JAMA Network Open

Original Investigation | Oncology

Development and Validation of a Deep Learning Model for Non–Small Cell Lung Cancer Survival

Yunlang She, MD; Zhuochen Jin, BS; Junqi Wu, MD; Jiajun Deng, MD; Lei Zhang, MD; Hang Su, MD; Gening Jiang, MD; Haipeng Liu, PhD; Dong Xie, MD, PhD; Nan Cao, PhD; Yijiu Ren, MD; Chang Chen, MD, PhD

Abstract

IMPORTANCE There is a lack of studies exploring the performance of a deep learning survival neural network in non–small cell lung cancer (NSCLC).

OBJECTIVES To compare the performances of DeepSurv, a deep learning survival neural network with a tumor, node, and metastasis staging system in the prediction of survival and test the reliability of individual treatment recommendations provided by the deep learning survival neural network.

DESIGN, SETTING, AND PARTICIPANTS In this population-based cohort study, a deep learning–based algorithm was developed and validated using consecutive cases of newly diagnosed stages I to IV NSCLC between January 2010 and December 2015 in a Surveillance, Epidemiology, and End Results database. A total of 127 features, including patient characteristics, tumor stage, and treatment strategies, were assessed for analysis. The algorithm was externally validated on an independent test cohort, comprising 1182 patients with stage I to III NSCLC diagnosed between January 2009 and December 2013 in Shanghai Pulmonary Hospital. Analysis began January 2018 and ended June 2019.

MAIN OUTCOMES AND MEASURES The deep learning survival neural network model was compared with the tumor, node, and metastasis staging system for lung cancer–specific survival. The C statistic was used to assess the performance of models. A user-friendly interface was provided to facilitate the survival predictions and treatment recommendations of the deep learning survival neural network model.

RESULTS Of 17 322 patients with NSCLC included in the study, 13 361 (77.1%) were white and the median (interquartile range) age was 68 (61-74) years. The majority of tumors were stage I disease (10 273 [59.3%]) and adenocarcinoma (11 985 [69.2%]). The median (interquartile range) follow-up time was 24 (10-43) months. There were 3119 patients who had lung cancer–related death during the follow-up period. The deep learning survival neural network model showed more promising results in the prediction of lung cancer–specific survival than the tumor, node, and metastasis stage on the test data set (C statistic = 0.739 vs 0.706). The population who received the recommended treatments had superior survival rates than those who received treatments not recommended (hazard ratio, 2.99; 95% CI, 2.49-3.59; $P < .001$), which was verified by propensity score–matched groups. The deep learning survival neural network model visualization was realized by a user-friendly graphic interface.

CONCLUSIONS AND RELEVANCE The deep learning survival neural network model shows potential benefits in prognostic evaluation and treatment recommendation with respect to lung

(continued)

Key Points

Question Can deep learning architecture be applied for individual prognosis evaluation and treatment recommendation?

Findings In this cohort study of 17 322 patients with non–small cell lung cancer. The performance of a deep learning model was assessed on real-life clinical data sets. The ability of a deep learning model to learn complex associations between an individual's characteristics and the outcome benefits of different treatments was also elucidated; particularly, a deep learning network identified persons with non–small cell lung cancer and survival more accurately than tumor, node, metastasis staging.

Meaning Findings suggest that this novel analytical approach may have great potential in providing individual prognostic information and treatment recommendations in real clinical scenarios.

+ Invited Commentary

+ Video

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

cancer-specific survival. This novel analytical approach may provide reliable individual survival information and treatment recommendations.

JAMA Network Open. 2020;3(6):e205842. doi:10.1001/jamanetworkopen.2020.5842

Introduction

Lung cancer is the most commonly diagnosed cancer in China and the second in the United States, approximately 85% of which is non–small cell lung cancer (NSCLC).[1] The precise stratification of patients with NSCLC into groups according to survival outcomes represents a crucial step in treatment. The staging system in the 8th edition of the American Joint Committee on Cancer classifies patients based on tumor, node, and metastasis (TNM) staging.[2] However, the survival rate within the same stage cohort varies widely.[3-5] It has been found that other independent prognostic factors including age, sex, histology, and treatment choices could significantly contribute to individualized predictions of survival.[6]

To improve the precision of lung cancer survival estimations, Cox proportional hazard models have gained popularity as a way of predicting outcomes.[7] For example, the nomogram is a reliable tool that has demonstrated the ability to quantify risk by combining and clarifying significant clinical characteristics for clinical oncology.[6] By drafting a concise chart of an outcome-risk predictive model, a nomogram derives the risk probability of a specific event, such as lung cancer–specific survival (LCSS). In various cancers, nomograms possess the ability to derive more precise risk predictions when incorporated with TNM staging.[8,9] However, these models have several limitations with respect to time-to-event prediction for the clinical management of patients with cancer, including the precise evaluation of overall survival and time to progression.[10] Moreover, these models make linearity assumptions rather than perform nonlinear analyses that reflect real-world clinical characteristics.[11] Therefore, there is a need for better solutions that focus on nonlinear variables.[12]

Deep learning networks can learn the highly intricate and linear/nonlinear associations between prognostic clinical characteristics and an individual's risk of death from LCSS.[13] In application, these networks have even shown potential for providing individual recommendations based on the calculated risk.[14] For example, by analyzing clinical data in the Surveillance, Epidemiology, and End Results (SEER) cancer registry, Bergquist et al[15] assembled computerized methods including random forests, lasso regression, and neural networks to achieve 93% accuracy in predicting lung cancer stages. In another study, Corey et al[16] developed a software package (Pythia) based on machine learning models that incorporated a patient's age, sex, clinical baseline, race/ethnicity, and comorbidity history to determine the risk of postoperative complications or deaths. Matsuo et al[17] also developed a deep learning network model that has demonstrated a higher C statistic than the traditional proportional hazard regression model (C statistic = 0.795 vs 0.784) for progression-free survival analysis. Furthermore, Katzman et al[18] developed a novel deep learning method for survival analysis that uses a deep learning network to integrate Cox proportional hazards, which is referred to as the learning survival neural network (DeepSurv). The authors demonstrated that DeepSurv performed as well as published survival models and could be used to provide treatment recommendations for better survival outcomes.

