Module (DCM)

- Database Mapping and Normalization Module (DMN)

- Database Management Module (DBM).

Building a automated software supporting the diagnosis and treatment of lung cancer

- Building AI model to detect mutations from NGS raw data: Research, design, and development of software for detecting mutations from NGS raw data.

- Building variant-treatment mapping model: Buiding variant-treatment mapping model to identify gene mutations in patient samples and support personalized targeted therapy.

2. The second research objective

Evaluation of treatment outcomes based on software support to identify and predict targeted therapeutic in NSCLC. A multi-center retrospective cohort study was conducted in two large health centers including National Lung Hospital and National Cancer Hospital from January 2019 to June 2023. The patients who were diagnosed with NSCLC carrying EGFR 19del mutation or EGFR 21L858R mutation were enrolled. The inclusion criteria were as follows: (I) sensitive EGFR mutations detected by next-generation sequencing (NGS) from tumor tissue or liquid biopsy; (II) NSCLC patients are receiving initial treatment with gefitinib, erlotinib, or afatinib, supported by software; (III) their response to treatment evaluated after at least 2 months of supervision. We excluded the patients under three months of age and incomplete medical records. Collection and analysis of single EGFR mutation and concurrent genetic alterations. Detection of more than 2800 hotspot mutations in 50 cancer-related genes (Appendix) was performed using a semiconductor-based next-generation sequencing (NGS) platform.

Outcomes

Progression-Free Survival (PFS) of EGFR-TKIs was defined as the duration from the initiation of EGFR-TKI treatment to disease progression or death from any cause. All patients were regularly monitored during the treatment process to assess clinical response and diagnostic imaging every 8 weeks (or earlier for significant progression appeared). The best clinical response to treatment was evaluated based on the RECIST guidelines (version 1.1) by a fully trained clinician or radiologist of the participating. Objective response rate (ORR) included complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). Disease Control Rate (DCR) was determined by the sum of objective response and stable disease (CR + PR + SD). The predictive outcomes were determined based on progression-free survival (PFS) and overall survival (OS), classifying them into two distinct categories: favorable and unfavorable outcomes.

Data analysis

Survival curves were calculated using the Kaplan-Meier method was employed to from the beginning of the advanced NSCLC diagnosis to mortality or the last follow-up documented. p values were calculated using Fisher's exact test and Pearson's test for categorical and continuous variables, respectively. Continuous variables and binary variables were compared using the Wilcoxon test. All statistical analyses were performed using SPSS 20.0 software (IBM Corporation, NY, USA). A p-value <0.05 was considered as significant.

3. Research Ethics

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of the Hanoi Medical University (No. 912/ GCN-HĐĐĐNCYSH-ĐHYHN). The study was approved to collect data from the National Lung Hospital, National Cancer Hospital.

III. RESULTS

1. Compare the accuracy of algorithms in constructing text classification models.

The collected dataset includes 225,149 articles of which 124,850 articles belong to the positive class (with 4079 articles from HGMD as the rest are similarly cited articles from PubMed) and 100,299 articles belong to the negative class. It is important to note that all the articles within each class are distinct and there are no duplication. The articles are downloaded and their titles and summaries are recorded.

The results of the classification model using different algorithms are presented below:

Table 1. Summarize the recall comparison of the models

Model name	Recall
K-Nearest Neighbors	96.47%
Bernoulli Naive Bayes	84.99%
Multinomial Naive Bayes	98.12%
Complement Naive Bayes	97.74%
Logistic Regression	95.46%
Support Vector Machine (SVM)	96.79%

Table 1 illustrates the Recall index across different models. Among them, Multinomial Naive Bayes exhibits the highest Recall at 98.12%, followed sequentially by Complement Naive Bayes, Support Vector Machine (SVM), K- Nearest Neighbors, and Logistic Regression, with Recall index of 97.74%, 96.79%, and 96.47%, respectively. The Bernoulli Naive Bayes algorithm attains the lowest Recall at 84.99%.

