Using Electronic Medical Records to Study Lung Cancer Prognosis

USING ELECTRONIC MEDICAL RECORDS TO STUDY LUNG CANCER PROGNOSIS

QIANYU YUAN

Population Health Sciences Program
Department of Environmental Health
Harvard T.H. Chan School of Public Health

April 2020

Dissertation advisor: David C. Christiani Qianyu Yuan

Using electronic medical records to study lung cancer prognosis

Abstract

Lung cancer is the most commonly diagnosed malignancy and is a leading cause of cancer-related deaths worldwide. In the US, the current five-year survival is about 20.6%, which is significantly lower than most leading cancers, such as prostate cancer (99%), breast cancer (91%), and colon cancer (66%). Survival of lung cancer patients is heterogeneous, even within the stage group. The identification of stable and reliable prognostic variables and the development of prediction tools are needed to identify the subgroup with better or worse outcomes. Electronic medical records (EMRs) provide a low-cost means of accessing rich longitudinal data on large populations for research. It allows us to evaluate multiple risk factors including clinical, demographic, treatment, molecular, behavior information, and lung cancer progression simultaneously, enabling development of predictive models.

In chapter 1, we assembled a lung cancer cohort using EMRs from a large healthcare system (Partners HealthCare). Phenotyping algorithm was applied to identify lung cancer patients. Extraction strategies combining structured and unstructured data were used to collect demographics, clinical outcomes, prognostic factors, and treatment information for lung cancer patients. Data completeness was evaluated, and data accuracy was assessed by comparing with the Boston Lung Cancer Study (BLCS) database and chart review results.

In chapter 2, a prognostic model for 5-year overall survival (OS) was developed and validated for newly diagnosed non-small cell patients. We identified age, sex, smoking status, histological type, stage, BMI, albumin, ALP, creatinine, HGB, RDW, WBC, NLR, calcium and sodium as significant predictors of 5-year OS. Our model achieved higher discrimination compared with the model based on sex, age, stage, and histological type. A more accurate outcome prediction model, which can be applied upon the diagnosis of NSCLC, would be essential for informed decisions making regarding clinical care and practice.

Finally, in chapter 3, we aimed to identify advanced NSCLC patients who likely benefit from PD-1/L1 inhibitors. We proposed a prognostic score to stratify advanced NSCLC patients treated with PD-1/L1 inhibitors into poor, intermediate, and good groups for progression free survival.

Added up, we assembled a large lung cancer cohort, investigated how clinical factors influence the prognosis of non-small cell lung cancer, and develop integrative prediction algorithms for clinical outcomes.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank many people for their support in completing this dissertation.

First, I would like to express my deepest gratitude to my advisor, Dr. David Christiani, for his continuous support of my doctoral study and research, for his endless encouragement along the way. I appreciate the trust he has given me to explore my research projects. Without his persistent help, this dissertation would not have been possible. I greatly appreciate my committee members. I thank Dr. Tianxi Cai, for introducing me to the field of EMR, for her insightful, professional guidance and immense knowledge. I thank Dr. Liming Liang, for being my master's advisor and providing me with outstanding training in molecular epidemiology and statistical modeling, for his patience and motivation throughout this journey.

I thank all my colleagues at Dr. David C. Christiani's lab for their suggestions towards this dissertation, for providing emotional support and comfort, for all the beautiful moments we spent together. I would thank my collaborators in Dr. Tianxi Cai's lab, for their insightful inputs and work in this dissertation. It has been great fun to work with such talented people.

Finally, from the bottom of my heart, I would like to thank my parents, Xiaoping Yuan and Xiaojun Tang, for their unwavering love and unconditional support on my studies and life.

.

Using electronic medical records to assemble a cohort for studying lung cancer prognosis

Abstract

Background: Electronic medical records (EMRs) provide a low-cost means of accessing longitudinal data on large populations with detailed information regarding diagnosis, clinical procedures, medications, and laboratory tests. A lung cancer cohort assembled from EMR represents a powerful resource for studying prognosis.

Method: A classification algorithm was developed to identify lung cancer patients from a large healthcare system (Partners HealthCare) between 1988 to October 2018. Data were extracted from both structured data and unstructured clinical notes processed by natural language processing (NLP) tools. We developed a new tool, NLP Interpreter for Cancer Extraction (NICE), to extract stage, histological type, and tumor mutations. Data completeness was evaluated, and data accuracy was assessed by comparing with the Boston Lung Cancer Study (BLCS) database and chart review results.

Results: The initial population contains 76,643 patients with at least one diagnostic code related to lung cancer. A total of 42,069 lung cancer patients were identified as lung cancer cases through classification model. Excluding patients with lung cancer history and patients with less than 14 days of follow-up after diagnosis resulted in a final cohort of 35,375 patients with data on demographics, clinical outcomes, prognostic factors, and treatment information. Analysis for overall survival showed high consistency between the EMR and BLCS cohorts, as Cox regression models controlled for age, sex, race, smoking status, histological type, stages yielded similar estimates.

Conclusion: We assembled a large scale EMR-based lung cancer patient cohort with detailed longitudinal measurements of clinical factors over time. This study would help to better understand how clinical factors influence the progression of lung cancer.

Introduction

Globally, lung cancer has been the most common cancer diagnosed and the leading cause of death from cancer for several decades.[1] In the US, the current five-year survival is about 20.6%, and has only been improved slightly from 12.2% over the past four decades.[2] The prognosis of lung cancer is heterogeneous with various prognostics factors. [3-7] The identification of stable and reliable prognostic variables would help to identify the subgroup with a better or worse response and serve as the evidence for decision-making regarding specific therapeutic interventions.

Many studies have leveraged epidemiology cohorts for lung cancer prognosis research, such as Surveillance, Epidemiology, and End Results (SEER) and The International Lung Cancer Consortium (ILCCO).[8-11] They used cancer registries, questionnaires, or ask clinical staff to obtain high-quality data, which require substantial time and effort. The growing availability of electronic medical records (EMR) data offers a timely and low-cost alternative with potential of efficiently including considerable study populations.[12,13] In addition, some studies use EMR to provide complementary data from real-life treatment populations to support clinical trials.[14-17] Rich longitudinal data regarding diagnosis, clinical procedures, medications, and tests in EMR offers new opportunities for lung cancer research. On the other hand, when bringing rich sources for analysis, the large amount and diversity of EMR data also introduce further difficulty to perform mining for cancer-related data. These facts make natural language processing a requisite technology for data extraction.[18,19]

In this work, we firstly developed a classification algorithm that accurately identified lung cancer patients from a large healthcare system (Partners HealthCare). Variables of interest were extracted from both structured data and unstructured clinical notes. We developed an NLP tool named NLP interpreter for cancer extraction (NICE) to apply data mining for cancer-related characteristics including clinical stage, TNM stage, histology, cancer first reported date, mutation variables. In addition, we adopted our previously developed tool EXTEND for extraction of height, weight, human body mass index (BMI), and Eastern Cooperative Oncology Group (ECOG) status. Data completeness was evaluated, and data accuracy was assessed by comparing with the Boston Lung Cancer Study (BLCS) database and chart review results. Our primary goal is to build a lung cancer cohort that is reliable for prognosis study using EMRs and also serves as a general approach for assembling EMR cohort to study cancer progression.

Materials and Methods

Data source and study population

EMR data were from the Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH) using Partners HealthCare System Research Patient Data Registry (RPDR). The Institutional Review Board of Partners HealthCare (Protocol Number: 1999P004935/PHS) approved this study. Initial data mart contains 76,643 patients with at least one diagnostic code related to lung cancer (International Classification of Diseases-10 [ICD-10]: 162, 1620, 1622, 1623, 1624, 1625, 1628, 1629, 20921, 2312, ICD10:C33, ICD10:C34.00, ICD10:C34.10, ICD10:C34.2, ICD10:C34.30, ICD10:C34.80, ICD10:C34.90, ICD10:C7A.090, ICD10:D02.20, ICD10:Z85.118, V1011).

The Boston Lung Cancer Study (BLCS) is a cancer epidemiology cohort of lung cancer cases enrolled at MGH and the Dana-Farber Cancer Institute from 1992-present. 6,225 patients from MGH can be linked to EMR database and were used for comparison for our study. Demographics, smoking status, clinical characteristics as well as first-line treatments immediately after diagnosis were collected at baseline from questionnaire, pathology reports, and clinical notes. Follow up medical records review was conducted to gather the survival information.

Identification of lung cancer patients

Training labels. To develop the classification algorithm to identify lung cancer patients, a total of 200 individuals were randomly selected as the gold-standard set. Medical record reviews of 200 charts were performed by two reviewers Qianyu Yuan (200 charts) and Andrea Shafer (200 charts) separately.

Features. A list of candidate features including codified data and informative medical concepts, were created for the classification algorithm. 1) The total number of ICD codes for lung cancer were counted for each patient. 2) We followed the previously published Surrogate-Assisted Feature Extraction (SAFE) method to generate a list of candidate lung cancer concepts.[20]

Algorithm training and evaluation. We developed the classification algorithm for lung cancer disease status using LASSO penalized logistic regression, which further reduced the number of variables in the model and optimized external validity. The final features feed into the classification algorithm include the total number of ICD codes for lung cancer and the total number of mentions of medical concepts "lung carcinoma" and "malignant lung neoplasm" in clinical notes. Comparing against chart-review gold-standard labels, performance characteristics of the classification algorithm were reported using the area under the receiver operating characteristic curve (AUC), positive predictive value (PPV), sensitivity, specificity, and F-score (harmonic mean of PPV and sensitivity). We chose 90% specificity as the threshold for a binary classifier. Cross-validation with 70:30 splits averaged over 100 random partitions was used to correct for overfitting bias. The algorithm assigned each patient a probability of having lung cancer. Those with probabilities above a threshold that achieves 90% specificity were classified as having lung cancer. We used established phenotyping package *PheCAP* in R version 3.5.0 for algorithm development.[21]

Data extraction

The RPDR includes patient demographics, vital signs, laboratory test results, problem list entries, prescribed medications, billing codes, and clinical notes.[22] Based on research purposes, we extracted demographics, clinical outcomes, prognostic factors, and treatment information from both structured data and clinical notes using the NLP tools (Table 1.1). We use the published tool Extraction of EMR Numerical Data (EXTEND)[23] to extract ECOG and BMI information. For other variables including cancer stages, histology, and mutation information, we developed a new tool named NLP Interpreter for Cancer Extraction (NICE) to perform data extraction from clinical notes including discharge summary, progress notes. For variables that use both structured data and clinical notes or have multiple values, we used rule-based approaches or prediction models to gather the final value.