The present study design follows the American Joint Committee on Cancer criteria for model adoption and the transparent report of a deep learning architecture for individual prognosis and treatment recommendation. In this study, we first describe the performance of DeepSurv[18] on real-life clinical data sets. Second, we elucidate how the DeepSurv model can learn complex associations between an individual's characteristics and the outcome benefits of different treatments.

Methods

Eligibility Criteria and Clinical Information

All patients gave informed oral consent prior to data collection. After obtaining institutional review board approval from Shanghai Pulmonary Hospital, we selected patients from the SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2017 Sub, which includes clinical records on cancer occurrences in 18 areas of the United States and contains approximately 27.8% of the population. Clinical cases were included if the following criteria were met: pathologically confirmed primary stage I to IV NSCLC (only adenocarcinoma and squamous cell carcinoma) between January 2010 and December 2015 and the presence of 1 malignant primary lesion. From the SEER database (eTable 1 in the Supplement), we collected the baseline information of cases (sex, age, and marriage status), tumor characteristics (location, size, histologic grade, histologic type, TNM stage, SEER code (CS extension, CS mets at dx, regional nodes examined, regional nodes positive, lung–pleural/elastic layer invasion by H and E or elastic stain, lung–separate tumor nodules–ipsilateral lung, lung–surgery to primary site]), and treatment details (surgical type).[19,20] Patients were excluded if any of the included clinical characteristics status were unknown or missing. The outcome of interest in this study included LCSS according to specific codes provided by SEER (defined as the interval from surgery until death as a result of lung cancer). These patients were randomly divided into the training and validation cohort at a ratio of 8 to 2. To validate the DeepSurv model, an external test cohort was provided by the CHINA database. The cohort comprised 1182 patients with stage I to III NSCLC diagnosed between January 2009 and December 2013 in Shanghai Pulmonary Hospital, which are completely distinct from the patients in SEER database. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline.

Deep Learning Model Design

In this study, DeepSurv was used to analyze patient-individual survival outcomes, which is a deep learning algorithm that can predict individual survival risk values (**Figure 1**).[18] We use deep feed-forward neural network and the Cox proportional hazards model in survival analysis. The DeepSurv model contained a core hierarchical structure with fully connected feed-forward neural networks with a single output node to calculate the survival risks $h\theta(x_i)$ of patients using the negative log-partial likelihood function. More details about the DeepSurv were described in the eMethods in the Supplement. Using the provided data set, we compared the performances of the TNM staging model and our deep learning model with respect to 2 tasks (LCSS predictions) with 3 different data sets (Figure 1).

Data Analysis

First, we developed a 6-layer neural network for predicting patient LCSS in the NSCLC training data set (n = 12 912). To validate the prediction performance, we used Harrell C statistic and calibration plots to evaluate the network discrimination and calibration in the NSCLC validation data set (n = 3228) and CHINA data set (n = 1182).

Next, we trained a personalized treatment recommendation system using separately developed lobar and sublobar resection risk prediction models with a 3-layer neural network in the lobectomy (n = 10 766) and sublobectomy (n = 1444) training data sets. For each patient in the lobectomy and sublobectomy validation data set (SEER: n = 3064; CHINA: n = 1142), we chose the lower-risk value of the model's treatment as the recommendation.

Finally, we categorized the patients into 2 groups according to the consistency of the treatment received and recommended. For survival analysis, we used the Kaplan-Meier method to analyze LCSS between different groups and the log-rank test to compare survival curves.

An additional Cox proportional hazard regression model with nonneural network methods[6,17] was performed following the simple backward-stepwise approach using all the variables included in

the DeepSurv model. It estimated the risk function $h(x_i)$ of the event occurring (LCSS) for patient i based on included features x_i using a linear function: $h(Xi) = \Sigma\, x_i\beta_i$ (β = the coefficient of x_i).