2. The PubMedBERT and BioBERT models constructed in the study with similar models of the NER module

Table 2. Compare the results of the NER module on the BioBERT dataset

Model	F1-score
BioBERT - CRF	88.7
PubMedBERT -CRF	89.3
BiLSTM - CRF	87.1
PubMedBERT(Ours)	88.8
BioBERT (Ours)	87.6
BERT-GT	56.5
Pubmed BERT	58.9

We compared the constructed models of PubMedBERT and BioBERT with similar variants in the NER module. The results indicated that the F1-scores of Our PubMedBERT and Our BioBERT are 88.8% and 87.6%, respectively, surpassing BiLSTM – CRF (87.1%), BERT-GT (56.5%), and Pubmed BERT (58.9%). However, the F1-scores of our models were lower than PubMedBERT – CRF (88.7%) and BioBERT – CRF (89.3%). Some results deviated slightly from the article due to some discrepancies such as different data preprocessing, different hyperparameters, etc.

3. Characteristics of the constructed database:

By applying our BioBERT model, we have successfully constructed a database, in which a total of 59 genes were synthesized with 286 gene variations, and 101 therapeutic drugs had their drug licensing status determined. Of these therapeutic drugs, 32 are licensed by both the Ministry of Health of Vietnam and the FDA, 28 are licensed by the FDA, while the remaining 37 are still in the process of clinical trials of which 2 drugs have been approved by the National Medical Products Administration and National Assembly of China (NMPA) licensed.

Table 3. Gene mutations, drug therapies, and treatment responses in the database

Gene	Variant	Targeted therapy	Relationship	Resource
EGFR	T790M	Gefitinib	Resistant	PubMed:18676761
		Afatinib	Resistant	PubMed:26862733
		Erlotinib	Resistant	PubMed:21430269
	C797G	Orsimetinib	Resistant	PubMed:29807405
NTRK1	SQSTM1	Entrectinib	Response	PubMed:26565381
	SQSTM1	Entrectinib	Response	PubMed:28183697
	G595R & G667S	Larotrectinib	Resistant	PubMed:30624546

KRAS	G12A	Erlotinib,Gefitinib	Resistant	PubMed:19794967
	G12C	Sotorasib	Response	PubMed:34096690
	A146V	Abemaciclib	Response	PubMed:24836576
	G12V	Gefitinib	Resistant	PubMed:17409929
	Non-specific	Erlotinib + Tivantinib	Response	PubMed:NCT00777309
BCL2L11	Non-specific	Gefitinib + Vorinostat	Response	PubMed:NCT02151721

Table 3: Illustrates some examples of mutated genes, targeted therapy, and the response potential of the targeted therapy when carrying that gene mutation. For example, the EGFR T790M mutation is resistant to Erlotinib, while the KRAS non-specific mutation strongly responds to the Erlotinib + Tivantinib therapy.

4. Target therapy treatment outcomes for non-small cell lung cancer based on prediction software

Among 109 patients with advanced NSCLC, with targeted NGS test, exon 19 deletion (Ex19del) mutation was detected in 75 patients, exon 21 L858R point mutation was in 32 and both Exon19del mutation and exon 21 L858R point mutation were in 2 (1.8%).

Table 4. Comparison of clinical profile between single *EGFR* mutation and concurrent gene alteration patients

Characteristics	Single *EGFR* mutation (N=66) n (%)	Concurrent alteration (N=43) n (%)	*p*-values
Gender			0.02*
Male	23 (34.8)	25 (58.1)	
Female	43 (65.2)	18 (41.9)	
Age group			0.70
<60	29 (43.9)	17 (39.5)	
≥60	37 (56.1)	26 (60.5)	
Smoking status			0.03*
Never	19 (28.8)	17 (39.5)	
Former/current	37 (56.1)	26 (60.5)	
Histology			0.40
Adenocarcinoma	66 (100)	42 (97.7)	
No-adenocarcinoma	0 (0)	1 (2.3)	

Characteristics	Single *EGFR* mutation (N=66) n (%)	Concurrent alteration (N=43) n (%)	*p*-values
Stage at *EGFR*-TKI treatment			0.74
IIIB	5 (7.6)	4 (9.3)	
IV	61 (92.4)	39 (90.7)	
EGFR mutation type			0.03*
Exon 19 deletion	51 (77.3)	24 (55.8)	
L858R	15 (22.7)	17 (39.5)	
Exon 19 deletion+Exon 21 L858R	0 (0.00)	2 (4.7)	
Performance score at *EGFR*-TKI treatment			0.56
0-1	65 (98.5)	41 (95.3)	
2-3	1 (1.5)	2 (4.7)	