Table 1.1 Data sources and description

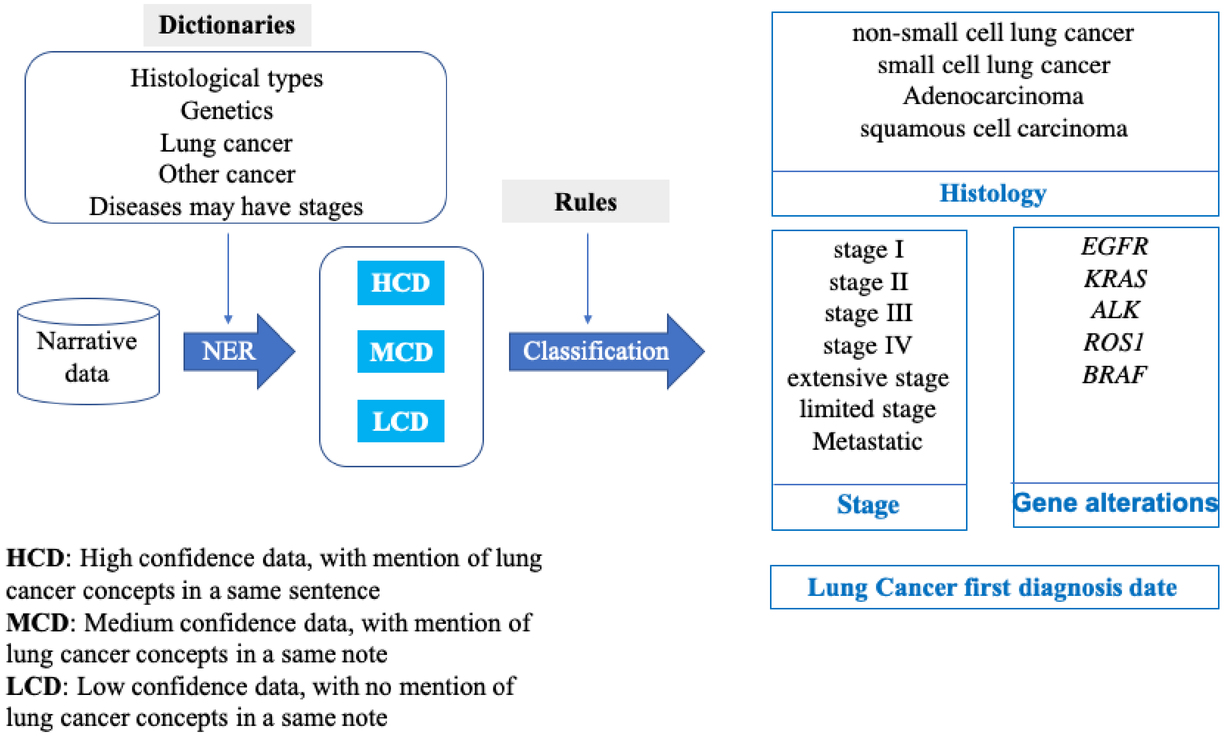
Variable	Data sources and extraction method		Variable description
	Structured data	Unstructured data	
Demographic			
Birthday	Demographics		
Sex	Demographics		
Race & Ethnicity	Demographics		
Clinical outcomes			
Diagnosis date	Diagnosis codes (ICD-9/10 codes)	NICE	Date of the lung cancer diagnosis
Overall survival	Visits or deaths		Death time/last visit time - diagnosis date
Prognostic factors			
Stage		NICE	TNM stage and clinical stage
Histology type		NICE	Non-small cell lung cancer (adenocarcinoma, squamous cell carcinoma, other non-small cell carcinoma), small cell lung cancer
Smoking status			Smoker and non-smoker
BMI	Vital signs	EXTEND	Weight in kilograms divided by the square of height in meters
ECOG		EXTEND	Grade 0 to 4
Laboratory test	Laboratory tests codes		Complete blood count, metabolic panel, prothrombin time, lipid panel, liver panel, thyroid stimulating hormone, hemoglobin A1C and urinalysis
Tumor mutation		NICE	Genetic alterations in *EGFR, KRAS, ALK, ROS1, MET,* and *BRAF*
Medical history	Diagnosis codes (ICD-9/10 codes)		Respiratory disease (e.g., COPD, asthma), cardiovascular disease and etc.
Treatment			
Surgery	Procedure codes (CPT/ICD-10 codes)		Surgery procedure (lobectomy, segmentectomy, wedge resection, wedge resection, video-assisted thoracic surgery (VATS))
Radiation therapy	Procedure codes (CPT/ICD-10 codes)		Radiation therapy procedure
Chemotherapy	Procedure codes (CPT/ICD-10 codes) and medication names		Chemotherapy procedures, lung cancer chemotherapy drugs
Target therapy/ immune therapy	Medication names		Lung cancer target therapy and immunotherapy drugs

Variables extracted from structured data

Birthday, sex, and race/ethnicity were stored in the structured coded data. Birthday was used for calculating age of diagnosis. Race was categorized into five categories: White, Asian, Black, Hispanic and others. Common treatments for lung cancer patients include surgery, chemotherapy, radiation therapy, target therapy, and immunotherapy can be found in prescribed medications, billing codes. Surgery and radiation therapy were extracted using ICD-9/10 and Common Procedural Terminology (CPT) codes. Chemotherapy, target therapy, and immunotherapy were extracted from ICD-9/10-CM, CPT, and medication codes. Common laboratory tests include complete blood count, metabolic panel, prothrombin time, lipid panel, liver panel, thyroid stimulating hormone, hemoglobin A1C and urinalysis. The numeric value with its measurement dates were extracted using structured codes.

Variables extracted from unstructured data

NLP Interpreter for Cancer Extraction (NICE) was developed to extract lung cancer concepts, cancer stages, histology, and mutation information (Figure1.1).



HCD: High confidence data, with mention of lung cancer concepts in a same sentence
MCD: Medium confidence data, with mention of lung cancer concepts in a same note
LCD: Low confidence data, with no mention of lung cancer concepts in a same note

Figure1.1 Workflow of NICE

For the extraction of lung cancer concepts, we built a dictionary containing all synonyms of the concept ‘lung cancer’ such as ‘lung cancer’ and ‘lung carcinoma’ using the Unified Medical Language System (UMLS). All notes were processed to identify the positive mention of the ‘lung cancer’ concept via Named Entity Recognition (NER). E.g. ‘Lung cancer’ in a sentence like ‘The patient denies lung cancer history’ was ignored. Date information was also extracted if a date was mentioned in the same sentence as the concept of ‘lung cancer’ was located. E.g., “Lung cancer (HCC) 11/18/2014.”. The date “11/18/2014” was assigned to the concept mention of ‘lung cancer’. The most mentioned dates of the ‘lung cancer’ concept was combined with ICD-9/10 codes time to choose the earlier time as the cancer diagnosis date.

For the extraction for stage and histology, we built dictionaries for both that are similar as for lung cancer concept. Then we processed notes to identify positive mention of stage and histology. Because stage information can also be mentioned for other diseases such as various other cancers, sleep status, bed sore, and chronic kidney disease, we ignored the mention of stage with mention any of these diseases in the same sentence. We also excluded the mention of histology if there was a mention of other cancer instead of lung cancer. The mention of stage and histology will be categorized into three confidence levels: high, medium, and low. A high confidence level was assigned when the lung cancer concept appeared in the same sentence. A medium confidence level was assigned when the lung cancer concept appeared in the same note instead of the same sentence. We assigned la ow confidence level to stage or histology concept if there was no mention of lung cancer concept in the same note. We built regular expression patterns for extracting TNM stages as additional stage information then convert TNM stages to clinical stages. For histological type, phrases were grouped into four categories: non-small cell lung cancer, small cell lung cancer, adenocarcinoma, squamous cell carcinoma. If at least two histological types were mentioned, the most commonly occurring phrases were selected. If the most commonly occurring type is non-small cell lung cancer, we choose the most occurring subtype: adenocarcinoma or squamous cell carcinoma. If none of these subtypes were mentioned, the histological type was defined as non-small cell unspecified. For stage, phrases were extracted and grouped into the seven categories: stage I, stage II, stage III, and stage IV, extensive stage, limited stage, and metastatic.

Gene alterations that listed in the NCCN guideline and have been identified that impact therapy selection including *EGFR, KRAS, ALK, ROS1,* and *BRAF*. For patients who received tumor diagnostic tests from Partners including Snapshot assay, Fluorescent in situ hybridization (FISH), Immunohistochemistry (IHC), we extracted results from the molecular pathology reports. The process of the extraction is similar to for stage and histology, but we don’t perform the categorization of confidence level because the mention of these genetic variables in pathology notes is specific without ambiguation.

For smoking status, each patient was assigned as a smoker or non-smokers and was predicted using classification model combining structured coded data and clinical notes. To calculate BMI, height and weight or calculated BMI recorded with measurement date were extracted using structured data. BMI and ECOG performance status documented in clinical notes were extracted using the NLP tool EXTEND.[23]

Patient selection criteria

This study excluded patients whose lung cancer history ICD codes (ICD10: Z85.118, V1011) were earlier than lung cancer ICD codes, under the assumption that they were recurrent or secondary primary lung cancer patients. Patients with follow up less than 14 days after diagnosis were also excluded.

Data quality assessment

To assess the utility of using EMR data in cancer research, the quality of the database should be evaluated, here we assessed the completeness and accuracy of the dataset.

Completeness

First, the percentage of completeness was calculated for each variable on patient-level to measure whether patients had at least one measurement for the variable. Second, two months' time window before and after diagnosis was cut to measure the data availability at the time of diagnosis. We further investigated the relationship between completeness and year of diagnosis as well as the visit days they were present in hospitals.

Accuracy

The accuracy of basic characteristics was assessed by comparison with two datasets. One is random samples from the data mart, which is manually reviewed. Chart review gathered data retrospectively from EMR system and can be viewed as a gold standard. The accuracy was tested when comparing chart review and EMR data. Another is the large sample size BLCS cohort, which prospectively collect data with its multiple data collection sources. Agreement would be tested by comparing data curated from BLCS and data from EMR.

Firstly, we assessed the distribution pattern across for basic characteristics. Secondly, we specifically tested the accuracy for diagnosis date, histological type, and clinical stages as these variables are extracted from clinical notes and are essential for prognosis study. For assessment of the diagnosis date, absolute discrepancies were calculated by the difference between EMR estimated dates and diagnosis dates from chart review or cohort data. Distributions of the date difference was shown using histograms. Percentage of absolute discrepancy of more than 90 days, 180 days and one year was calculated. Contingency tables comparing chart review/BLCS data and EMR data were compared to test the accuracy/agreement for histological and clinical stages. To assess whether EMR data yield valid estimates, Cox proportional models control for age, sex, race, smoking status, histological type, stage were conducted within the population of BLCS, using data curated from BLCS and EMR data separately. The hazard ratio and p values that test the effect of each variable on overall survival would be compared to test the consistency of results.

Results

Study population and extracted variables

Among the 200 reviewed charts, we identified 142 true lung cancer cases, 55 non lung cancer patients, and three uncertain patients. The best classification model identified 42,069 lung cancer patients with a sensitivity of 75.2%, specificity of 90.0%, PPV of 94.4, F score of 0.837, and AUC of 0.927. Excluding patients with lung cancer history (n=2,876), and patients with less than 14 days of follow-up after diagnosis (n=5,302) resulted in a final cohort of 35,375 patients. Over the study period, the number of patients diagnosed across calendar years increased (Figure 1.2).

