**AI-Guided Chemotherapy Optimization in Lung Cancer Using Genomic and Survival Data**

Moon, Phanny, Jenny, Minho, Liz

**Abstract**

This study integrates advanced bioinformatics and machine learning methodologies to develop prognostic genomic markers for optimizing chemotherapy efficacy in non-small cell lung cancer patients. By utilizing high-dimensional genomic datasets from multiple independent clinical trials available in the NCBI Gene Expression Omnibus repository, we constructed a meta-database to address population heterogeneity. Our approach employes a bagging with regularized Cox proportional hazards models, Random Survival Forests, and Deep Survival Networks to improve treatment decision-making. Our results demonstrate that AI-driven survival analysis enhances patient stratification for adjuvant chemotherapy, minimizing unnecessary exposure while improving survival outcomes. This framework significantly contributes to personalized medicine by refining chemotherapy treatment options, potentially reducing toxicity risks, and improving clinical decision-making.

1. **Introduction**

Lung cancer remains the leading cause of cancer-related mortality worldwide, posing a significant health challenge. In 2018, the Global Cancer Observatory (GLOBOCAN) reported approximately 2.09 million new cases of lung cancer, making it the most diagnosed malignancy across all populations (Bray et al., 2018). In the United States, lung cancer accounts for 12.7% of all cancer diagnosis, with an estimated 229,000 new cases reported in 2020 (Thandra et al., 2021). Among lung cancer subtypes, non-small cell lung cancer (NSCLC) constitutes nearly 85% of cases, encompassing a broad stectrum of histological types, including adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Despite advances in screening and treatment, long-term survival remains poor, primarily due to high recurrence rates and limited therapeutic efficacy in certain patient subgroups.

For patients with Stage I and II resectable NSCLC, surgical resection remains the primary curative treatment. However, adjuvant chemotherapy (ACT) is commonly recommended for Stage II and select Stage IB-III cases, particularly for patients with high-risk features such as lymph node involvement, large tumor size, or specific molecular markers (Lemjabbar-Alaoui et al., 2015). While ACT has demonstrated a survival benefit of 4% to 15% in Stage IB-IIIA NSCLC (Pirker and Filipits 2019), not all patients respond favorably to chemotherapy. The decision to administer ACT is complicated by substantial interpatient variability, where some patients experience significant survival benefits, while others endure severe toxicities with minimal therapeutic gain. This implies the critical need for personalized treatment strategies that accurately identify patients who are most likely to benefit from chemotherapy, minimizing unnecessary exposure to toxicity.

Advancements in genomic profiling have provided novel insights into the molecular drivers of NSCLC and the factors influencing chemotherapy response. Traditional clinical parameters, such as tumor size, nodal status, and histology, do not fully capture the biological heterogeneity of NSCLC. In contrast, genomic biomarkers offer a deeper understanding of tumor-specific molecular alterations that govern disease progression and treatment response. The integration of genomic data into clinical decision-making has become a central focus of precision oncology, allowing for personalized treatment approaches based on tumor-specific gene expression patterns.

Genomic analysis enables the identification of predictive biomarkers that can distinguish chemotherapy-responsive subgroups from those who may not benefit from ACT. This approach facilitates a precision medicine framework, where treatment is customized to the unique molecular characteristics of each patient, leading to improved efficacy and reduced toxicity. Additionally, genomic profiling has facilitated the identification of mutations driving cancer growth, which have paved the way for the development of targeted therapies that directly inhibit tumor progression. However, despite the growing body of evidence supporting the utility of genomic biomarkers, there is still no standardized framework for integrating genomic and clinical data into chemotherapy decision-making in NSCLC.

The application of artificial intelligence (AI) and machine learning (ML) in oncology presents a transformative opportunity for optimizing ACT selection in NSCLC. AI-driven algorithms have the capacity to process large-scale genomic datasets, identify complex interactions between molecular and clinical variables, and develop predictive models that can refine chemotherapy recommendations. Unlike traditional methods, AI-based approaches can detect non-linear relationships, capture subtle genomic variations, and enhance the accuracy of patient stratification.