*significant at 0.05

All 109 patients were analyzed for KRAS, PIK3CA, BRAF, MET mutations and ALK, ROS1 fusion genes. We compared the clinical features of 66 patients harboring single EGFR mutation with those of the 43 patients harboring co-alterations with an EGFR mutation. The analysis of data show that a significant differences in EGFR del19 group compared with EGFR exon 21 group and those who never smoke compared with those who smoke .

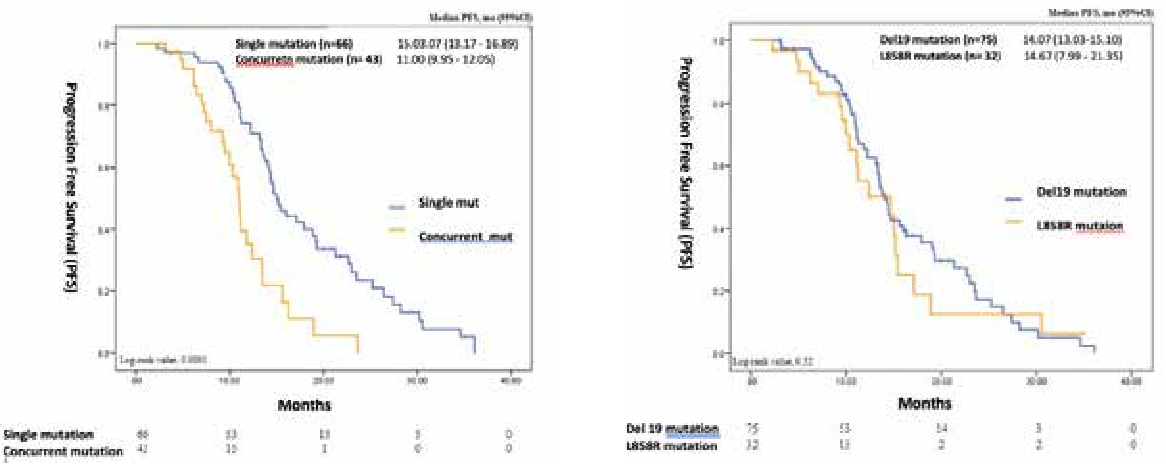

Figure 2. Comparison of PFS with EGFR-TKI treatment between single EGFR mutation and concurrent gene alterations patients

Furthermore, when we compared the clinical features and treatment effect of the two groups, we found the significant differences in PFS. The median PFS in all the 109 patients was 13.73 months (95% CI: 12.70-14.76). The PFS in the group with single EGFR mutation and concurrent gene alterations group were 15.03 months (95% CI: 13.17-16.89) và 11.00 months (95% CI: 9.95-12.05) (p = 0.0001).

Table 5. Clinical efficacy comparison of *EGFR*-TKI in single *EGFR* mutation and concurrent gene alterations

Best response	Single *EGFG* mutation (N=66) n (%)	Concurrent gene alterations (N=43) n (%)	*p*-values
CR	1 (1.5)	0 (0.0)	
PR	43 (65.2)	29 (67.4)	
SD	9(13.6)	8 (18.6)	
PD	13 (19.7)	6 (14.0)	
ORR	66.7	67.4	1.00
DCR	80.3	86.0	0.61
Median PFS (month)	15.03	11.00	0.0001*
Median OS (month)	NA	39.2.2	0.131

*significant at 0.05

The Objective Response Rate (ORR) of the total patient was 67%, while 82.6% of patient experienced Disease Control Rate (DCR). 43 patients with single EGFR mutation showed partial responses [PR] (65.2%), one with complete response [CR] (1.5%) and 9 showed stable disease [SD] (13.6%); 13 patients had progressive disease [PD] (19.7%).