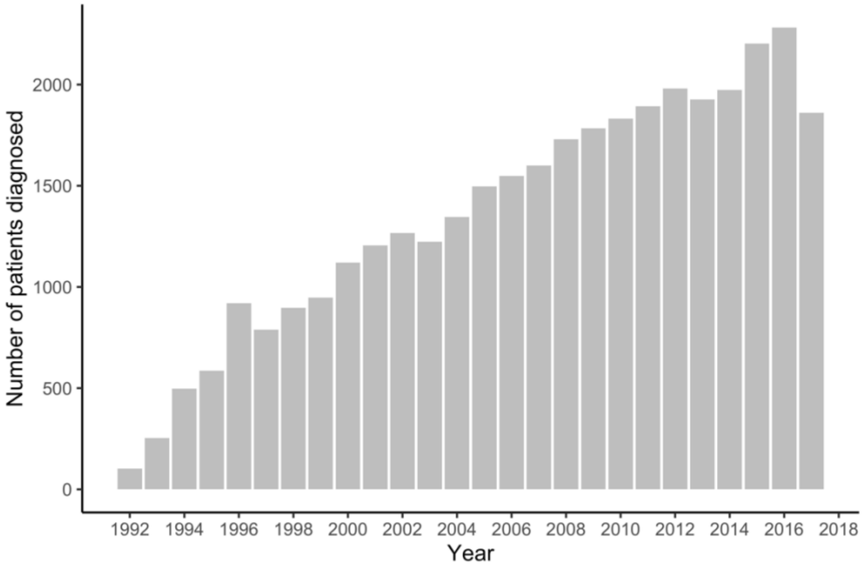


Figure 1.2 Number of lung cancer patients identified from Partners

A summary of demographic and baseline characteristics of the full cohort and final cohort was presented in Table 1.2. In the final cohort, the median age at diagnosis was 66.0 years; around half of the patients were female (53.0%, n=18,754), the majority of them were white (85.1%, n=30,097) and most had a history of smoking (92.3%, n=32,650). 89.8% of the patients were non-small cell patients (59.5% adenocarcinoma patients, 18.9% squamous cell, and 11.7% other non-small cell lung cancer) and 9.9% were small cell lung cancer patients. 42.5% of patients were diagnosed at early stage and 57.5% were diagnosed at late stage.

Table 1.2 Basic characteristics of the full cohort and final cohort

	Full cohort(n=42,069)	**Final cohort(n=35,375)**
Characteristics	**Number (%)**	**Number (%)**
Age at initial diagnosis	66.0±11.6	66.0±11.5
Gender		
Female	22,158 (52.7)	18,756 (53.0)
Male	19,898 (47.3)	16,613 (47.0)
Unknown	13 (0.0)	6 (0.0)
Ethnicity		
White	35155 (83.6)	30,140 (85.2)
Black	1210 (2.9)	1,040 (2.9)
Asian	1007 (2.4)	857 (2.4)
Hispanic	395 (0.9)	323 (0.9)
Other	321 (0.8)	267 (0.8)
Unknown	3981 (9.5)	2,748 (7.8)
Smoking Status		
Smoker	38,492 (91.5)	32,650 (92.3)
Non-Smoker	3,577 (8.5)	2,725 (7.7)
Histology		
Completeness (%)	82.2	87.1
Adenocarcinoma	20,256 (58.6)	18,331 (59.5)
Squamous cell	6,401 (18.5)	5,816 (18.9)
NSCLC unspecified	4,409 (12.7)	3,601 (11.7)
Small cell	3,535 (10.2)	3,065 (9.9)
Stage		
Completeness (%)	69.7	75.9
1	7,714 (26.3)	7,083 (26.4)
2	3,357 (11.5)	3,069 (11.4)
3	6,363 (21.7)	5,889 (21.9)
4	9,380 (32.0)	8,495 (31.6)
Limited	1,324 (4.5)	1,222 (4.6)
Extensive	1,177 (4.0)	1,085 (4.0)

For patients in the final cohort, 38.5%, 39.7%, and 41.6% of the patients received the surgery, chemotherapy, and radiation therapy within the Partners system with ICD 9/10 codes, procedure codes, or medication codes available. 29.5%, 20.4%, and 24.4% of the patients received the surgery, chemotherapy, and radiation therapy within three months after diagnosis (Table 1.3). Target therapies and immunotherapies are often used for advanced patients. In the final cohort, 396 patients received angiogenesis inhibitors, 1455 patients received EGFR inhibitor, 232 patients received ALK inhibitors, and 11 patients received BRAF inhibitor. In addition, 503 patients received PD-L1/PD-1 inhibitors.

Table 1.3 Percent of patients receiving treatments within Partners HealthCare

	Any treatment	**Primary treatment** [a]
Therapy type	**Number (%)**	**Number (%)**
Surgery	13,628(38.5)	10,446(29.5)
Chemotherapy	14,039(39.7)	7,204(20.4)
Radiotherapy	14,710(41.6)	8,627(24.4)
Target therapy	2,631(7.4)	667(1.9)
Immunotherapy	504(1.4)	94(0.3)

[a] Primary treatment: surgery received within one month before diagnosis or within three months after diagnosis, chemotherapy/radiation therapy/target therapy/immunotherapy received within three months after diagnosis

Genetic mutation results from different molecular tests were shown in Table 1.4. Among 4,655 patients tested using SNaPshot assay, 46.9% of patients were positive for at least one mutation in three genes, including 26.7% *KRAS*, 18.4% *EGFR*, and 3.7% *BRAF*. Translocation of *ALK*, *ROS1* were tested among 3,791, 2,436 patients with a positive rate of 5.4% and 2.1%.

Table 1.4 Patients tested for NCCN listed driven genes

Gene	Platform	# of patients tested	Mutation frequencies (%)
EGFR	SNaPshot	4,655	18.4
KRAS	SNaPshot	4,655	26.7
BRAF	SNaPshot	4,655	3.7
ALK	FISH/IHC assay	3,791	5.4
ROS	FISH/IHC assay	2,436	2.1

The estimated median overall survival (OS), defined as the time from the date of diagnosis to the date of death, the date of the latest follow-up, whichever came first. Median OS was 2.51 (95% CI: 2.45 - 2.57) years. The median OS for stage 1 to 4 in non-small cell lung cancer patients were 9.29 (95% CI: 8.98 - 2.57), 5.29 (95% CI: 4.88 - 5.61), 2.38 (95% CI: 2.26 - 2.48) and 1.31 (95% CI: 1.28 - 1.39) years, respectively.

Data completeness

In structured data, birthday, sex, and race were available for 100%, 99.97%, and 92.2% of the study population, respectively. In treatment data, patients without specific treatment procedure codes and medication codes can be truly absent of treatment or because they received treatment in other healthcare centers. 59.9%(n=21,189) of the patients have at least one lung cancer-related therapy within Partners. For common laboratory tests, 82.5% of patients have at least one measurement at any time, and 67.7% of patients have at least one measurement within 60 days before or after diagnosis date (Table 1.5).

Variables from unstructured data that need to be extracted from clinical notes frequently have more missing values (Table 1.5). 87.1% of patients have an extracted histological type, and 75.9% of patients have extracted stage. For longitudinal measurements, 49.6%, 28.1% of patients have at least one measurement of BMI, ECOG performance status measured during their stay in Partners. 38.9%, 14.8% of patients have at least one measurement of BMI, ECOG performance status measured within 60 days before and after diagnosis time.

Table 1.5 Completeness of variables in the final cohort

	Completeness (%)	
	Total [a]	**Baseline** [b]
Demographic	100	100
Clinical outcomes		
Diagnosis date	100	100
Overall survival	100	100
Prognostic factors		
Stage	75.9	75.9
Histology type	87.1	87.1
Smoking status	100	100
BMI	49.6	38.9
ECOG	28.1	14.8
Laboratory test [c]	82.5	65.3

a. Data available at any time
b. Data available within three months before or after diagnosis time
c. At least one measurement for common lab test

There was an increase in the completeness for selected variables based on the year of diagnosis (Figure 1.3). The increase follows the gradual process of EHR adoption within Partners. There was also an increase in the completeness as the number of days when they were present in the hospital increase (Figure 1.4).

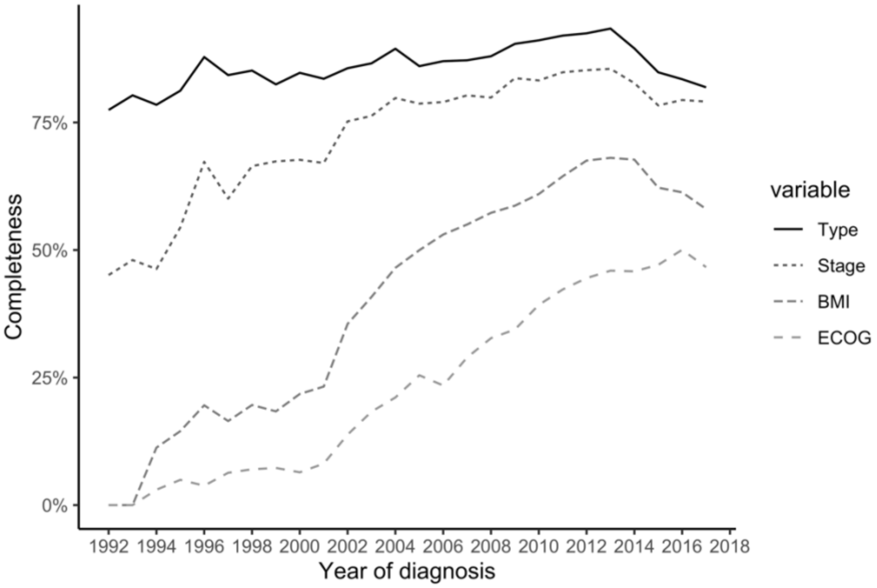


Figure 1.3 Completeness of type, stage, BMI, and ECOG performance status improvement over time. The completeness of variables has improved as EMR adoption has increased

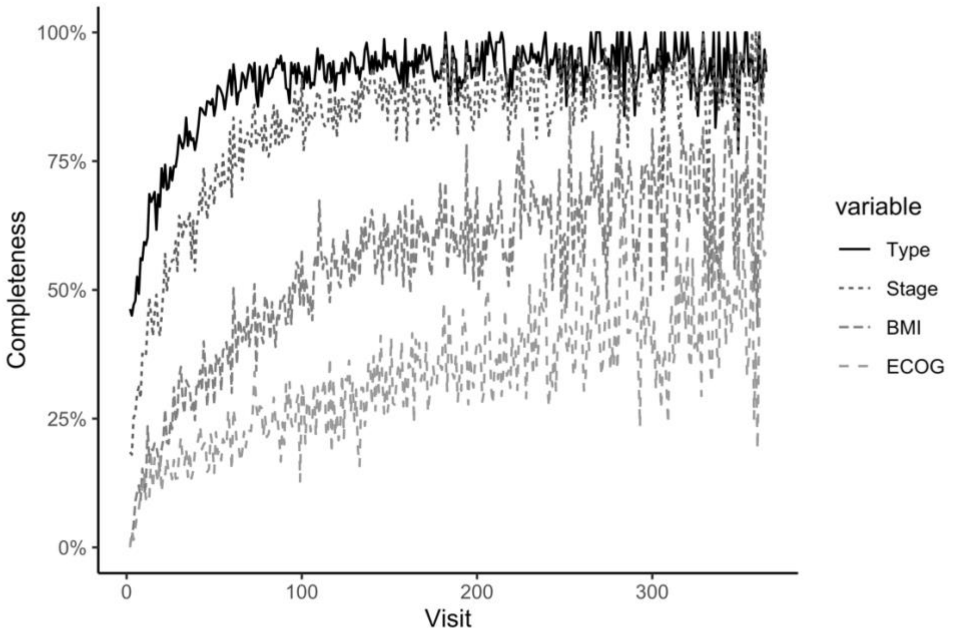


Figure 1.4 Completeness of type, stage, BMI, and ECOG performance status improved with days patients present hospitals.