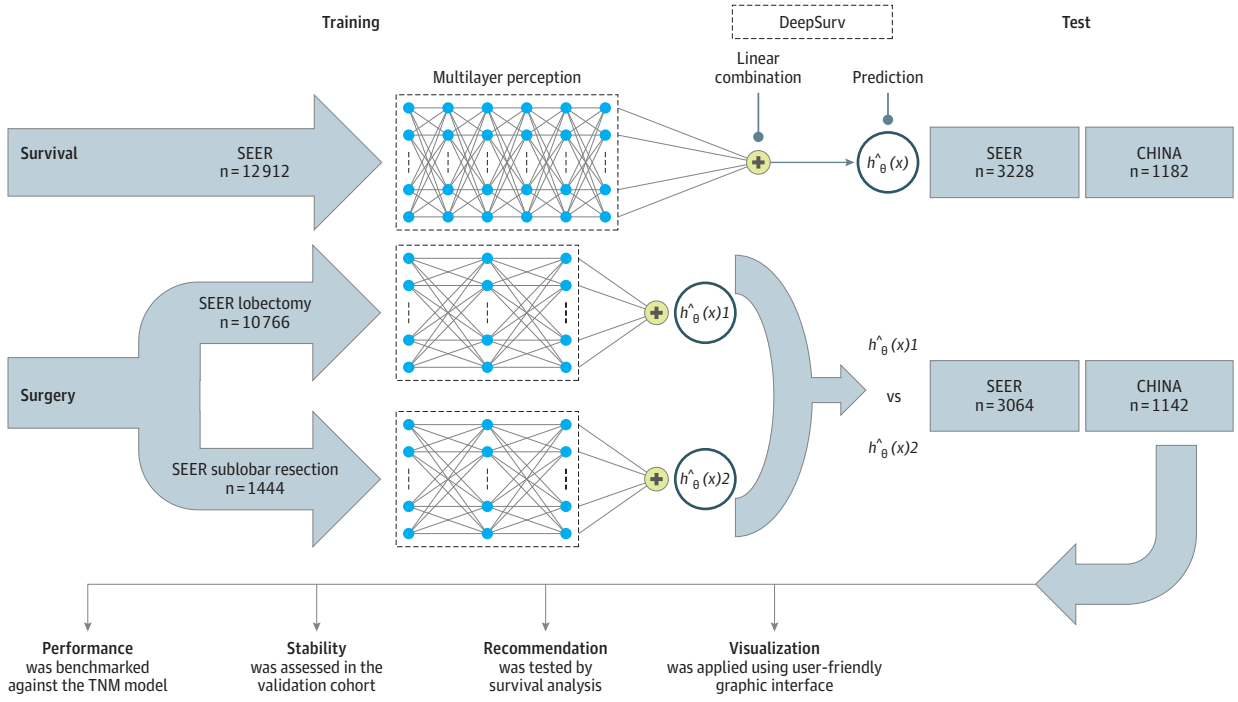
Model Visualization

We also developed a user-friendly interface to facilitate the survival predictions and treatment recommendations of the DeepSurv model. This interface consists of 3 views: (1) the user input view, (2) the survival prediction view, and (3) the treatment recommendation view. The user input view is designed to help users input all entries regarding patient characteristics using the XML schema constructed based on the features input into DeepSurv models. The user input view allows users to predict the survival probability and obtain a treatment recommendation based on specific patient information by clicking the predict and recommendation buttons, respectively. All SEER codes followed the SEER guideline.[21]

Statistical Analysis

A 2-sided *P* value less than .05 was considered to be statistically significant. The Akaike information criterion was calculated to assess the risk of overfitting. The likelihood-based method was applied to the type I censoring design.[22] All statistical analyses were performed with SPSS version 23 (IBM Corporation) software. The C statistic was performed by comparing C package with R statistical software (R Project for Statistical Computing), and the survival curves were plotted using GraphPad Prism 7 (GraphPad Software) software. Codes in our study are available online (https://github.com/thoraciclang/Deep_Lung).[23] Analysis began January 2018 and ended June 2019.

Figure 1. Diagram of the Study Procedure



Deep learning networks were trained end to end on 3 data set groups. The training and testing of these networks were all conducted on independent data sets. Four further experiments were conducted on the networks to test their performances against tumor, node, and metastasis (TNM) models, assess their degrees of stability, formulate recommendations, and finally, accomplish model visualization. SEER indicates Surveillance, Epidemiology, and End Results.

Results

Screening Process and Clinicopathology

A total of 17 322 patients with NSCLC were included in the study. According to the screening criteria, a total of 16 140 patients diagnosed as having NSCLC from the SEER database were included (eFigure 1A in the Supplement). **Table 1** shows the patients' main baseline clinical characteristics (eTables 2 and 3 in the Supplement). The majority of patients were white (13 361 [82.8%]), and the median (interquartile range) age was 68 (61-74) years. The majority of tumors were stage I disease (9327 [57.8%]) and adenocarcinoma (11 037 [68.4%]). The median (interquartile range) follow-up time was 24 (10-43) months. There were 2893 patients (17.9%) who had events (deaths from NSCLC) during the follow-up time. There were 1182 patients diagnosed with NSCLC from CHINA database (eFigure 1B in the Supplement). There were 226 events (deaths from NSCLC) over a median (interquartile range) follow-up time of 63.3 (53-70) months.

Table 1. Main Characteristics of Patients in the Whole Data Sets of Survival Analysis

Characteristic[a]	Data set, No. (%)		
	Training	SEER (test 1)	CHINA (test 2)
Age at diagnosis, median (range), y	68 (28-95)	68 (19-92)	60 (30-87)
Sex			
Female	6657 (51.6)	1639 (50.8)	642 (54.3)
Male	6255 (48.4)	1589 (49.2)	540 (45.7)
Histologic type			
Adenocarcinoma	8794 (68.1)	2243 (69.5)	948 (80.2)
Squamous cell carcinoma	4118 (31.9)	985 (30.5)	234 (19.8)
Marital status at diagnosis			
Unmarried	5304 (41.1)	1843 (57.1)	526 (44.5)
Married	7608 (58.9)	1385 (42.9)	656 (55.5)
T			
T1a	563 (4.4)	139 (4.3)	128 (10.8)
T1b	3156 (24.4)	804 (24.9)	396 (33.5)
T1c	2342 (18.1)	641 (19.9)	346 (29.3)
T2a	3258 (25.2)	791 (24.5)	208 (17.6)
T2b	594 (4.6)	141 (4.4)	56 (4.7)
T3	1994 (15.4)	445 (13.8)	40 (3.4)
T4	1005 (7.8)	267 (8.3)	8 (0.7)
N			
N0	9712 (75.2)	2439 (75.6)	1030 (87.1)
N1	1732 (13.4)	418 (12.9)	54 (4.6)
N2	1422 (11)	356 (11)	98 (8.3)
N3	46 (0.4)	15 (0.5)	0
M			
M0	12 559 (97.3)	3132 (97)	1182 (100)
M1a	143 (1.1)	41 (1.3)	0
M1b	202 (1.6)	52 (1.6)	0
M1c	8 (0.1)	3 (0.1)	0
LCCS			
Alive	10 581 (81.9)	2666 (82.6)	956 (80.9)
Dead	2331 (18.1)	562 (17.4)	226 (19.1)
Surgery to primary site			
Pneumonectomy	613 (4.7)	132 (4.1)	40 (3.4)
Lobectomy	10 766 (83.4)	2695 (83.5)	872 (73.8)
Sublobar	1444 (11.2)	369 (11.4)	270 (22.8)
None	89 (0.7)	32 (1.0)	0

Abbreviations: LCCS, lung cancer–specific survival; M, metastasis; N, node; SEER, Surveillance, Epidemiology, and End Results cancer registry; T, tumor.