IV. DISCUSSION

As mentioned, our research objective is to develop a software support system for the targeted treatment of NSCLC utilizing genetic mutation analysis data. In this report, we discuss the algorithm chosen for text classification, the accuracy of the PubMed BERT model in building the Mapping database and the characteristics of the database we built. This lays the foundation for developing AI to detect target therapies and the relationship between gene-drug responses, which will be applied in choosing optimal targeted therapy for lung cancer treatment.

Currently, in the medical field, extensive research is conducted on text classification models. In Olalekan A Uthman et al used deep learning algorithms such as parallel CNN, stacked CNN, parallel-stacked CNN, RNN and CNN- RNN to automatically classify potential research on interventions for primary prevention in cardiovascular disease.[11] The study labelled 16,611 articles (4.0% were tagged as ‘relevant’ and 96% were tagged as “irrelevant”). Among them, the algorithm with best-performing is parallel CNN, with a recall of 96.4%. To choose the best algorithm for text classification, in this study, we also have conducted research, developed, and rigorously tested several text classification models using the algorithms K-Nearest Neighbors, Bernoulli Naive Bayes, Multinomial Naive Bayes, Complement Naive Bayes, Logistic Regression, Line Support Vector Machine. After comparing the effectiveness of these models, we chose the most optimal model which is the Multinomial Naive Bayes with a Recall index of 98.12%. Such models have the potential to significantly enhance the quality of medical diagnostics, treatment, and research, offering a promising avenue for future development in the healthcare sector.

In this report, we have presented the deep learning models to build variant- gene/drug mapping modules related to common cancer types in Vietnam. Based on the results we have achieved, we see that the PubMedBERT model is the model displaying the best results for the two modules NER and RE with the F1- score for NER module reached 88.8, for example. With the results obtained above along with monitoring the operating model, it can be concluded that the model is highly effective with stable performance. The results obtained are important in helping doctors quickly map information between genes, drugs, diseases and mutations to effectively research and treat patients.[12]

Currently, lung cancer is one of the leading global causes of mortality and its incidence is rising, resulting in a significant economic burden on society. Traditional cancer treatments have primarily revolved around surgical resection, radiotherapy, and cytotoxic chemotherapy. Cytotoxic drugs inhibit carcinogenesis and tumor growth by interfering with rapidly dividing cells, but most of them have limited therapeutic efficacy and can cause serious side effects. Therefore, the development of more effective and less harmful treatment methods is extremely necessary. To address this challenge, the expert database on the relationship between gene mutations and treatment drugs was built by the authors based on open databases including COSMIC and OncoKB. COSMIC is considered the most detailed and comprehensive data resource on the effects of somatic mutations on cancer.[13] The latest 2023 release includes 23,854,105 mutations across 1,520,321 tumor samples, curated from 29,024 publications worldwide. Data is mainly managed manually, ensuring high quality, accuracy, and reliability from research works and published scientific articles. The data then compiled into a substantial data repository that comprehensively maps the responses of targeted drugs across various cancer types, encompassing those under clinical investigation and those already authorized for clinical application.

Around the world, there have been many studies conducted to build databases to serve the goal of providing evidence on targeted treatments for cancer. In 2010, author Simone Mocellin and his colleagues compiled and built a database on targeted therapy for some types of malignant cancer (Targeted Therapy Database-TTD) based on scientific publications and articles. Scientific articles on Medline, Embase, Cancerlit and Cochrane databases. In 2019, Xue Bai and his colleagues also built a targeted therapy database to store and extract data of 1088 clinical trials for 15 types of cancer and this database was granted free access.[14]