Data accuracy

The diagnosis date combining ICD time and NICE extracted dates agreed with the chart review and cohort data with median 0 days. 10.4%, 9.0%, and 7.5% of the population having an absolute discrepancy of more than 90 days, 180 days, and one year when compared with chart review results. 12.4%, 8.8%, and 6.4% of the population having an absolute discrepancy of more than 90 days, 180 days, and one year when comparing with BLCS results (Supplemental Table 1.1 and Supplemental Figure 1.1). The accuracy of the diagnosis dates was higher than using ICD time only when comparing with chart review results (Supplemental Table 1.2). Histological type shows great accuracy and agreement with 4.5% discrepancies comparing with chart review and 9.3% discrepancies comparing with BLCS (Supplemental Table 1.3 and Supplemental Table 1.4). Most of the discrepancies exist in classifying adenocarcinoma from non-small cell unspecified. For stages, there were 19.3% discrepancies comparing with chart review and 18.4% discrepancies comparing with BLCS; Compare to chart review, seven patients were one stage category off, three patients were two stage categories off, and one patient was three categories off; Compared to BLCS, 10.9% were one stage category off, 3.6% were two stage categories off, and 3.7% were three categories off (Supplemental Table 1.5 and Supplemental Table 1.6).

For both non-small cell lung cancer patients and small cell patients, two cox proportional models yield similar estimates of the hazard ratio for age, sex, histological type, and stage. (Table 1.6 and Table 1.7).

Table 1.6 Multivariate Cox proportional hazards regression for non-small cell lung cancer patients in BLCS and EMR data

	BLCS cohort Data		**EMR data match with BLCS cohort**		**EMR data[b]**	
	(n=5056)		**(n=4,377)**		**(n=23,420)**	
	HR	**P-value**	**HR**	**P-value**	**HR**	**P-value**
Age at diagnosis	1.02	<0.001	1.02	<0.001	1.02	<0.001
Sex						
Female	ref					
Male	1.38	0.002	1.31	<0.001	1.24	<0.001
Race						
White	ref					
Other	0.93	0.45	0.88	0.21	0.96	0.27
Smoking status						
Never smoker	ref					
Smoker	1.41	<0.001	1.80	<0.001	1.68	<0.001
Type						
Adenocarcinoma	ref					
Squamous cell	1.42	<0.001	1.35	<0.001	1.21	<0.001
Others	1.29	<0.001	1.71	<0.001	1.77	<0.001
Stage						
1	ref					
2	1.50	<0.001	1.52	<0.001	1.67	<0.001
3	2.78	<0.001	3.05	<0.001	2.77	<0.001
4	6.22	<0.001	5.43	<0.001	4.89	<0.001

Note: a.b Patients with complete data of age, sex, stage, type, race, smoking status in both cohort and EMR were included in analysis c. Complete cases of EMR data

Table 1.7 Multivariate Cox proportional hazards regression for small cell lung cancer patients in BLCS and EMR data

	BLCS cohort data		EMR data match with BLCS cohort		EMR data[b]	
	(n=475)		(n=412)		(n=3356)	
	HR	P-value	HR	P-value	HR	P-value
Age at diagnosis	1.02	<0.001	1.02	<0.001	1.02	<0.001
Sex						
Female	ref					
Male	1.38	0.002	1.31	0.02	1.12	0.01
Race						
White	ref					
Other	1.44	0.22	1.08	0.83	1.07	0.45
Smoking status						
Never smoker	ref					
Smoker	1.30	0.57	1.78	0.32	1.55	<0.001
Stage						
limited	ref					
extensive	2.89	<0.001	2.78	<0.001	2.60	<0.001

Note: a. Patients with complete data of age, sex, stage, type, race, smoking status in both cohort and EMR
b. Complete cases of EMR data

Discussion

In this study, we firstly applied a classification algorithm to identify a cohort of lung cancer patients with high PPV and sensitivity. Various EMR components from both structured and unstructured data were utilized to extract demographics, clinical factors, laboratory tests, treatments, and follow-up data.

Besides the tool EXTEND were further developed for extracting BMI and ECOG status, the new NLP tool NICE successfully retrieved essential prognostics factors including cancer stage, location, mutation variables and histological type which demonstrated the benefit of using unstructured EMR data. NICE is able to extract detailed tumor stage information including clinical stage, TNM stage and further grouped stage information such as early stage and advanced stage. Comparing with using structured data only, NICE helped to improve the accuracy of first diagnosis date information. Additionally, NICE was developed for extraction cancer-related data from all types of notes other than only pathology notes that is important to retrieve target data only existing in progress notes, discharge summaries or other types of notes.

We further examined the data completeness and accuracy. The magnitude of data completeness has been increased over the diagnosis year. In 2006, Partners Community Healthcare, Inc. (PCHI) Board resolved to require all PCHI primary care physicians to adopt an EMR by the end of 2008, and all specialists by the end of 2009. With the gradual adoption of EMR, the completeness is expected to increase. The completeness of data is also associated with a certain number of days that they have been present in the hospital. Understanding the variability and pattern of completeness will be essential for choosing the study population and imputation method for further analysis. Variable extracted from EMR generally show high accuracy and agreements, as compared with the manual chart review results and large epidemiology cohort data. In addition, we reported results in terms of the effect size of basic characteristics on overall survival to facilitate interpretation inference.

This large scale EMR-based cohort with detailed longitudinal measurements of clinical factors and patient care data over time would help to better understand how clinical factors influence the progression of lung cancer, investigate the response and patterns of treatments and develop integrative prediction algorithms for clinical outcomes. Our study handles a variety of tasks involving phenotyping, extraction strategies, quality assessment to assemble the cohort that can be applied to other diseases.

This study has serval limitations. First, mortality data collected in structured data as a part of routine clinical care is incomplete as patients may leave the healthcare system and loss to follow up. Not every patient's death was captured. Missing death report inflated estimates of median survival time in our cohort, but hazard ratios estimated from Cox model remain stable, which is consistent with Carrigan's study. [24] One possible solution for this is to augment incomplete mortality data with other sources. Second, the determination of diagnosis date for patients with recurrences or patients transferred from other hospitals is still challenging. Although we exclude patients with lung cancer history code prior to lung cancer-related code, the current data are adequate to identify patients who were initially diagnosed elsewhere. Third, extracting stage through NLP is also challenging, as there were many discrepancies and uncertainties in the notes.

For example, upstage of early stage tumor after surgery, frequent use of “metastatic” without metastatic sites (brain/bone/lymph nodes). Forth, structured data maybe not enough to capture the complete information of treatment, as patients may receive treatments in different healthcare systems. Assumption that patients without a code for drug prescription or treatment procedure were not treated would be violated in this case. Survival analyses comparing cohorts with differential treatment exposure need further consideration. One solution for this is to extract treatment from unstructured data under the assumption that clinicians will document oncology history in clinical notes. Finally, patients within MGH and BWH, the two largest hospitals in Boston but are not a random sample from the population at the US level. It varied depending on the location of the medical institution and may result in biases in the patient demographics and the health condition of admitted patients.

Conclusions

We assembled a large lung cancer cohort from EMRs using phenotyping algorithm and extraction strategies combining structured and unstructured data. The quality of analytic data from EMRs were compared with the well-curated epidemiology study to ensure their suitability for lung cancer prognosis research.

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;68:394-424.

2. SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD,. based on November 2018 SEER data submission, posted to the SEER web site, April 2019., at https://seer.cancer.gov/csr/1975_2016/.)

3. Ashworth AB, Senan S, Palma DA, et al. An individual patient data metaanalysis of outcomes and prognostic factors after treatment of oligometastatic non-small-cell lung cancer. Clin Lung Cancer 2014;15:346-55.

4. Brundage MD, Davies D, Mackillop WJ. Prognostic factors in non-small cell lung cancer: a decade of progress. Chest 2002;122:1037-57.

5. Gaspar LE, McNamara EJ, Gay EG, et al. Small-cell lung cancer: prognostic factors and changing treatment over 15 years. Clin Lung Cancer 2012;13:115-22.

6. 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;5:620-30.

7. Sarraf KM, Belcher E, Raevsky E, Nicholson AG, Goldstraw P, Lim E. Neutrophil/lymphocyte ratio and its association with survival after complete resection in non-small cell lung cancer. J Thorac Cardiovasc Surg 2009;137:425-8.

8. Subramanian J, Morgensztern D, Goodgame B, et al. Distinctive characteristics of non-small cell lung cancer (NSCLC) in the young: a surveillance, epidemiology, and end results (SEER) analysis. J Thorac Oncol 2010;5:23-8.

9. Shepshelovich D, Xu W, Lu L, et al. Body Mass Index (BMI), BMI Change, and Overall Survival in Patients With SCLC and NSCLC: A Pooled Analysis of the International Lung Cancer Consortium. J Thorac Oncol 2019;14:1594-607.

10. Kates M, Swanson S, Wisnivesky JP. Survival following lobectomy and limited resection for the treatment of stage I non-small cell lung cancer<=1 cm in size: a review of SEER data. Chest 2011;139:491-6.

11. Chen VW, Ruiz BA, Hsieh MC, Wu XC, Ries LA, Lewis DR. Analysis of stage and clinical/prognostic factors for lung cancer from SEER registries: AJCC staging and collaborative stage data collection system. Cancer 2014;120 Suppl 23:3781-92.

12. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. Promises and pitfalls of electronic health record analysis. Diabetologia 2018;61:1241-8.

13. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annu Rev Public Health 2016;37:61-81.