[a] Other detailed clinical characteristics can be found in eTables 2 and 3 in the Supplement.

Training Curves

eFigure 2 in the Supplement demonstrated the training curves of networks in 3 submodels. The accuracy during the training course was indicated by validation and training lines. The curve was plotted to monitor the training course as the weights of the network were adjusted over each epoch, which represented the algorithm runs through the entire training and test data sets. After fine tuning, the change trend of loss and accuracy tended to become smoother and the algorithm maintained high accuracy on the validation set without significant overfitting. With a 500-epoch limit, we chose the model with the best performance on the test data set.

Calibration and Validation of the Prognostic DeepSurv for LCSS in the Test Set

We compared the TNM staging model to DeepSurv for LCSS in the test data sets (**Table 2**). The calibration plot indicated the calibration and how far the predictions of DeepSurv deviated from the actual event (**Figure 2**). In general, the actual outcomes in our databases of all patients with NSCLC for 3-year and 5-year LCSS were highly consistent with those predicted by the DeepSurv model, with most points falling almost directly on the 45° line. The DeepSurv model generated significantly better predictions than the TNM staging model (C statistic for TNM stage vs DeepSurv = 0.70; 95% CI, 0.681-0.731 vs 0.739; 95% CI, 0.713-0.764 [$P < .001$]). In the test group (CHINA data set), the DeepSurv model (C statistic = 0.742; 95% CI, 0.709-0.775) showed significantly better prediction than TNM model (C statistic = 0.706; 95% CI, 0.681-0.731; $P < .001$). High C statistic was observed for the results of the lobectomy and sublobar resection test data sets (Table 2). The feature component weightings in DeepSurv model are listed at eTable 4 in the Supplement.

The Cox proportional hazard regression model (eTable 5 in the Supplement) was compared with the DeepSurv model for LCSS. The DeepSurv model had significantly better predictions compared with the Cox proportional hazard regression model (C statistic for Cox proportional hazard regression model vs DeepSurv model = 0.716; 95% CI, 0.705-0.727 vs 0.739; 95% CI, 0.713-0.764). The Akaike information criterion value of TNM stage model, Cox proportional hazard regression model, and DeepSurv model were 10741.89, 10307.08, and 10310.52, respectively.

Treatment Recommender

First, we plotted 2 Kaplan-Meier survival curves: the outcome of test cases whose actual treatments were the same as those recommended and those whose were not (eFigure 3 in the Supplement). The population that followed the recommendation experienced significantly better survival rates than those who did not (SEER: hazard ratio [HR], 2.99; 95% CI, 2.49-3.59; $P < .001$ vs CHINA: HR, 2.14; 95% CI, 1.65-2.77; $P < .001$). Furthermore, patients in the test data sets were classified into lobectomy and sublobar resection groups according to the received treatment. Consistent with prior analyses, LCSS favored lobectomy compared with sublobar resection in the subgroup with the lobectomy recommendation (SEER: HR, 1.79; 95% CI, 1.28-2.50; $P = .001$ vs CHINA: HR, 1.92; 95% CI, 1.30-2.83; $P = .001$). No significant distinction in survival results were observed for lobectomy and sublobar resection in the subgroup with the sublobar resection recommendation (SEER: HR, 0.65; 95% CI, 0.41-1.02; $P = .06$ vs CHINA: HR, 0.75; 95% CI, 0.44-1.77; $P = .28$).

Table 2. Comparison of TNM Stage and DeepSurv Model for Survival Prediction in Test Data Sets

LCCS outcome	Model	C statistic (95% CI)	*P* value
SEER	TNM	0.706 (0.681-0.731)	NA
	DeepSurv	0.739 (0.713-0.764)	<.001
CHINA	TNM	0.691 (0.659-0.724)	NA
	DeepSurv	0.742 (0.709-0.775)	<.001
Treatment	DeepSurv (lobectomy, SEER)	0.725 (0.698-0.751)	NA
	DeepSurv (sublobar resection, SEER)	0.698 (0.672-0.725)	NA

Abbreviations: LCCS, lung cancer-specific survival; NA, not applicable; SEER, Surveillance, Epidemiology, and End Results cancer registry; TNM, tumor, node, and metastasis.

