Editorial on Immunotherapy and Tumor Microenvironment

Multiomics and machine learning in lung cancer prognosis

Yanan Gao[#], Rui Zhou[#], Qingwen Lyu

Department of Information, Zhujiang Hospital, Southern Medical University, Guangzhou, China

[#]These authors contributed equally to this work.

Correspondence to: Dr. Qingwen Lyu. Department of Information, Zhujiang Hospital, Southern Medical University, 253 Industrial Avenue, Guangzhou 510282, China. Email: 832241@qq.com.

Submitted Mar 10, 2020. Accepted for publication Jul 13, 2020.
doi: 10.21037/jtd-2019-itm-013
View this article at: http://dx.doi.org/10.21037/jtd-2019-itm-013

Worldwide, lung cancer accounts for 11.6% of total cancer cases; it is the most common cancer type and the leading cause of cancer death (1). Despite the development of technology and treatment, the prognosis of lung cancer remains poor (2-5). With the development of artificial intelligence technology and the advent of omics, including radiomics, proteomics, genomics, and transcriptomics (6-8), multiomics analysis based on machine learning has great potential to improve lung cancer prognosis. In this paper, schemes based on multiomics and machine learning for improving the prognosis of lung cancer are reviewed.

Radiomics, pathology, demographics, clinical data and machine learning in lung cancer prognosis

Currently, radiomics research and medical or biological research are usually carried out separately by researchers in different disciplines. However, with the emergence of the field of radiomics, the association between biomarkers and radiomics features has attracted increasing research interest (9).

As shown in *Table 1*, in 2010, Jayasurya *et al.* (10) used radiomic features from positron emission tomography (PET) images, pathological features, and performance status (WHO-PS) to develop two personalized prediction models based on Bayesian networks (BNs) and support vector machines (SVMs) to predict the 2-year survival rate of patients with inoperable non-small cell lung cancer (NSCLC). The authors validated the models in three external validation cohorts from three centres. Among them, the area under the curve (AUC) of the prediction model based on BNs reached 0.82 in a cohort of 28 patients. However, the validation cohort in this work was small, and more research on the clinical utility of the model is needed to confirm the results. In 2013, Sun *et al.* (11) attempted to differentiate benign from malignant lung cancer according to computed tomography (CT) images during early diagnosis to improve prognosis. This group used SVMs and other classifiers, including neural networks, LASSO regressions, boosting, random forests, decision trees, and k-nearest neighbours, to establish prediction models. A set of radiomics features, including 476 textural and 9 morphological features, and demographic parameters were used as input data. The AUC values for the SVM, neural networks, LASSO regressions, boosting, random forests, decision trees, and k-nearest neighbours were 0.94, 0.92, 0.91, 0.86, 0.85, 0.73, and 0.72, respectively. Although the experimental results showed that the SVM-based model was effective, only 57 patients were included in the validation cohort. Recently, Hyun *et al.* (12) performed a similar study in which a total of 44 demographic and radiomic features were used as input data for a machine-learning model to predict tumour histological subtype. To reduce feature dimensions, they applied a ranking-based feature selection method with the Gini coefficient. By evaluating radiomic and demographic features' associations with the histological class, they obtained Gini coefficient scores and then ranked these features based on the Gini coefficient. Nine feature subsets were selected to identify the optimal feature selection size, ranging from 5 to 44 in increments of 5. Five different machine-learning algorithms for binary

Table 1 Summation of journal publications

Author	Data type	Model	Test-set size	AUC
Jayasurya *et al.* (10)	Radiomic features; pathological features; WHO-PS	Bayesian networks	28	0.76
		Support vector machines		0.82
Sun *et al.* (11)	Radiomic features; demographic features	Support vector machines	57	0.94
		Neural networks		0.92
		LASSO regression		0.91
		Boosting		0.86
		Random forest		0.85
		Decision tree		0.73
Hyun *et al.* (12)	Radiomic features; demographic features	K-nearest neighbours	119	0.85
		Random forest		0.79
		Neural network		0.854
		Naïve Bayes		0.755
		Logistic regression		0.859
		Support vector machines		0.766

AUC, area under the curve.

classification, namely, a random forest, a neural network, a naïve BN method, a logistic regression model, and SVM, were evaluated. When using a subset with 15 features, the logistic regression model (AUC =0.859) performed better than other classifiers.

Genomics, transcriptomics, genetics, proteomics and machine leaning in lung cancer prognosis

In addition to radiomics, pathology, and demographics, there is research interest with regard to the genomics, transcriptomics, genetics and proteomics of lung cancer prognosis.