14. Nadler E, Espirito JL, Pavilack M, Boyd M, Vergara-Silva A, Fernandes A. Treatment Patterns and Clinical Outcomes Among Metastatic Non-Small-Cell Lung Cancer Patients Treated in the Community Practice Setting. Clin Lung Cancer 2018;19:360-70.

15. Khozin S, Miksad RA, Adami J, et al. Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non-small cell lung cancer. Cancer 2019;125:4019-32.

16. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;106:1-9.

17. Arunachalam A, Li H, Bittoni MA, et al. Real-World Treatment Patterns, Overall Survival, and Occurrence and Costs of Adverse Events Associated With Second-Line Therapies for Medicare Patients With Advanced Non-Small-Cell Lung Cancer. Clin Lung Cancer 2018;19:e783-e99.

18. Cai T, Giannopoulos AA, Yu S, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiographics 2016;36:176-91.

19. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural Language Processing for EHR-Based Computational Phenotyping. IEEE/ACM Trans Comput Biol Bioinform 2019;16:139-53.

20. Yu S, Chakrabortty A, Liao KP, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inform Assoc 2017;24:e143-e9.

21. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc 2019;14:3426-44.

22. Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc 2017;24:339-44.

23. Cai T, Zhang L, Yang N, et al. EXTraction of EMR numerical data: an efficient and generalizable tool to EXTEND clinical research. BMC Med Inform Decis Mak 2019;19:226.

24. Carrigan G, Whipple S, Taylor MD, et al. An evaluation of the impact of missing deaths on overall survival analyses of advanced non-small cell lung cancer patients conducted in an electronic health records database. Pharmacoepidemiol Drug Saf 2019;28:572-81.

Developing a Prognostic Model to Predict Non-Small Cell Lung Cancer 5-Year Overall Survival

Abstract

Background: Survival of non-small cell lung cancer (NSCLC) patients with the same stage varies widely. A more accurate outcome prediction model, which can be applied upon the diagnosis of non-small cell lung cancer, would be essential for informed decisions making regarding clinical care and practice.

Methods: We identified 16,648 NSCLC patients between Jan 2000 and Jan 2015 from Partners HealthCare and collected variables from electronic medical records. Patients were randomly assigned to nonoverlapping training and testing sets to train models and evaluate the performance. Prognostics factors were selected by penalized Cox proportional hazard model with group minimax concave penalty (MCP) penalty and were used to develop the prognostic model and build the nomogram. Model performance was evaluated by the time-dependent area under the receiver operating curves (AUC) and calibration plots.

Results: A total of 11,724 NSCLC patients were included in the analysis. Age, sex, smoking status, histological type, stage, BMI, albumin, ALP, creatinine, HGB, RDW, WBC, NLR, calcium, and sodium were identified as significant predictors of 5-year overall survival (OS). AUCs reached 0.828, 0.825, 0.814, 0.814 and 0.812 for 1- to 5-year prediction, respectively, in

the testing set. The calibration plots showed great agreements between prediction and actual observation survival probability.

Conclusions: We developed and validated a prognostic model for NSCLC patients based on inexpensive and readily available variables collected in routine clinical care. This prognostic tool can be conveniently used to facilitate the prediction of NSCLC survival.

Introduction

Lung cancer is the most commonly diagnosed malignancy and is a leading cause of cancer-related deaths worldwide, with NSCLC accounting for approximately 85% of all diagnosed patients.[1] In the US, the current five-year survival is about 20.6%, and has only been improved slightly from 12.2% over the past four decades.[2] Cancer stage remains the most widely used prognostic factor for NSCLC. The five-year survival rate for NSCLC is about 60% for early stages patients and 6% for advanced-stage patients.[2] However, survival of patients with the same stage varies widely and using TNM staging system as the only predictor for lung cancer survival is imprecise.[3-7] A more accurate outcome prediction model, which can be applied upon the diagnosis of NSCLC, would be essential for informed decisions making regarding clinical care and practice.

To date, no single prognostic model has achieved widespread clinical utility. Some prediction tools were based on small samples of clinical trials with homogeneous patient characteristics, thus were not applicable to real-world patients in oncology practice.[8-11] Some models proposed to include various molecular biomarkers in prediction model.[12-14] However, most of these new molecular markers are not yet available in routine clinical practice. Most of the prediction models utilized demographics (e.g., age, gender, and race) and tumor characteristics (e.g., tumor stage and histology type) and were limited by the lack of other clinical data collected such as laboratory tests results. Routing clinical variables, including laboratory tests, disease history, and BMI, probably play a role in prognosis and may help increase predictive power for NSCLC survival. To our knowledge, there has not been a published study systematically assessed these comprehensive routing clinical variables.

Electronic medical records (EMRs) provide a low-cost means of accessing rich longitudinal data on large populations for research.[15] The data include demographics, health behaviors, outpatient/inpatient/emergency encounters, laboratory data, medication orders, procedures, problem list entries, and clinical notes for healthcare services provided within the system.[16] It allows us to evaluate multiple risk factors and lung cancer prognosis simultaneously, enabling the development of predictive models. This study aimed to develop a prognostic tool using routine clinical variables from EMR to aid physicians and patients in estimating NSCLC survival.

Methods

Study population and data sources

EMR data are from the Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH) using Partners HealthCare System Research Patient Data Registry (RPDR). Institutional Review Board of Partners HealthCare (Protocol Number: 1999P004935/PHS) approved this study. Patients were identified between Jan 2000 and Jan 2015 with histologically and stage confirmed NSCLC. We limited the age range from 18 to 90, excluded patients without routine blood test results within 60 days before or after diagnosis dates.

Data collection

Demographic (age, sex, race), smoking status (smoker, nonsmoker), body mass index (BMI), Eastern Cooperative Oncology Group (ECOG) performance status, tumor characteristics (histological type, stage), history of COPD, history of asthma, history of type 2 diabetes and common laboratory tests were collected. Measurements that within 60 days before or after diagnosis time was considered as the baseline measurements for longitudinal variables. For variables with multiple measurements, the measurement closest to diagnosis dates were used in the analysis. Laboratory tests were from complete blood count (CBC) and comprehensive metabolic panel (CMP). CBC includes white blood count (WBC), neutrophil, lymphocyte, monocyte, and eosinophil and their ratio such as neutrophil-lymphocyte ratio (NLR), red blood count (RBC), red cell distribution width (RDW), hemoglobin (HGB), hematocrit (HCT), platelet count (PLT), mean corpuscular volume (MCV). Routine CMP panel includes albumin, total bilirubin, alkaline phosphatase (ALP), alanine aminotransferase (ALT) and aspartate aminotransferase (AST), blood urea nitrogen (BUN), creatinine, glucose, calcium, sodium, potassium, and chloride. Missing values were coded as a separate missing category. Variables that were categorized include PS (≤ 1, ≥ 2), BMI (underweight: BMI < 18.5 kg/m^2; normal: 18.5 kg/m^2 ≤ BMI < 25 kg/m^2; overweight: 25 kg/m^2 ≤ BMI < 30 kg/m^2; obese: BMI ≥ 30 kg/m^2) and laboratory tests (under, normal, above clinical range) to facilitate easier clinical interpretation. Date of death was collected until Feb 2020.

Statistical analysis

Our outcome of interest is 5-year OS. For 5-year OS, patients who died or who were alive at the last follow-up or 5 years after diagnosis without evidence of death were censored. 5-year OS is defined as the time from the date of diagnosis to the date of death, the date of the latest follow-up, 5 years after diagnosis whichever came first. Patients were randomly assigned to nonoverlapping training (75%) and testing (25%) sets to train models and evaluate the performance.

Penalized regression with group selection of the multi-level categorical covariates was applied for variable section in Cox proportional hazard model. In this study, we used group minimax concave penalty (MCP) as the penalty function.[17] For training set, 10-fold cross-validation was used to select the value of the penalty parameter in a way that minimized the model deviance. Features with non-zero coefficients selected by MCP from the training dataset were used for multivariate Cox proportional hazards analyses and nomogram construction.

Model's discrimination accuracy for predicting 5-year OS was assessed by constructing the time-dependent receiver operating characteristic (ROC) curves and AUC.[18] Time-dependent AUC was calculated each year from the first to the fifth year. The value of the AUC ranges from 0.5 to 1.0, with 0.5 indicating a random prediction and 1.0 indicating that the model perfectly discriminates the outcome with the model. The AUCs of final models were compared with the other two models: 1) model with stage only; 2) model with age, sex, stage, and histology type. Model's calibration capability was assessed by the agreements between predicted and observed death rates at 1-, 3- and 5-year, respectively. A perfect prediction would result in a 45-degree calibration curve.

To facilitate the utility of the models in the clinical setting, nomograms were used to create an intuitive graph of the prediction model, which will give rise to a numerical probability of the overall survival.[19] The results of multivariate Cox regression model incorporating variables from panelized regression were used to build the final nomogram and generate probabilities of OS at 1-, 3- and 5-year after diagnosis. We also created a user-friendly webserver for our nomogram, which calculates survival probabilities for each year and plots the survival curve.

To evaluate the robustness of our model to missing data, we performed a sensitivity analysis on the patients with complete data of the final model and the excluded patients with missing laboratory lab test results. AUCs were recalculated and calibration plots were showed to assess the robustness of the final model. Statistical analyses were conducted using the R software. P values were two-sided with a value of less than 0.05 were considered statistically significant.

Results

A total of 16,648 patients were identified between Jan 2000 and Jan 2015 with histologically and stage confirmed NSCLC. Sixty-one patients younger than 18 or older than 90 were excluded. Besides, 4,854 patients without routine blood test results within 60 days before or after diagnosis dates were excluded. A total of 11,724 patients were included in the final analysis, with 8,793 patients in training set and 2,931 patients in testing sets. Median follow up and median OS were 2.43 and 3.08 (95% CI: 2.94 - 3.25) years, respectively, in total population, 2.41 and 3.08 (95% CI: 2.88 - 3.28) years, respectively, in the training set, and 2.51 and 3.09 (95% CI: 2.87 - 3.44) years, respectively, in the testing set. Baseline characteristics of the patients in the training and testing sets were summarized in Table 2.1. The distribution of these variables showed no difference between the training and testing sets. The distributions of laboratory variables in this study were summarized in Supplemental Table 2.1. Around half of the patients were female; the median age was around 67 years; the majority of them were white and smokers. Among the collected variables, ECOG performance status and LDH had missing values higher than 30% and were excluded from further analysis. The missing values for the remaining variables were coded as a separate missing group.