Model Visualization

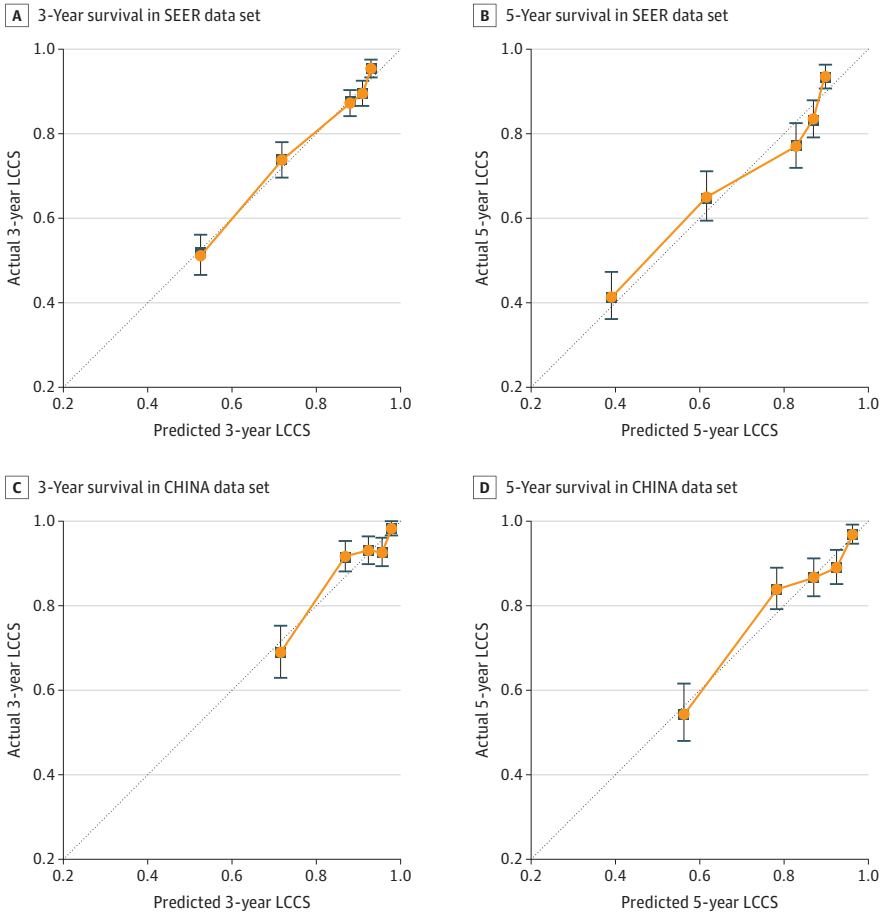
In the prediction view, the system invokes a prediction model (**Figure 3**; **Video**), and the DeepSurv model is used to predict patients' survival probability. The analysis results are visualized in a graphic view as a survival curve, which indicates the survival probability of the patient input over time. In the recommendation view, a recommendation model is adopted by the system, which can provide different patient survival probabilities for different treatments (lobectomy or sublobar resection) (Figure 3; **Video**). The survival curves of lobectomy and sublobar resection are also presented in a graphic view to facilitate visual comparison.

Discussion

The results of our pilot study proved that the deep learning network model (DeepSurv) performed better than conventional linear regression modeling (TNM staging model) in postoperative outcome prediction for patients with newly diagnosed NSCLC. Also, this model may serve as a useful analytical tool for treatment recommendation in patients with NSCLC, given its evidence of the significant prognostic benefits of following the treatment recommendation, which clearly outweigh those associated with not following the recommendation.

Previous studies have reported a series of linear models to predict the survival of patients with lung cancer.[24-27] However, few risk factors have been selected to these models, which is significantly associated with the survival or recurrence. For example, Liang et al[25] constructed a nomogram based on 6 factors. On the other hand, our Cox analysis demonstrated the contribution of 16 factors in the

Figure 2. Calibration Plots for Lung Cancer–Specific Survival (LCCS) for the DeepSurv Model



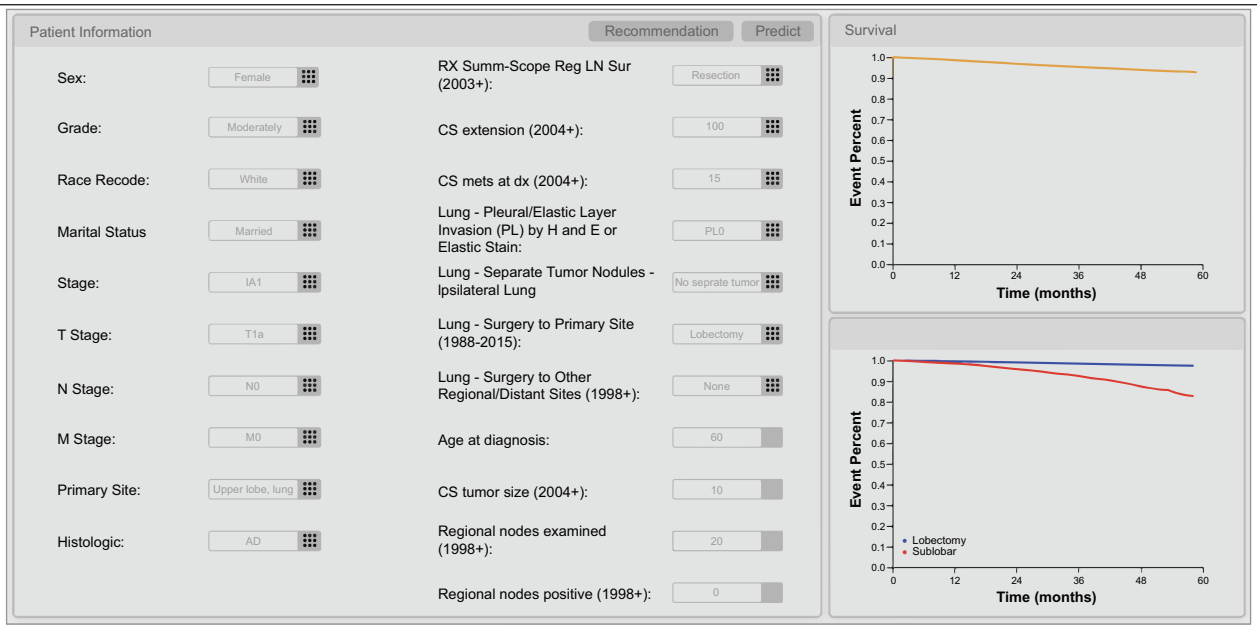
Three-year survival (A) and 5-year survival (B) of Surveillance, Epidemiology, and End Results (SEER) data set and 3-year survival (C) and 5-year survival (D) of CHINA data set are shown.