Machine learning models such as regularized Cox regression (Zou and Hastie, 2005) and Random Survival Forests (RSF) (Ishwaran et al., 2008) offer distinct advantages in predicting patient-specific ACT benefit. These approaches have been shown to outperform conventional survival models, particularly in the context of high-dimensional genomic data, where feature selection and dimensionality reduction are crucial for improving model robustness. AI can also facilitate subgroup analysis, identifying clusters of patients with distinct survival trajectories based on their genomic profiles. Such insights can lead to more informed treatment decisions, ensuring that ACT is administered selectively to those who will derive meaningful survival benefits.

Preliminary research efforts have sought to integrate genomic markers with statistical modeling to improve chemotherapy decision-making. Moon et al. (2018) employed lasso-regularized Cox regression and Random Forests (Breiman, 2001) classifiers to predict ACT benefit using gene expression signatures in early-stage NSCLC. The study demonstrated that genomic-based predictive modeling could improve treatment stratification, potentially reducing unnecessary chemotherapy exposure in patients with a low probability of benefit.

Moon et al. (2020) developed a statistical decision support tool to identify key risk factors and assess the likelihood of benefit from ACT in early-stage NSCLC patients. By utilizing genome-wide microarray data, their study identified genomic markers predictive of treatment response, enabling stratification of patients who would benefit from chemotherapy in addition to surgery versus those for whom surgical resection alone would suffice. The study employed Accelerated Failure Time (AFT) models to estimate the probability of chemotherapy benefit based on genomic markers, facilitating individualized treatment recommendations. To enhance model performance, tree-based ensemble algorithms and Cox regression with elastic net regularization were applied to gene expression data, refining the identification of relevant genomic markers. Overall, their study demonstrated the potential of genomic-guided ACT selection in optimizing therapeutic strategies for lung cancer patients, thereby improving prognosis and survival outcomes.

Our recent study (Moon et al., 2021) continues to aim at identifying predictive genomic biomarkers that can distinguish subgroups of early-stage lung cancer patients likely to benefit from ACT. A modified-covariate regularized Cox regression model with a lasso penalty was employed on the JBR.10 dataset (Zhu et al., 2010). The study identified ACT as a more favorable prognosis in advance cases by highlighting the importance of the predictive model in guiding personalized treatment recommendations and improving survival outcomes for NSCLC patients.

Our study sets the stage for significant advancement in the field of clinical decision support for lung cancer treatment. By introducing innovative methodologies, including a bagging approach based on the penalized Cox PH models, the Random Survival Forest (RSF) algorithm, and a deep learning survival network (DeepSurv), our goal is to enhance the robustness and accuracy of our predicted treatment recommendation system. These advanced machine learning algorithms not only enable more precise predictions of treatment outcomes but also conforms recommendations to individual patient profiles, ultimately optimizing therapeutic strategies. This meth(odology holds the potential to substantially improve clinical outcomes by facilitating more informed, personalized treatment decisions for patients with NSCLC.

Despite these advances, further refinement of AI-based predictive models is necessary to enhance their clinical utility. The integration of diverse multi-omics datasets, incorporation of clinical data, and external validation of predictive models in independent patient cohorts are crucial next steps. In this paper, we demonstrate how utilizing AI-driven survival analysis and genomic profiling enhances personalized chemotherapy decision-making, ultimately leading to improved clinical outcomes for early-stage NSCLC patients.