Table 2.1 Characteristics of patients in the training set and testing set

Patient characteristic	Training set N=8793 (%)	Testing set N=2931 (%)	*p-value*
Age, median, (years)	66.88	66.57	0.323
Sex			
Female	4720 (53.7)	1586 (54.1)	0.7
Male	4073 (46.3)	1345 (45.9)	
Race			0.937
White	8100 (92.1)	2698 (92.1)	
Others	693 (7.9)	233 (7.9)	
Smoking			0.615
Smoker	8292 (94.3)	2756 (94.0)	
Nonsmoker	501 (5.7)	175 (6.0)	
Histological Type			0.364
Adenocarcinoma	6009 (68.3)	1976 (67.4)	
Squamous	933 (10.6)	302 (10.3)	
NSCLC not specified	978 (11.0)	317 (10.7)	
Stage			0.465
1	2645 (30.1)	868 (29.6)	
2	1136 (12.9)	385 (13.1)	
3	1983 (22.6)	699 (23.8)	
4	3029 (34.4)	979 (33.4)	
BMI			0.586
Normal	2311 (26.3)	777 (26.5)	
Obese	1527 (17.4)	487 (16.6)	
Over	2221 (25.3)	773 (26.4)	
Under	201 (2.3)	58 (2.0)	
Missing	2533 (28.8)	836 (28.5)	
History of COPD			0.307
No	6791 (77.2)	2291 (78.2)	
Yes	2002 (22.8)	640 (21.8)	
History of asthma			0.854
No	6956 (79.1)	2324 (79.3)	
Yes	1837 (20.9)	607 (20.7)	
History of diabetes			0.226
No	8723 (99.2)	2900 (98.9)	
Yes	70 (0.8)	31 (1.1)	

Fifteen variables with non-zero coefficients were finally retained by the group MCP in the training set, including age, sex, smoking status, histological type, stage, BMI, albumin, ALP, creatinine, HGB, RDW, WBC, NLR, calcium, and sodium. These variables were used for multivariate Cox proportional hazards analysis and construction of the nomogram. As shown in Table 2.2, these variables were independent predictors of OS with significant *P* values. In the nomogram, the final risk score was calculated by adding up the point of each item using the nomogram depicted in Figure 2.1 and aligned to the total point axis to estimate the 1-, 3- and 5-year survival probabilities. The nomogram showed that stage contributed most to the survival prediction, followed by calcium, albumin, smoking, and histological type.

Table 2.2 Predictors of 5-year overall survival in training set by multivariate Cox proportional hazards regression

	HR	95% CI	P-value
Sex			
Female	Ref		
Male	1.23	1.16-1.3	<0.001
age	1.02	1.01-1.02	<0.001
Smoking			
Non-smoker	Ref		
Smoker	1.66	1.46-1.89	<0.001
Stage			
1	Ref		
2	1.73	1.54-1.94	<0.001
3	2.92	2.66-3.2	<0.001
4	5.08	4.65-5.54	<0.001
Type			
Adenocarcinoma	Ref		
Squamous cell carcinoma	1.05	0.98-1.13	0.16
Other	1.52	1.4-1.65	<0.001
BMI			
Normal	Ref		
Obese	0.88	0.8-0.97	0.01
Over	0.9	0.83-0.98	0.02
Under	1.28	1.06-1.55	0.01
Missing	1.38	1.28-1.49	<0.001

Table 2.2 (Continued)

Albumin			
<=3.5 g/dl	Ref		
>3.5 g/dl	0.66	0.61-0.71	<0.001
Missing	0.48	0.41-0.57	<0.001
ALP			
<=140 IU/L	Ref		
>140 IU/L	1.4	1.27-1.54	<0.001
Missing	1.17	0.99-1.38	0.06
Creatinine			
Normal	Ref		
Above	1.02	0.95-1.1	0.57
Under	1.45	1.26-1.67	<0.001
Missing	1.19	0.9-1.57	0.22
HBG			
Normal	Ref		
Above	1.35	1.08-1.7	0.01
Under	1.16	1.09-1.24	<0.001
Missing	1.56	0.96-2.55	0.07
RDW			
<=14.5%	Ref		
>14.5%	1.12	1.05-1.2	<0.001
Missing	0.74	0.46-1.2	0.23
WBC			
4.5-11*10^9/L	Ref		
>=11*10^9/L	1.17	1.09-1.25	<0.001
Missing	1.01	0.64-1.59	0.97
NLR			
<=4	Ref		
>4	1.34	1.25-1.43	<0.001
Missing	0.85	0.76-0.96	0.01
Calcium			
8.5-10.5 mg/dl	Ref		
<=8.5 mg/dl	0.72	0.66-0.79	<0.001
>=10.5 mg/dl	1.37	1.17-1.59	<0.001
Missing	1.08	0.84-1.4	0.54
Sodium			
135-145 mEq/L	Ref		
<=135 mEq/L	1.32	1.23-1.43	<0.001
>=145 mEq/L	0.95	0.77-1.16	0.59
Missing	1.04	0.73-1.48	0.82

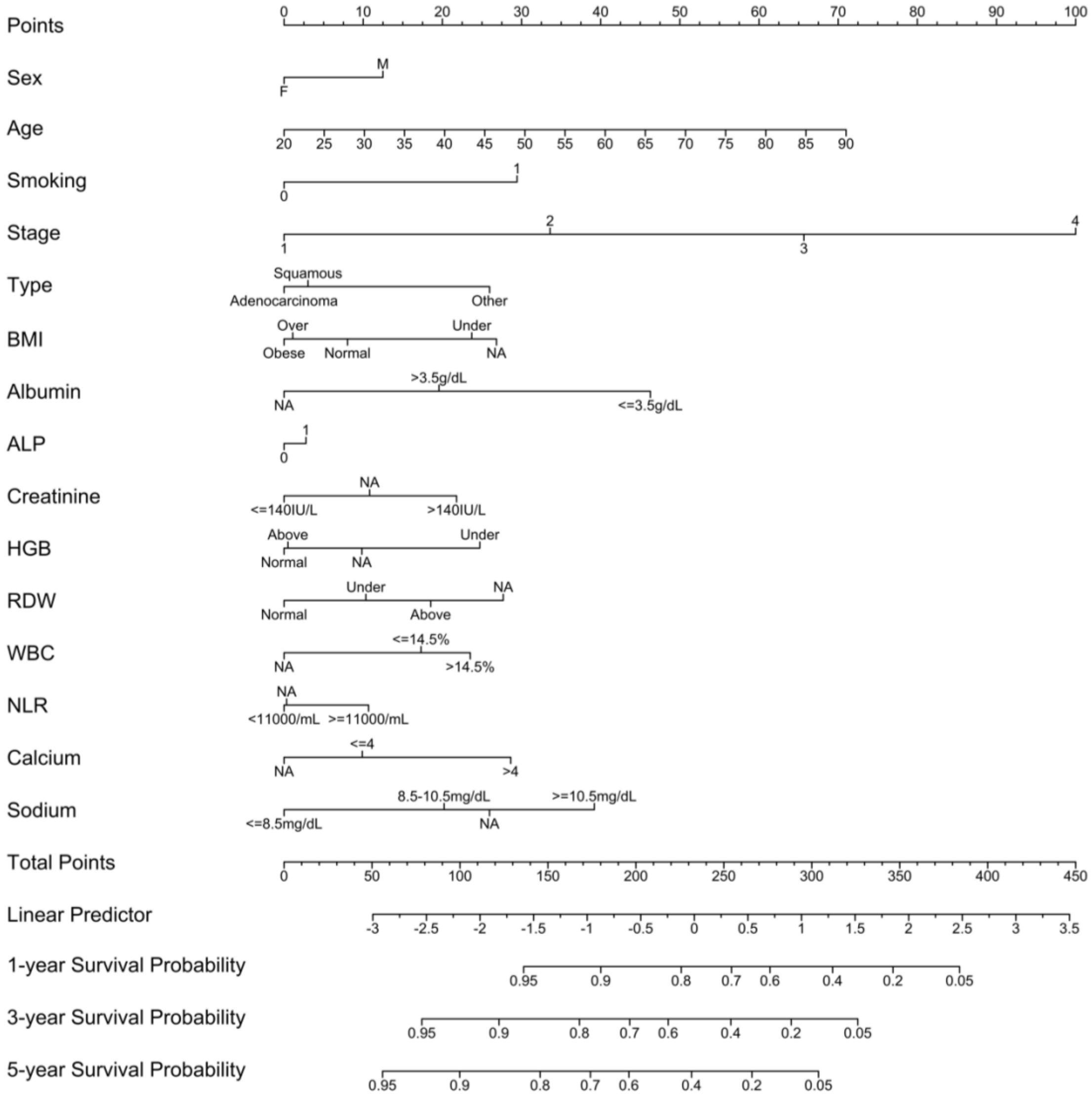


Figure 2.1. Prognostic nomogram for NSCLC patients

AUCs for 1 to 5-year overall survival were calculated to assess the discrimination of the final model. As shown in the Figure 2.2, the AUCs reached 0.830, 0.820, 0.819, 0.817 and 0.813 for 1- to 5-year prediction in the training, 0.828, 0.825, 0.814, 0.814 and 0.812 for 1- to 5-year prediction in the testing set. The prognostic ability of the proposed model was better the basic model with sex, age, histology type, and stage. In training set, the integrated AUC of the proposed model was 0.820, whereas that of the model with sex, age, histology type, and stage was 0.783. In testing set, the integrated AUC of the proposed model was 0.819, whereas that of model with sex, age, histology type, and stage was 0.779. The calibration plots showed that the observed probabilities of survival were generally within 95% CI of the predicted probabilities of survival at 1-, 3- and 5-years after diagnosis, respectively (Figure 2.3). The model slightly underestimated survival among individuals having survival probability higher than 0.8 in the first year after diagnosis and overestimated survival among individuals having survival probability around 0.6 in the fifth year. An online version of our final model can be accessed at https://qyyuan.shinyapps.io/lcprog.

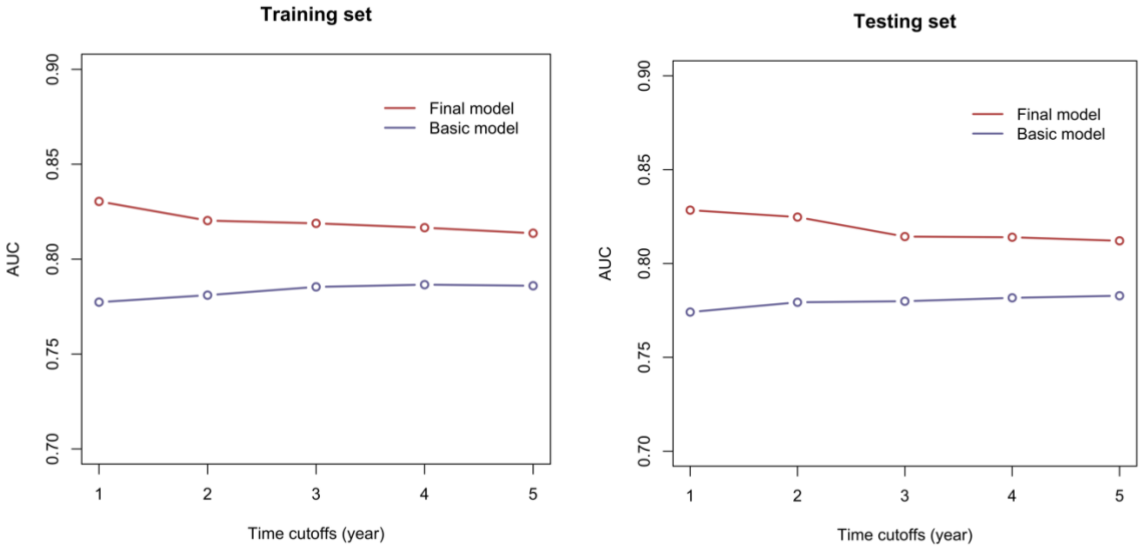


Figure 2.2 Time-dependent AUCs for 1- to 5-year in training set and testing set. The final model included age, sex, smoking status, histological type, stage, BMI, albumin, ALP, creatinine, HGB, RDW, WBC, NLR, calcium, and sodium. Basic model included age, sex, histological type, and stage.