DeepSurv model. Obviously, a more comprehensive analysis could be performed by a nonlinear deep learning model. After reviewing the most relevant advanced research, we found many studies to have already applied deep learning models in their analytical approaches to surgical oncology research.[13] However, most studies have focused on diagnostic application,[28-30] such as radiographic image automated quantification,[14,31-35] digital histopathology image interpretation,[30,36-39] or biomarker analysis.[11,40] Examples of published research using deep learning models for prognostic prediction in surgical oncology are rare, to our knowledge. In NSCLC research, the deep learning technique has been applied to digital histopathology image interpretation, driver mutation risk detection,[40] and image characteristics discrimination,[33,41] but only a few studies have focused on postoperative outcome prediction or surgical recommendation, to our knowledge.

As a new analytic tool, the deep learning network model will likely become more widely applied to support clinical decision-making.[13] The performance of deep learning models in improving treatment outcomes is a key question and requires solid validation in the real world. In an analysis of 1194 patients with NSCLC, Hosny et al[14] evaluated the prognostic signatures of quantitative imaging features, which were extracted using deep learning networks. Based on their study of the TNM stages of postoperative patients, the authors' main finding was that deep learning networks significantly outperformed previous models. In our study, we selected a larger patient cohort with NSCLC of unselected consecutive cases including I to IV stages for model training and testing, which provide more solid results for interpretation. The advantages of the deep learning network model for postoperative outcome prediction in surgical research can be summarized as follows.[42,43] First, DeepSurv shows improved adaptability to variables with a nonlinear association, which includes real-world clinical factors. Unlike other models, deep learning algorithms can integrate the nonlinear risk functions associated with outcomes. Second, DeepSurv possesses flexibility in dealing with complex clinical factors. DeepSurv models cannot only automatically learn feature representations from uninterpreted clinical data but also analyze censored factors. Also, the predictions of the DeepSurv model have been proven to perform better in big data analysis. Owing to its ability to learn factor representation, the advantages of the DeepSurv model in dealing with both large factors and sample size may play a big role in biomedical marker analyses.[44-46]

It is a surgeon's duty to introduce clinical information to patients. To facilitate discussion of different potential surgical options, surgeons and patients need an informative tool that focuses on

Figure 3. User-Friendly Interface of DeepSurv Model, Which Facilitates Survival Prediction and Treatment Recommendation



survival benefits. In real cases, the establishment of a user-friendly graphic interface based on a patient communication framework will be key to effectively conveying results and illustrating complex analyses, including prognostic prediction, treatment recommendation to patients and family members, and improving the surgeons' understanding of deep learning models.[44,47] With its fast application and convenient operation, the user-friendly graphic interface example established in our study (Figure 3; **Video**) shows potential for use with any type of surgical care. To date, identifying patients who are appropriate for initial surgical management and conveying individualized prognostic analyses of postoperative outcomes has been an elusive goal. Instead, most published models are guided by patient characteristics to generate prognostic factors and are influenced by biases for different treatments.[48] The DeepSurv model and its user-friendly graphic interface has the potential to address this clinical dilemma and better share individual outcomes following different surgical procedures.

Limitations

Since the innovation of deep learning models, many limitations have been recognized. First, deep learning network models are computationally expensive to train and validate. The process of predictions can be hard to interpret because the deep learning networks function much like black boxes, which make it difficult to determine how the network arrives at its decisions. We also recognize that single-clinical data sources have limited clinical characteristics compared with the automated quantification of radiographic images. In this study, we examined 127 features of 21 characteristics in the model. Some important factors including preoperative elements were neglected, which makes the recommendation system need more improvements and stay at feasibility trial status. Also, external validation is lacking in this study. Further study is needed to validate the advantages of deep learning networks in survival prediction.

Conclusions

To our knowledge, this study is the first to explore the performance of a deep learning network that integrates Cox proportional hazards (DeepSurv) in NSCLC and to obtain promising results in outcome prediction. In addition, we demonstrated DeepSurv's potential to provide personalized treatment recommendations based on real clinical data.

ARTICLE INFORMATION

Accepted for Publication: March 20, 2020.

Published: June 3, 2020. doi:10.1001/jamanetworkopen.2020.5842

Corresponding Author: Chang Chen, MD, PhD (changchenc@tongji.edu.cn), and Yijiu Ren, MD (yjscott@hotmail.com), Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai 200443, China.

Author Affiliations: Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China (She, Wu, Deng, Zhang, Su, Jiang, Xie, Ren, Chen); College of Design and Innovation, Tongji University, Shanghai, China (Jin, Cao); Shanghai Key Laboratory of Tuberculosis, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China (Liu); Computer Science, NYU Shanghai, Shanghai, China (Cao).

Author Contributions: Dr Chen had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Drs She, Wu and Jin equally contributed to this work.

Concept and design: Deng, Liu, Ren, Chen.

Acquisition, analysis, or interpretation of data: She, Jin, Wu, Deng, Zhang, Su, Jiang, Xie, Cao, Ren, Chen.

Drafting of the manuscript: Wu, Zhang, Jiang, Ren, Chen.

Critical revision of the manuscript for important intellectual content: She, Jin, Deng, Su, Liu, Xie, Cao, Ren, Chen.

Statistical analysis: She, Jin, Wu, Deng, Zhang, Xie, Ren, Chen.