Wang *et al.* (13) presented a method to construct a prediction model of EGFR mutation-induced drug resistance in lung cancer by combining pathological and demographic data and EGFR-inhibitor interaction patterns. In this method, they initially translated mutations into 3D structures, after which the binding free energies of the mutants and inhibitors were evaluated and the dynamics of the kinase mutant-inhibitor systems were simulated. The EGFR-inhibitor interaction was characterized by binding free energy components, including polar and nonpolar interactions, van der Waals forces and electrostatic interactions. The classification model was built by extreme learning machines, and they also conducted a comparison between a model involving only the mutation feature and a model involving multiomics features, with the latter (classification accuracy of 95.13%) being much better than the former (classification accuracy of 79.17%). In 2015, Emaminejad *et al.* (14) integrated two genomic biomarkers and radiomic features to predict recurrence risk in patients with stage I NSCLC: they trained a multilayer perceptron-based model using two genomic features (protein expression scores of RRM1 and ERCC1) and a naïve BN classifier using eight redundant radiomic features to predict cancer recurrence risk. The AUC values of the multilayer perceptron classifier and naïve BN classifier were 0.68±0.06 and 0.78±0.07, respectively. Moreover, the AUC value increased significantly (0.84±0.05, $P<0.05$) when an equal weighting factor to fuse the prediction scores generated by the two models was used. In 2017, Yu *et al.* (15) used random forest, transcriptomics, and proteomics signatures to predict histology grade (AUC >0.80), building integrative models by using histopathologic and transcriptomic features as input data of the regularized Cox proportional hazards model; the integrative model outperformed transcriptomics or histopathology alone for prognostic prediction ($P=0.0182\pm0.0021$). Additionally, Liu *et al.* (16)

Table 2 Accuracy and AUC value of journal publications based on different datasets and models

Author	Data type	Models	Accuracy (%)	AUC
Wang *et al.* (13)	Mutation features	Extreme learning machines	79.17	NA
	Mutation features; pathological features; demographic features		95.83	
Emaminejad *et al.* (14)	Genomic features	Multilayer perceptron; Naïve Bayes	NA	0.68
	Radiomic features			0.78
	Integrated dataset			0.84
Yu *et al.* (15)	Genomic features; transcriptomics/proteomics features; histopathology features	Random forest	NA	0.81
Matsubara *et al.* (17)	PPI network; gene expression	Convolutional networks	83.16	NA
		Radom forest	82.63	
		Support vector machines	81.58	
Malik *et al.* (18)	Copy number variations; mutation; protein; RNA; mi-RNA	Support vector machines	72.7	NA
		Neural network	92.9	
		RUS ensemble boost	66.7	
Giang *et al.* (20)	Gene expression	Support vector machines	62.50	0.6964
	DNA methylation		71.88	0.6235
	mi-RNA expression		65.63	0.722
	Integrated dataset		78.13	0.7227

AUC, area under the curve; RUS, random undersampling; NA, not available; PPI, protein-protein interaction.

identified a novel cluster of prognostic biomarkers for lung adenocarcinoma (LAC) by multiomics analysis. In this work, five microarray datasets downloaded from the Gene Expression Omnibus database were progressively processed by genome-wide relative significance and global significance, and 200 genes able to stably distinguish between nontumour and tumour cells were determined by SVM assessment. These genes were then subjected to gene coexpression and protein-protein interaction (PPI) network analyses. CENPA, CDC20 and CDC20 were identified and validated as having high coexpression and strong PPI patterns in clinical samples, and CENPA, CDC20 and CDK1 might serve as a novel cluster of prognostic biomarkers in LAC. In 2018, Matsubara *et al.* (17) proposed an approach to lung cancer classification that integrates PPI network and gene expression profile data as input features of a convolutional network; comparisons between convolutional networks and other machine-learning models (random forest and SVM) were also conducted. The model-based convolutional network (accuracy rate was 83.16%) outperformed the model-based SVM and random forest methods (accuracy rates were 81.58% and 82.63%, respectively). Malik *et al.* (18) in 2019 utilized copy number variations (CNVs), mutations, proteins, RNAs and mi-RNAs to develop three prediction models for LAC prognosis based on SVMs, neural network and random undersampling (RUS) ensemble boosts, with accuracies of 72.7%, 92.9% and 66.7%, respectively. To acquire more omics information, Lee *et al.* (19) investigated four data features, including DNA methylation, RNA-Seq, CNVs and miRNA-Seq, to build a survival risk stratification model for LAC patients. They proposed an autoencoding approach to predict survival subtype, compared to other approaches, principal component analysis (PCA), Cox-ph and iClusterPlus. As the autoencoding approach has a better log-rank P value (4.08e-09) and C-index (0.65), autoencoding exhibits better prediction performance. Recently, Giang *et al.* (20) presented a method that combines gene expression, miRNA expression and DNA methylation data features to construct a classification model of lung cancer patient stratification. SVM was used for building a classification model, and a comparison between the approach involving an integrated dataset and that in which only a single dataset was used was also conducted. *Table 2* shows the accuracy and AUC value

of the models based on different datasets and models.