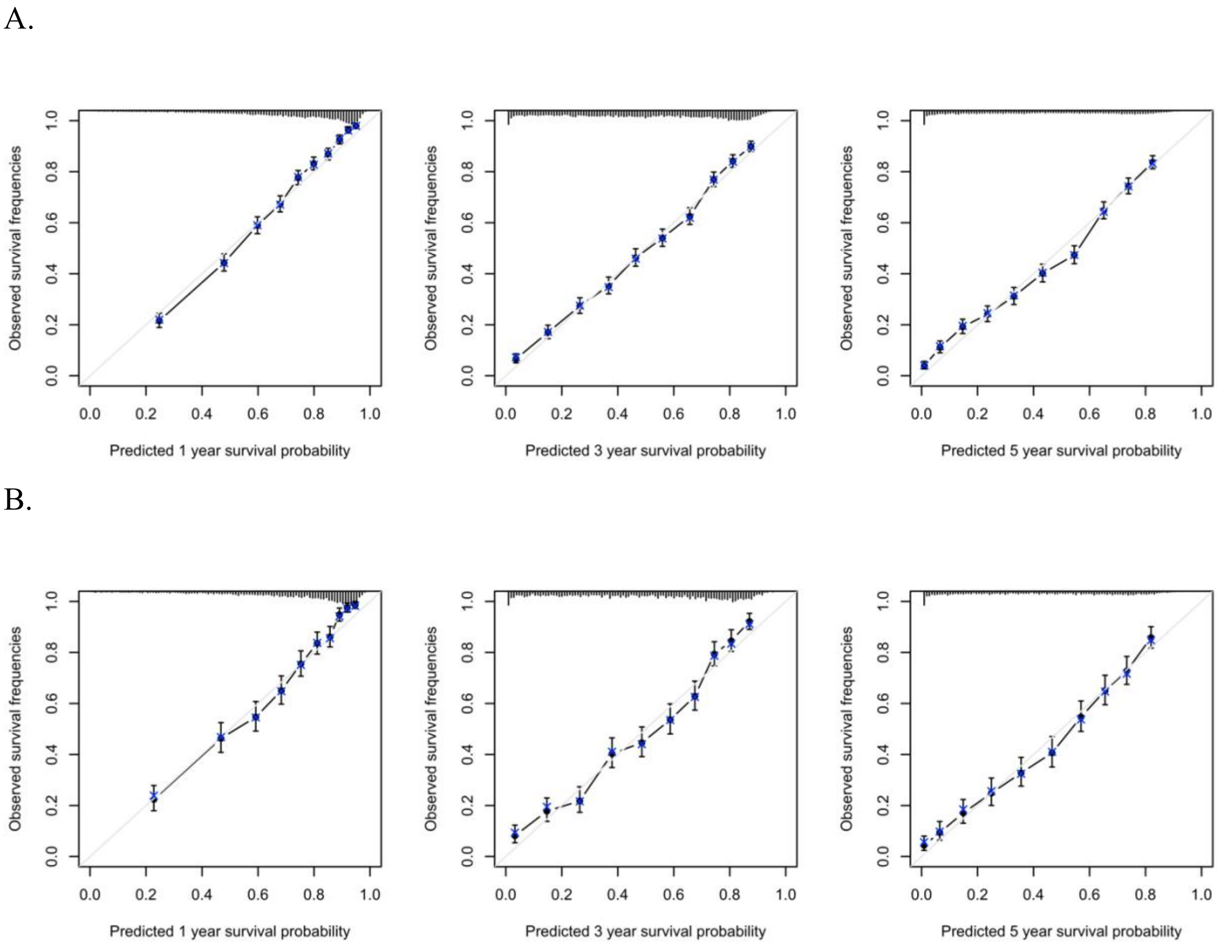


Figure 2.3 Calibration curves compare predicted and actual survival probabilities at 1-year, 3-year, and 5-years. A plot along the 45-degree line would indicate a perfect calibration model in which the predicted probabilities are identical to the actual outcomes. (A). Training set; (B). Testing set.

In the sensitivity analysis, the integrated AUCs were 0.802 and 0.818 for patients with complete data (n=6,548) and patients without laboratory results (n=4,854), respectively. The calibration curves at 1-, 3-, 5-years (Supplemental Figure 2.1) still showed high consistency between predicted survival probability and actual survival proportion. The results proved the robustness of this model to missing data.

Discussion

Accurate assessment of patient's prognosis is essential for clinicians and patients to guide disease management. Prediction of survival using staging is not enough since NSCLC is remarkably heterogeneous. In this study, a prognostics model was developed and validated using a large cohort of NSCLC patients from real-world clinical care. The cohort was obtained from two largest hospitals (MGH and BWH) in Boston; variables were extracted from electronic medical records using structured data and clinical notes. In the validation, our model achieved high discrimination and calibration. Discrimination ability was revealed by the higher time-dependent AUCs compared with the model based on sex, age, stage, and histological type. The calibration plot showed great agreements between prediction and actual observation survival probability.

Through penalized regression, age, sex, smoking status, histological type, stage, BMI, albumin, ALP, creatinine, HGB, RDW, WBC, NLR, calcium, and sodium were identified as independent prognostics factors. These findings were consistent with previously reported prognosis factors. Specifically, older age, male sex, and advanced stages have been associated with poor survival.[20] Compared to adenocarcinoma, squamous cell carcinoma and other NSCLC have shown worse survival.[21] Patients who were underweight (BMI < 18.5 kg/m^2) have been associated with worse prognosis and patients who were overweight (25 kg/m^2 ≤ BMI < 30 kg/m^2) or obese (BMI ≥ 30 kg/m^2) have been reported to be associated with improved survival.[22] High albumin level that measures nutritional status has been reported to be associated with better survival.[23] Increased serum alkaline phosphatase has been reported to be associated with bone metastasis and poor survival.[24] Compared to patients with normal HGB level, patients with decreased HGB have shown poor overall survival.[25] Patients with higher RDW values had poorer prognoses than those with lower RDW values.[26] WBC and NLR are essential markers of immune functions and the inflammatory response. Previous studies have shown that high WBC, high NLR contributed to decreased survival.[27-29] Besides, patients with hypercalcemia or hyponatremia of malignancy often had a poor prognosis.[30,31] Our study also identified new prognostics factors. Low creatinine levels reflect low muscle mass, or malnutrition was associated with poor survival.[32] Increased HGB level that probably caused by COPD contributed to poor survival.

Prognostic models with considerable heterogeneity in the selection of prognostic factors have been proposed in many studies. Most of the prognostics model included age, sex, histological type and tumor stage.[3,33] Smoking, BMI, routine laboratory tests or comorbidity data were included in some of the models.[8,27,34-37] Compare with other published prognostic models, the main strength of our study is that it assessed routine clinical variables including demographics, tumor characteristics, disease history as well as routine blood-based laboratory test results, thus representing a more comprehensive prognostic model. In our study, we didn't include treatment information because the model was built based on the time of diagnosis before any treatment.

This study has several limitations. First, morality data is incomplete as patients may leave the healthcare system and loss to follow up. Missing death information inflated inflated estimates of median survival time in our cohort, but hazard ratios estimated from Cox model remain stable. The second limitation was a lack of ECOG performance status and LDH data due to the high missing rate, which probably plays a role in prognosis. Incorporation of these relevant variables would probably help to improve this model. Third, the missingness of the covariates is not missing completely at random. Patients who missed baseline laboratory test results were more likely to come to Partners for a consultation and not for disease management. In sensitivity analyses, we examined an alternative population based on the missingness of variables; the discrimination and calibration ability remained, supporting the robustness of our final model. Finally, patients were from MGH and BWH and do not represent a random sample from the population at the US level; this can affect the generalizability of the models trained on our data to other medical institutions. External validation of our model in the other population would be desirable.

Conclusions

We developed and validated a prognostic model for NSCLC patients using a large lung cancer patient cohort. This proposed model based on inexpensive and readily available clinical data may provide a more precise survival estimation. The nomogram implemented on an online web server could be useful for clinical counseling.

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;68:394-424.

2. SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD,. based on November 2018 SEER data submission, posted to the SEER web site, April 2019., at https://seer.cancer.gov/csr/1975_2016/.)

3. Putila J, Remick SC, Guo NL. Combining clinical, pathological, and demographic factors refines prognosis of lung cancer: a population-based study. PLoS One 2011;6:e17493.

4. Sarraf KM, Belcher E, Raevsky E, Nicholson AG, Goldstraw P, Lim E. Neutrophil/lymphocyte ratio and its association with survival after complete resection in non-small cell lung cancer. J Thorac Cardiovasc Surg 2009;137:425-8.

5. 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;5:620-30.

6. Brundage MD, Davies D, Mackillop WJ. Prognostic factors in non-small cell lung cancer: a decade of progress. Chest 2002;122:1037-57.

7. Ashworth AB, Senan S, Palma DA, et al. An individual patient data metaanalysis of outcomes and prognostic factors after treatment of oligometastatic non-small-cell lung cancer. Clin Lung Cancer 2014;15:346-55.

8. Florescu M, Hasan B, Seymour L, Ding K, Shepherd FA, National Cancer Institute of Canada Clinical Trials G. A clinical prognostic index for patients treated with erlotinib in National Cancer Institute of Canada Clinical Trials Group study BR.21. J Thorac Oncol 2008;3:590-8.

9. Hoang T, Dahlberg SE, Sandler AB, Brahmer JR, Schiller JH, Johnson DH. Prognostic models to predict survival in non-small-cell lung cancer patients treated with first-line paclitaxel and carboplatin with or without bevacizumab. J Thorac Oncol 2012;7:1361-8.

10. Mandrekar SJ, Schild SE, Hillman SL, et al. A prognostic model for advanced stage nonsmall cell lung cancer. Pooled analysis of North Central Cancer Treatment Group trials. Cancer 2006;107:781-92.

11. Hoang T, Xu R, Schiller JH, Bonomi P, Johnson DH. Clinical model to predict survival in chemonaive patients with advanced non-small-cell lung cancer treated with third-generation chemotherapy regimens based on eastern cooperative oncology group data. J Clin Oncol 2005;23:175-83.