Obtained funding: She, Xie, Ren, Chen.

Administrative, technical, or material support: She, Jin, Su, Jiang, Xie, Ren, Chen.

Supervision: Xie, Cao, Ren, Chen.

Conflict of Interest Disclosures: Dr Chen reported grants from Shanghai Hospital Development Center during the conduct of the study. No other disclosures were reported.

Funding/Support: This work was supported by projects from Shanghai Hospital Development Center (SHDC12012716), Shanghai Municipal Health Commission (2018ZHYL0102), Tongji University AI Program (12712150026), and Shanghai Pulmonary Hospital Innovation Program (FKCX1906).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. *CA Cancer J Clin*. 2016;66(2):115-132. doi:10.3322/caac.21338

2. Asamura H, Chansky K, Crowley J, et al. The International Association for the Study of Lung Cancer Staging Project: proposals for the revision of the n descriptors in the forthcoming 8th edition of the TNM classification for lung cancer. J Thorac Oncol. 2015;10(12):1675-1684. doi:10.1097/JTO.0000000000000678

3. Chansky K, Sculier JP, Crowley JJ, Giroux D, Van Meerbeeck J, Goldstraw P; International Staging Committee and Participating Institutions. The International Association for the Study of Lung Cancer Staging Project: prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. *J Thorac Oncol*. 2009;4(7):792-801. doi:10.1097/JTO.0b013e3181a7716e

4. Kawaguchi T, Takada M, Kubo A, et al. Performance status and smoking status are independent favorable prognostic factors for survival in non-small cell lung cancer: a comprehensive analysis of 26,957 patients with NSCLC. *J Thorac Oncol*. 2010;5(5):620-630. doi:10.1097/JTO.0b013e3181d2dcd9

5. Sculier JP, Chansky K, Crowley JJ, Van Meerbeeck J, Goldstraw P; International Staging Committee and Participating Institutions. The impact of additional prognostic factors on survival and their relationship with the anatomical extent of disease expressed by the 6th Edition of the TNM Classification of Malignant Tumors and the proposals for the 7th Edition. J Thorac Oncol. 2008;3(5):457-466. doi:10.1097/JTO.0b013e31816de2b8

6. Liang W, Zhang L, Jiang G, et al. Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. *J Clin Oncol*. 2015;33(8):861-869. doi:10.1200/JCO.2014.56.6661

7. Xie D, Marks R, Zhang M, et al. Nomograms predict overall survival for patients with small-cell lung cancer incorporating pretreatment peripheral blood markers. *J Thorac Oncol*. 2015;10(8):1213-1220. doi:10.1097/JTO.0000000000000585

8. Gold JS, Gönen M, Gutiérrez A, et al. Development and validation of a prognostic nomogram for recurrence-free survival after complete surgical resection of localised primary gastrointestinal stromal tumour: a retrospective analysis. *Lancet Oncol*. 2009;10(11):1045-1052. doi:10.1016/S1470-2045(09)70242-6

9. Callegaro D, Miceli R, Bonvalot S, et al. Development and external validation of two nomograms to predict overall survival and occurrence of distant metastases in adults after surgical resection of localised soft-tissue sarcomas of the extremities: a retrospective analysis. *Lancet Oncol*. 2016;17(5):671-680. doi:10.1016/S1470-2045(16)00010-3

10. Randall RL, Cable MG. Nominal nomograms and marginal margins: what is the law of the line? *Lancet Oncol*. 2016;17(5):554-556. doi:10.1016/S1470-2045(16)00072-3

11. Li B, Cui Y, Diehn M, Li R. Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non-small cell lung cancer. *JAMA Oncol*. 2017;3(11):1529-1537. doi:10.1001/jamaoncol.2017.1609

12. Kopecky KE, Urbach D, Schwarze ML. Risk calculators and decision aids are not enough for shared decision making. *JAMA Surg*. 2019;154(1):3-4.

13. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268(1):70-76. doi:10.1097/SLA.0000000000002693

14. 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

15. Bergquist SL, Brooks GA, Keating NL, Landrum MB, Rose S. Classifying lung cancer severity with ensemble machine learning in health care claims data. *Proc Mach Learn Res*. 2017;68:25-38.

16. Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. *PLoS Med*. 2018;15(11):e1002701. doi:10.1371/journal.pmed.1002701

17. Matsuo K, Purushotham S, Jiang B, et al. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am J Obstet Gynecol*. 2019;220(4):381.e1-381.e14. doi:10.1016/j.ajog.2018.12.030

18. 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 Med Res Methodol*. 2018;18(1):24. doi:10.1186/s12874-018-0482-1

19. Shi S, Hua J, Liang C, et al. Proposed modification of the 8th edition of the AJCC staging system for pancreatic ductal adenocarcinoma. Ann Surg. 2019;269(5):944-950. doi:10.1097/SLA.0000000000002668

20. Wong SM, Freedman RA, Sagara Y, Aydogan F, Barry WT, Golshan M. Growing use of contralateral prophylactic mastectomy despite no improvement in long-term survival for invasive breast cancer. *Ann Surg*. 2017;265(3):581-589. doi:10.1097/SLA.0000000000001698

21. Dai C, Ren Y, Xie D, et al. Does lymph node metastasis have a negative prognostic impact in patients with NSCLC and M1a disease? *J Thorac Oncol*. 2016;11(10):1745-1754. doi:10.1016/j.jtho.2016.06.030

22. Leung K-M, Elashoff RM, Afifi AA. Censoring issues in survival analysis. *Annu Rev Public Health*. 1997;18(1):83-104. doi:10.1146/annurev.publhealth.18.1.83