Our research is built on the study of available data from many reputable sources and databases in Vietnam and around the world including COSMIC, CIVic databases, and guidance documents. Lung cancer diagnosis and treatment guidelines from the US National Comprehensive Cancer Network (NCCN), combined with the FDA was approved status, thereby establishing a genetic relationship. -mutations-target drugs and the possibility of response or resistance to target drugs in lung cancer treatment and approved status to determine the order of priority for selecting target drugs.[21] As a result, a total of 59 genes were synthesized with 286 gene variations, and 101 therapeutic drugs had their drug licensing status determined. Of these, 32 drugs were approved by FDA, the remaining 37 drugs were still in clinical trials and 2 drugs were licensed by China's National Medical Products Administration (NMPA). Depending on each mutation and type of variation, the ability to respond and resist drugs is varied. For example, in two variants of two different genes, EGFR T790M mutation is resistant to erlotinib, while KRAS non-specific expression is highly responsive to erlotinib. In another case of the same KRAS gene, the KRAS G12A mutation variant is resistant to cetuximab while the PIK3CA_Exon 20 variant is treated effectively with the same drug. Thus, not only in one gene but also in variations of a gene, the response to treatment drugs is different[15]. Therefore, building a gene-mutation-target drug relationship to individualize patient treatment is extremely necessary.

Besides providing suggestions on appropriate targeted therapy drugs, a drug database can provide researchers with more inside on cancer metabolism and treatment mechanism. A database built by Lalu Muhammad Irham and his colleagues in 2020 on colorectal cancer target drugs has lead to 12 additional drugs be repurposed for colorectal cancer treatment, thanks to new understanding on cancer treatment.[16]

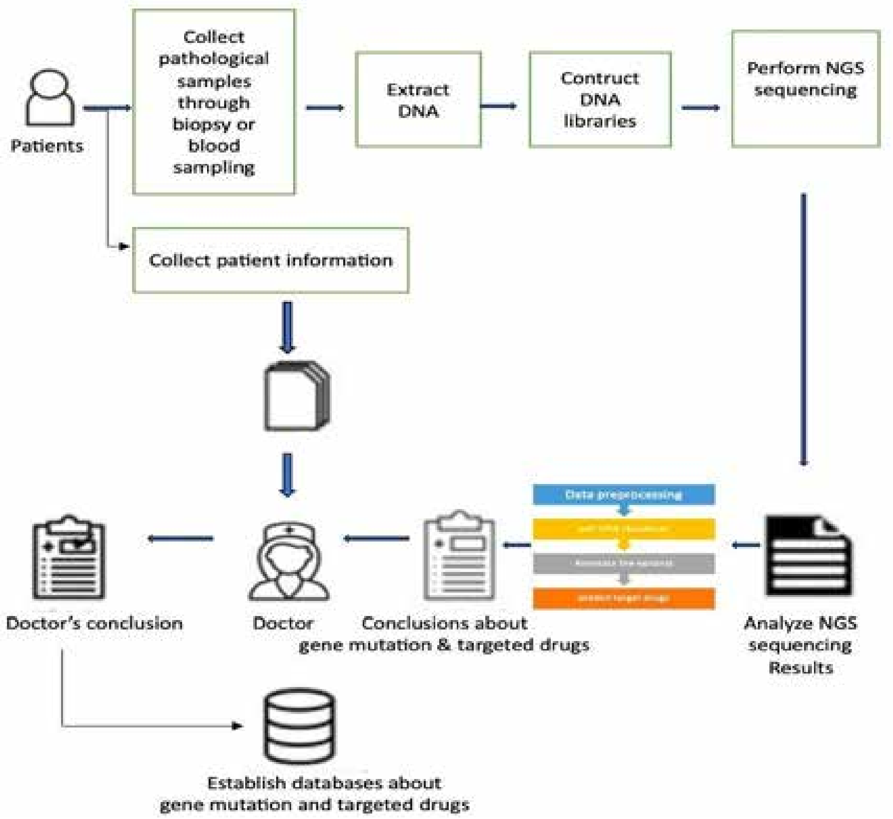
Patients
Collect pathological samples through biopsy or blood sampling
Extract DNA
Contruct DNA libraries
Perform NGS sequencing
Collect patient information
Data preprocessing
Analyze NGS sequencing Results
Conclusions about gene mutation & targeted drugs
Doctor
Doctor's conclusion
Establish databases about gene mutation and targeted drugs

Figure 3. Gene mutation and targeted drugs database aiding doctor in clinical practice

Our research has synthesized genetic variations of mutated genes, providing response or resistance to targeted drugs which serves as a valuable resource to assist clinicians in their decision-making processes when determining the most suitable treatment options. In clinical practice, after a patient's biopsy sample have been analyzed by NGS (next generation sequencing), the AI's database can provide physician with the targeted drugs corresponding to the gene mutation that have been found. In addition to FDA-approved drugs, our research also encompasses drugs that are currently undergoing clinical trials.