**Reference**

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424. doi: [10.3322/caac.21492](https://doi.org/10.3322/caac.21492).
2. Thandra KC, Barsouk A, Saginala K, Aluru JS, Barsouk A. Epidemiology of lung cancer. *Contemp Oncol (Pozn)*. 2021;25(1):45-52. doi: [33911981](https://pubmed.ncbi.nlm.nih.gov/33911981/).
3. Lemjabbar-Alaoui H, Hassan OU, Yang YW, Buchanan P. Lung cancer: Biology and treatment options. *Biochim Biophys Acta*. 2015;1856(2):189-210. doi: [10.1016/j.bbcan.2015.08.002](https://doi.org/10.1016/j.bbcan.2015.08.002).
4. Pirker R, Filipits M. Adjuvant Therapy in Patients With Completely Resected Non-small-cell Lung Cancer: Current Status and Perspectives. *Clin Lung Cancer*. 2019;20(1):1-6. doi: [10.1016/j.cllc.2018.09.016](https://doi.org/10.1016/j.cllc.2018.09.016) .
5. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301-20. doi: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)
6. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-60. <https://doi.org/10.1214/08-AOAS169>
7. Moon H, Zhao Y, Pluta D, Ahn H. Subgroup analysis based on prognostic and predictive gene signatures for adjuvant chemotherapy in early-stage non-small-cell lung cancer patients. *J Biopharm Stat.* 2018;28(4):750-762. doi: [10.1080/10543406.2017.1397006.](https://doi.org/10.1080/10543406.2017.1397006)
8. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi: 10.1023/A:1010933404324.
9. Moon H, Chao T, Ahn H. Identification of Risk Factors and Likelihood of Benefit from Adjuvant Chemotherapy for Early Stage Lung Cancer Patients. *J Biopharm Stat*. 2020;30(3):430-444. doi: [10.1080/10543406.2019.1684310](https://doi.org/10.1080/10543406.2019.1684310)
10. Moon H, Nguyen A, Lee E. Prognostic Genomic Predictive Biomarkers for Early-Stage Lung Cancer Patients. *The Open Biomarkers Journal*. 2021;11(1):69–78. [doi: 10.2174/1875318302111010069](http://dx.doi.org/10.2174/1875318302111010069)
11. Zhu CQ, Ding K, Strumpf D, Weir BA, Meyerson M, Pennell N, Thomas RK, Naoki K, Ladd-Acosta C, Liu N, Pintilie M, Der S, Seymour L, Jurisica I, Shepherd FA, Tsao MS. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J Clin Oncol*. 2010;28(29):4417-24. doi: [10.1200/JCO.2009.26.4325](https://doi.org/10.1200/jco.2009.26.4325)

4. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. Transl Lung Cancer Res 2014;3(4):242-249.

10. Cox, D. R. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187–202.

11. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67(2):301-20. doi:10.1111/j.1467-9868.2005.00503.x.

14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010;33(1):1–22. doi:10.18637/jss.v033.i01.

15. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841-60. doi:10.1214/08-AOAS169.

16. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966;50(3):163–70.

17. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018;18(1):24. doi:10.1186/s12874-018-0482-1.

18. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. Adv Neural Inf Process Syst. 2012;25:2951–9.

19. Jepkoech J, Mugo DM, Kenduiywo BK, Too EC. The effect of adaptive learning rate on the accuracy of neural networks. Int J Adv Comput Sci Appl. 2021;12(8). Available from: https://doi.org/10.14569/ijacsa.2021.0120885

20. Bjorck J, Gomes C, Selman B, Weinberger K. Understanding Batch Normalization. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS’18). Red Hook, NY: Curran Associates Inc.: 2018:7705–16. Available from: https://dl.acm.org/doi/pdf/10.5555/3327757.3327868.

21. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, CA: Curran Associates Inc.; 2017:971-80. Available from: https://proceedings.neurips.cc/paper/2017/file/5d44ee6f2c3f71b73125876103c8f6c4-Paper.pdf

22. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543–6. doi:10.1001/jama.1982.03320430047030.

23. Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ. On the C-statistic for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105–1117. doi:10.1002/sim.4154

24. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

25. Pencina MJ, D’Agostino RB. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. Stat Med. 2004;23(13):2109–2123. doi:10.1002/sim.1802

26. Steyerberg EW, Van Calster B, Pencina MJ. Performance measures for prediction models and their evaluation in the presence of censoring. Stat Methods Med Res. 2018;27(9):2504–2525. doi:10.1016/j.recesp.2011.04.017

27. SynGO Consortium. SynGO - ID Conversion Tool. SynGo.org; 2019. Available from: https://www.syngoportal.org/convert.

28.