23. GitHub. thoraciclang/Deep_Lung. Accessed May 11, 2020. https://github.com/thoraciclang/Deep_Lung

24. Wang Y, Qu X, Kam N-W, et al. An inflammation-related nomogram for predicting the survival of patients with non-small cell lung cancer after pulmonary lobectomy. *BMC Cancer*. 2018;18(1):692-692. doi:10.1186/s12885-018-4513-4

25. Liang W, Zhang L, Jiang G, et al. Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. *J Clin Oncol*. 2015;33(8):861-869. doi:10.1200/JCO.2014.56.6661

26. Yap W-K, Shih M-C, Kuo C, et al. Development and validation of a nomogram for assessing survival in patients with metastatic lung cancer referred for radiotherapy for bone metastases. *JAMA Netw Open*. 2018;1(6):e183242-e183242. doi:10.1001/jamanetworkopen.2018.3242

27. Zhang Y, Zheng D, Xie J, et al. Development and validation of web-based nomograms to precisely predict conditional risk of site-specific recurrence for patients with completely resected non-small cell lung cancer: a multiinstitutional study. *Chest*. 2018;154(3):501-511. doi:10.1016/j.chest.2018.04.040

28. Carin L, Pencina MJ. On deep learning for medical image analysis. *JAMA*. 2018;320(11):1192-1193. doi:10.1001/jama.2018.13316

29. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. 2019;25(3):433-438. doi:10.1038/s41591-018-0335-9

30. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054-1056. doi:10.1038/s41591-019-0462-y

31. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350. doi:10.1038/s41591-018-0107-6

32. Peake MD, Navani N, Baldwin DR. The continuum of screening and early detection, awareness and faster diagnosis of lung cancer. *Thorax*. 2018;73(12):1097-1098. doi:10.1136/thoraxjnl-2018-212189

33. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*. 2019;290(1):218-228. doi:10.1148/radiol.2018180237

34. Hwang EJ, Park S, Jin KN, et al; DLAD Development and Evaluation Group. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open*. 2019;2(3):e191095. doi:10.1001/jamanetworkopen.2019.1095

35. Topalovic M, Das N, Burgel PR, et al; Pulmonary Function Study Investigators. Pulmonary Function Study Investigators. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur Respir J*. 2019;53(4):1801660. doi:10.1183/13993003.01660-2018

36. Ren J, Karagoz K, Gatza ML, et al. Recurrence analysis on prostate cancer patients with Gleason score 7 using integrated histopathology whole-slide images and genomic data through deep neural networks. *J Med Imaging (Bellingham)*. 2018;5(4):047501. doi:10.1117/1.JMI.5.4.047501

37. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301-1309. doi:10.1038/s41591-019-0508-1

38. Gertych A, Swiderska-Chadaj Z, Ma Z, et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci Rep*. 2019;9(1):1483. doi:10.1038/s41598-018-37638-9

39. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol*. 2018;42(12):1636-1646. doi:10.1097/PAS.0000000000001151

40. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods*. 2018;15(10):816-822. doi:10.1038/s41592-018-0138-4

41. Tanaka N, Kanatani S, Tomer R, et al. Whole-tissue biopsy phenotyping of three-dimensional tumours reveals patterns of cancer heterogeneity. *Nat Biomed Eng*. 2017;1(10):796-806. doi:10.1038/s41551-017-0139-0

42. Yasaka K, Abe O. Deep learning and artificial intelligence in radiology: current applications and future directions. *PLoS Med*. 2018;15(11):e1002707. doi:10.1371/journal.pmed.1002707

43. Smith CC, Chai S, Washington AR, et al. Machine-learning prediction of tumor antigen immunogenicity in the selection of therapeutic epitopes. *Cancer Immunol Res*. 2019;7(10):1591-1604. doi:10.1158/2326-6066.CIR-19-0155

44. Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. *JAMA Intern Med*. 2019;179(3):293-294.

45. Shaffie A, Soliman A, Fraiwan L, et al. A generalized deep learning-based diagnostic system for early diagnosis of various types of pulmonary nodules. *Technol Cancer Res Treat*. 2018;17:1533033818798800. doi:10.1177/1533033818798800

46. Kantz ED, Tiwari S, Watrous JD, Cheng S, Jain M. Deep neural networks for classification of LC-MS spectral peaks. *Anal Chem*. 2019;91(19):12407-12413. doi:10.1021/acs.analchem.9b02983

47. Simon G, DiNardo CD, Takahashi K, et al. Applying Artificial Intelligence to Address the Knowledge Gaps in Cancer Care. *Oncologist*. 2019;24(6):772-782.

48. Coiera E. On algorithms, machines, and medicine. *Lancet Oncol*. 2019;20(2):166-167. doi:10.1016/S1470-2045(18)30835-0

SUPPLEMENT.

eMethods.

eTable 1. All Clinical Features Integrated in the Model

eTable 2. Characteristics of Patients in the Training Dataset of Survival Analysis

eTable 3. Characteristics of Patients in the Test Dataset of Survival Analysis

eTable 4. Feature Component Weightings in the DeepSurv Model

eTable 5. Survival Predictors in Cox PH Model

eFigure 1. Flow chart of datasets construction. (A) SEER dataset, (B) CHINA dataset

eFigure 2. Training curves of networks in the survival dataset of SEER database (A), lobectomy dataset (B), and sublobar resection dataset (C)

eFigure 3. Lung Cancer–Specific Survival Recommendation Comparisons of SEER Data set (A), SEER Lobectomy Test Data set (B), and SEER Sublobar Resection Test Data set (C); Lung Cancer–Specific Survival Recommendation Comparisons of CHINA Data set (D), CHINA Lobectomy Test Data set (E), and CHINA sublobar Resection Test Data set (F)