Furthermore, our research is updated with data compiled from various sources such as COSMIC, CIVic, treatment guidelines of NCCN 2023, Vietnam Ministry of Health, clinical trials. published on Pubmed, ClinicalTrials... This comprehensive approach ensures that our findings are not only current but also exceptionally objective, accurate, and complete which provides invaluable support to medical professionals as they make crucial decisions regarding the selection of appropriate drugs for their patients.

Our investigation reveals that the integration of AI into targeted therapy support for patients with NSCLC yields favorable outcomes in terms of PFS and response rates. Additionally, within both the single mutation and co-mutation patient cohorts, our therapeutic interventions present promising prospects. Global studies have demonstrated that targeted treatments extend PFS by 10-13 months, particularly evident in the EGFR co-mutations subgroup, which typically experiences a PFS of only 6 to 8 months.[17,18] However, our software-supported research, which facilitated precise treatment determination, significantly extended PFS to 15.03 months, notably within the 11-month group of co-mutant patients. The overall response rate (ORR) in our investigation yielded a notably high rate exceeding 80%, with no discernible discrepancies observed between the single mutation and co-mutation cohorts. Treatment outcomes within the EGFR Del19 and L858R mutation subgroups exhibited similarity, with no statistically significant differences noted. These findings underscore the potential of our research to improve treatment outcomes for a broader spectrum of patients in Vietnam and globally.

Our study has some limitations. First, the small sample was not very representative. Additionally, the availability of multiple EGFR TKIs for treating advanced NSCLC patients with EGFR mutations introduces complexity in accurately evaluating the efficacy of targeted medications. The frequency of gene mutations and the presence of missing data, as indicated in available records, may have influenced the results, although they were constrained by the limited original research information incorporated in our study. Subsequent large-scale investigations are imperative to establish a correlation between sample size and mutation frequency. Furthermore, the absence of mutation frequency and individual mutation frequencies restricts quantitative analysis. Nonetheless, as the pioneering study examining the role of multiple genes in patients with EGFR mutations, the findings carry significant clinical implications. We recommend conducting randomized clinical trials (RCTs) to precisely assess the impact of genetic mutations on the effectiveness of TKI therapy. Next-generation sequencing (NGS) should be applied to all patients with stage IV lung cancer to enhance understanding and guide treatment for more precise management of lung cancer.

V. CONCLUSSION

In conclusion, the software support system accurately predicts targeted therapy for patients with NSCLC, yielding favorable outcomes. Furthermore, our study highlights that treatment responses in patients with a sole EGFR mutation were notably more favorable compared to those with concurrent gene alterations.

FUNDING

This study was supported by the state-level scientific research project: ‘Research on the application of artificial intelligence and next-generation sequencing technologies in detecting gene mutations associated with drug response in common cancers in Vietnam’ under the program KC-4.0/19-25.

ACKNOWLEDGMENTS

We thank the patients for providing their samples in this study. The authors would like to thank the colleagues of National Lung Hospital, National Cancer Hospital, for supporting us during the research.

REFERENCES

1. Lee CK, Davies L, Wu YL, et al. Gefitinib or Erlotinib vs Chemotherapy for EGFR Mutation-Positive Lung Cancer: Individual Patient Data Meta-Analysis of Overall Survival. *J Natl Cancer Inst*. 2017; 109(6). doi:10.1093/jnci/djw279

2. Huang RX, Siriwanna D, Cho WC, et al. Lung adenocarcinoma-related target gene prediction and drug repositioning. *Front Pharmacol*. 2022; 13:936758. doi:10.3389/fphar.2022.936758.

3. Paz-Ares L, Tan EH, O’Byrne K, et al. Afatinib versus gefitinib in patients with EGFR mutation-positive advanced non-small-cell lung cancer: overall survival data from the phase IIb LUX-Lung 7 trial. *Ann Oncol Off J Eur Soc Med Oncol*. 2017; 28(2): 270-277. doi:10.1093/annonc/mdw611.