Overall, a growing number of studies have combined machine learning with multiomics analysis to improve the prognosis of lung cancer (21-25), and radiomics, genetics, genomics, proteomics, and transcriptomics are widely employed in the fields of lung cancer. Although the validation cohort in many related studies is relatively small, it is sufficient to indicate that multiomics analysis based on machine learning has great potential in lung cancer prognosis, and more schemes in this field will be developed to improve prognosis for these patients. I hope that this review will be of use to researchers who conduct related research.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editors (Peng Luo, Clare Y. Slaney and Jian Zhang) for the series "Immunotherapy and Tumor Microenvironment" published in *Journal of Thoracic Disease*. The article did not undergo external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jtd-2019-itm-013). The series "Immunotherapy and Tumor Microenvironment" was commissioned by the editorial office without any funding or sponsorship. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Mountzios G, Dimopoulos MA, Soria JC, et al. Histopathologic and genetic alterations as predictors of response to treatment and survival in lung cancer: a review of published data. Crit Rev Oncol Hematol 2010;75:94-109.
3. Ramalingam SS, Owonikoko TK, Khuri FR. Lung cancer: New biological insights and recent therapeutic advances. CA Cancer J Clin 2011;61:91-112.
4. Siegel RL, Fedewa SA, Miller KD, et al. Cancer statistics for Hispanics/Latinos, 2015. CA Cancer J Clin 2015;65:457-80.
5. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. CA Cancer J Clin 2016;66:115-32.
6. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89-109.
7. Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 2009;45:S199-209.
8. Zhang W, Li F, Nie L. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology 2010;156:287-301.
9. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006. Erratum in: Nat Commun. 2014;5:4644. Cavalho, Sara [corrected to Carvalho, Sara].
10. Jayasurya K, Fung G, Yu S, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys 2010;37:1401-7.
11. Sun T, Wang J, Li X, et al. Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set. Comput Methods Programs Biomed 2013;111:519-24.
12. Hyun SH, Ahn MS, Koh YW, et al. A Machine-Learning Approach Using PET-Based Radiomics to Predict the Histological Subtypes of Lung Cancer. Clin Nucl Med 2019;44:956-60.
13. Wang DD, Zhou W, Yan H, et al. Personalized prediction of EGFR mutation-induced drug resistance in lung cancer. Sci Rep 2013;3:2855.

14. Emaminejad N, Qian W, Guan Y, et al. Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients. IEEE Trans Biomed Eng 2016;63:1034-43.
15. Yu KH, Berry GJ, Rubin DL, et al. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. Cell Syst 2017;5:620-627.e3.
16. Liu WT, Wang Y, Zhang J, et al. A novel strategy of integrated microarray analysis identifies CENPA, CDK1 and CDC20 as a cluster of diagnostic biomarkers in lung adenocarcinoma. Cancer Lett 2018;425:43-53.
17. Matsubara T, Nacher JC, Ochiai T, et al. Convolutional Neural Network Approach to Lung Cancer Classification Integrating Protein Interaction Network and Gene Expression Profiles. 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE). Taichung; 29-31 Oct. 2018. IEEE, 2018:151-4.
18. Malik V, Dutta S, Kalakoti Y, et al. Multi-omics Integration based Predictive Model for Survival Prediction of Lung Adenocarcinoma. 2019 Grace Hopper Celebration India (GHCI). Bangalore, India; 6-8 Nov. 2019. IEEE, 2019.
19. Lee TY, Huang KY, Chuang CH, et al. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. Comput Biol Chem 2020;87:107277.
20. Giang TT, Nguyen TP, Tran DH. Stratifying patients using fast multiple kernel learning framework: case studies of Alzheimer's disease and cancers. BMC Med Inform Decis Mak 2020;20:108.
21. Song P, Cui X, Bai L, et al. Molecular characterization of clinical responses to PD-1/PD-L1 inhibitors in non-small cell lung cancer: Predictive value of multidimensional immunomarker detection for the efficacy of PD-1 inhibitors in Chinese patients. Thorac Cancer 2019;10:1303-9.
22. Zhong T, Wu M, Ma S. Examination of Independent Prognostic Power of Gene Expressions and Histopathological Imaging Features in Cancer. Cancers (Basel) 2019;11:361.
23. Levitsky A, Pernemalm M, Bernhardson BM, et al. Early symptoms and sensations as predictors of lung cancer: a machine learning multivariate model. Sci Rep 2019;9:16504.
24. Huang P, Park S, Yan R, et al. Added Value of Computer-aided CT Image Features for Early Lung Cancer Diagnosis with Small Pulmonary Nodules: A Matched Case-Control Study. Radiology 2018;286:286-95.
25. Alcala N, Leblay N, Gabriel AAG, et al. Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. Nat Commun 2019;10:3407.

Cite this article as: Gao Y, Zhou R, Lyu Q. Multiomics and machine learning in lung cancer prognosis. J Thorac Dis 2020;12(8):4531-4535. doi: 10.21037/jtd-2019-itm-013