4. Lou NN, Zhang XC, Chen HJ, et al. Clinical outcomes of advanced non-small-cell lung cancer patients with EGFR mutation, ALK rearrangement and EGFR/ALK co-alterations. *Oncotarget*. 2016; 7(40): 65185-65195. doi:10.18632/oncotarget.11218.

5. Lee CK, Kim S, Lee JS, et al. Next-generation sequencing reveals novel resistance mechanisms and molecular heterogeneity in EGFR-mutant non-small cell lung cancer with acquired resistance to EGFR-TKIs. *Lung Cancer Amst Neth*. 2017; 113: 106-114. doi:10.1016/j.lungcan.2017.09.005.

6. Girard N. New Strategies and Novel Combinations in EGFR TKI-Resistant Non-small Cell Lung Cancer. *Curr Treat Options Oncol*. 2022; 23(11): 1626-1644. doi:10.1007/s11864-022-01022-7.

7. Huang J, Zhuang C, Chen J, et al. Targeted Drug/Gene/Photodynamic Therapy via a Stimuli-Responsive Dendritic-Polymer-Based Nanococktail for Treatment of EGFR-TKI-Resistant Non-Small-Cell Lung Cancer. *Adv Mater Deerfield Beach Fla*. 2022; 34(27): e2201516. doi:10.1002/adma.202201516.

8. Forbes SA, Beare D, Bindal N, et al. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. *Curr Protoc Hum Genet*. 2016;91:10.11.1-10.11.37. doi:10.1002/cphg.21.

9. Agarwal SM, Nandekar P, Saini R. Computational identification of natural product inhibitors against EGFR double mutant (T790M/L858R) by integrating ADMET, machine learning, molecular docking and a dynamics approach. *RSC Adv*. 2022; 12(26): 16779-16789. doi:10.1039/d2ra00373b.

10. Hosny A, Parmar C, Coroller TP, et al. Deep learning for lung cancer prognostication:

A retrospective multi-cohort radiomics study. *PLoS Med*. 2018; 15(11):e1002711. doi:10.1371/journal.pmed.1002711.

11. Uthman OA, Court R, Enderby J, et al. Increasing comprehensiveness and reducing workload in a systematic review of complex interventions using automated machine learning. *Health Technol Assess Winch Engl*. Published online November 30, 2022. doi:10.3310/UDIR6682.

12. Doppalapudi S, Qiu RG, Badr Y. Lung cancer survival period prediction and understanding: Deep learning approaches. *Int J Med Inf*. 2021; 148:104371. doi:10.1016/j.ijmedinf.2020.104371.

13. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019; 47(D1): D941-D947. doi:10.1093/nar/gky1015.

14. Mocellin S, Shrager J, Scolyer R, et al. Targeted Therapy Database (TTD): a model to match patient's molecular profile with current knowledge on cancer biology. *PloS One*. 2010; 5(8): e11965. doi:10.1371/journal.pone.0011965.

15. Li S, Li L, Zhu Y, et al. Coexistence of EGFR with KRAS, or BRAF, or PIK3CA somatic mutations in lung cancer: a comprehensive mutation profiling from 5125 Chinese cohorts. *Br J Cancer*. 2014; 110(11): 2812-2820. doi:10.1038/bjc.2014.210.

16. Irham LM, Wong HSC, Chou WH, et al. Integration of genetic variants and gene network for drug repurposing in colorectal cancer. *Pharmacol Res*. 2020; 161: 105203. doi:10.1016/j.phrs.2020.105203.

17. Chen M, Xu Y, Zhao J, et al. Concurrent Driver Gene Mutations as Negative Predictive Factors in Epidermal Growth Factor Receptor-Positive Non-Small Cell Lung Cancer. *EBioMedicine*. 2019; 42: 304-310. doi:10.1016/j.ebiom.2019.03.023.

18. Hu W, Liu Y, Chen J. Concurrent gene alterations with EGFR mutation and treatment efficacy of EGFR-TKIs in Chinese patients with non-small cell lung cancer. *Oncotarget*. 2017; 8(15): 25046-25054. doi:10.18632/oncotarget.15337.