12. Der SD, Sykes J, Pintilie M, et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. J Thorac Oncol 2014;9:59-64.

13. Gyorffy B, Surowiak P, Budczies J, Lanczky A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. PLoS One 2013;8:e82241.

14. Hou J, Aerts J, den Hamer B, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 2010;5:e10312.

15. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annu Rev Public Health 2016;37:61-81.

16. Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc 2017;24:339-44.

17. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics 2010;38:894-942.

18. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000;56:337-44.

19. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol 2008;26:1364-70.

20. Lu T, Yang X, Huang Y, et al. Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. Cancer Manag Res 2019;11:943-53.

21. Tane S, Nishio W, Ogawa H, et al. Clinical significance of the 'not otherwise specified' subtype in candidates for resectable non-small cell lung cancer. Oncol Lett 2014;8:1017-24.

22. Shepshelovich D, Xu W, Lu L, et al. Body Mass Index (BMI), BMI Change, and Overall Survival in Patients With SCLC and NSCLC: A Pooled Analysis of the International Lung Cancer Consortium. J Thorac Oncol 2019;14:1594-607.

23. Gupta D, Lis CG. Pretreatment serum albumin as a predictor of cancer survival: a systematic review of the epidemiological literature. Nutr J 2010;9:69.

24. Li X, Li B, Zeng H, et al. Prognostic value of dynamic albumin-to-alkaline phosphatase ratio in limited stage small-cell lung cancer. Future Oncol 2019;15:995-1006.

25. Huang Y, Wei S, Jiang N, et al. The prognostic impact of decreased pretreatment haemoglobin level on the survival of patients with lung cancer: a systematic review and meta-analysis. BMC Cancer 2018;18:1235.

26. Koma Y, Onishi A, Matsuoka H, et al. Increased red blood cell distribution width associates with cancer stage and prognosis in patients with lung cancer. PLoS One 2013;8:e80240.

27. Zhang K, Lai Y, Axelrod R, et al. Modeling the overall survival of patients with advanced-stage non-small cell lung cancer using data of routine laboratory tests. Int J Cancer 2015;136:382-91.

28. Yu Y, Qian L, Cui J. Value of neutrophil-to-lymphocyte ratio for predicting lung cancer prognosis: A meta-analysis of 7,219 patients. Mol Clin Oncol 2017;7:498-506.

29. Ren F, Zhao T, Liu B, Pan L. Neutrophil-lymphocyte ratio (NLR) predicted prognosis for advanced non-small-cell lung cancer (NSCLC) patients who received immune checkpoint blockade (ICB). Onco Targets Ther 2019;12:4235-44.

30. Seccareccia D. Cancer-related hypercalcemia. Can Fam Physician 2010;56:244-6, e90-2.

31. Fiordoliva I, Meletani T, Baleani MG, et al. Managing hyponatremia in lung cancer: latest evidence and clinical implications. Ther Adv Med Oncol 2017;9:711-9.

32. Thongprayoon C, Cheungpasitporn W, Kashani K. Serum creatinine level, a surrogate of muscle mass, predicts mortality in critically ill patients. J Thorac Dis 2016;8:E305-11.

33. Blanchon F, Grivaux M, Asselain B, et al. 4-year mortality in patients with non-small-cell lung cancer: development and validation of a prognostic index. Lancet Oncol 2006;7:829-36.

34. Vincent MD, Ashley SE, Smith IE. Prognostic factors in small cell lung cancer: a simple prognostic index is better than conventional staging. Eur J Cancer Clin Oncol 1987;23:1589-99.

35. Park MJ, Lee J, Hong JY, et al. Prognostic model to predict outcomes in nonsmall cell lung cancer patients treated with gefitinib as a salvage treatment. Cancer 2009;115:1518-30.

36. Mou W, Liu Z, Luo Y, et al. Development and cross-validation of prognostic models to assess the treatment effect of cisplatin/pemetrexed chemotherapy in lung adenocarcinoma patients. Med Oncol 2014;31:59.

37. Alexander M, Wolfe R, Ball D, et al. Lung cancer prognostic index: a risk score to predict overall survival after the diagnosis of non-small-cell lung cancer. Br J Cancer 2017;117:744-51.

A Prognostics Score for Advanced Non-Small Cell Lung Cancer Treated with PD-1/L1 Inhibitors

Abstract

Background: Programmed death 1 (PD-1)/programmed death-legend 1(PD-L1) inhibitors have shown clinical benefits for a proportion of advanced lung cancer patients. This study aimed to develop a prognostic store with routinely available variables to identify non-small cell lung cancer (NSCLC) patients who likely benefit from PD-1/L1 inhibitors.

Methods: 412 patients who received PD-1/L1 inhibitors were retrospectively collected from Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH) between Nov 2013 and Jan 2018. Demographic, clinical, and common laboratory tests were collected. Penalized Cox proportional hazard model with group minimax concave penalty (MCP) penalty was applied for variable section and coefficient estimation. Model performance was assessed by calculating area under the receiver operating curves (AUC). A prognostic score was developed to categorize advanced NSCLC patients into good, intermediate and poor groups.

Results: Median PFS was 3.6 (95% CI: 2.9 - 4.3) months, median OS was 18.4 (95% CI: 14.2-24.9) months. A prognostic scoring model incorporating the weighted coefficients of PD-L1 expression level, EGFR mutation status, ECOG performance status, and albumin level was developed. Median PFS were 1.7 (95% CI: 1.5-2.1) vs 3.2 (95% CI: 2.7-6.3) vs 6.2 (95% CI: 4.9–9.7) months for the poor, intermediate and good groups, respectively. Median OS were 8.2

(95% CI: 4.5-14.1) vs 14.9 (95% CI: 11.8-30.0) vs 34.7 (95% CI: 26.2 -NE) months for the poor, intermediate and good groups, respectively.

Conclusions: a prognostic score combining PD-L1 expression level, EGFR mutation status, ECOG performance status, and albumin level may help to identify advanced NSCLC patients who likely benefit from PD-1/L1 inhibitors.

Introduction

Immunotherapy, especially programmed death 1 (PD-1)/programmed death-legend 1(PD-L1) inhibitors, has shown survival benefits over conventional chemotherapy.[1-6] In March 2015, nivolumab was approved by the U.S. Food and Drug Administration (FDA) as second line or above treatment. In October 2015, pembrolizumab was approved for patients whose tumors positively express PD-L1 after chemotherapy. In October 2016, pembrolizumab received approval as a first line. And in October 2016, atezolizumab was approved as second line or above treatment. Currently, PD-L1 expression is the most validated biomarker for patients treated with PD-1/L1 inhibitors.[7] Studies have generally shown that patients with positive PD-L1 expression had higher rates of objective response compared to patients with negative or weak PD-L1 expression.[2,6,8] Nevertheless, the benefit is not seen for all patients with positive PD-L1 expression, and some patients with negative PD-L1 expression can still achieve clinical benefit with PD-1/L1 inhibitors. Therefore, the development of prognostic biomarkers that can complement PD-L1 expression is essential to identify NSCLC patients who most likely to respond to immunotherapy, and to avoid unnecessary toxicity and high costs for these non-responders.[9]

Many prognostic score have been proposed such as the Gustave Roussy Immune Score (GRIm-Score), the Royal Marsden Hospital prognostic score (RMH score)[10], lung immune-based prognostics score(LIPI)[11], EPSILoN (ECOG PS, smoking, liver metastases, LDH, NLR)[12], advanced lung cancer inflammation index (ALI)[13], immunotherapy sex-ECOG-NLR-delta NLR (iSEND)[14] and systemic inflammation index (SII)[15], which mainly investigated the role of clinical characteristics and peripheral blood markers in the immunotherapy response. Most of them incorporated ECOG PS, neutrophil-to-lymphocyte (NLR) ratio, derived neutrophil-to-lymphocyte (dNLR) ratio, lactate dehydrogenase (LDH) level or albumin, that measure inflammatory or nutritional status and have shown notable association with clinical response. However, they were limited by the number of variables collected, and hence some factors important for prognosis were not included in the prediction models. For example, PD-1/L1 inhibitor has been found to be less effective in patients with EGFR mutation than in those without the mutation.[16] Also, baseline use of prednisone has been reported to be associated with poorer outcome given the immunosuppressive properties of corticosteroids.[17] Combining these factors into the scoring system may lead to a better ability to stratify patients to various prognostic groups.

Based on the rich longitudinal data on demographics, tumor characteristics, medical history, and treatment information in electronic medical records, we were able to identify advanced NSCLC patients who received PD-1/L1 inhibitors and conduct a comprehensive assessment of routine clinical variables on clinical outcomes. We aimed to construct a prognostic score combining baseline patient and clinical factors to identify patients who likely benefit from PD-1/L1 inhibitors.

Methods

Study population

We retrospectively collected advanced NSCLC patients (advanced NSCLC at diagnosis or early-stage NSCLC with a recurrence or progression) who received at least one dose of PD-1/L1 inhibitor (nivolumab, pembrolizumab and atezolizumab) between Nov 2013 and Jan 2018 from Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH). The Institutional Review Board of Partners HealthCare (Protocol Number: 1999P004935/PHS) approved this study. Patients who had at least one follow up visit after the first dose of PD-1/L1 inhibitor monotherapy were included in analysis.

Data collection

Treatment initiation time was extracted from the date of the first dose of PD-1/L1 inhibitors. The duration of the treatment was calculated by the time difference between the first and last dose of PD-1/L1 inhibitors. Demographics, histological type, prior treatments (surgery, chemotherapy, radiation therapy), PD-L1 expression level (0%, 1-49%, and ≥50%), driven mutations (*EGFR, KRAS*, and *ALK*), history of autoimmune disease, use of prednisone before treatment initiation, Eastern Cooperative Oncology Group (ECOG) performance status, smoking status (smoker, nonsmoker), body mass index (BMI) and common laboratory tests were collected using Partners HealthCare System Research Patient Data Registry (RPDR) or medical records review. Measurements that within 30 days prior to the initiation of the PD-1/L1 inhibitors were considered as the baseline measurements for longitudinal variables. Clinical outcomes were extracted using medical records review. The date of progression was extracted if progression was documented in oncology reports assessed by computed tomography scan. For patients who were not observed progression, the dates of last follow up, the dates of last computed tomography scan without disease progression were extracted.

Laboratory tests were from complete blood count (CBC) and comprehensive metabolic panel (CMP). CBC includes white blood count (WBC), neutrophil, lymphocyte, monocyte, and eosinophil and their ratio such as neutrophil-lymphocyte ratio (NLR), red blood count (RBC), red cell distribution width (RDW), hemoglobin (HGB), hematocrit (HCT), platelet count (PLT), mean corpuscular volume (MCV). Routine CMP panel includes albumin, total bilirubin, total protein, alkaline phosphatase (ALP), alanine aminotransferase (ALT) and aspartate aminotransferase (AST), blood urea nitrogen (BUN), creatinine, glucose, calcium, sodium, potassium, and chloride.

Statistical analysis

Progression free survival (PFS) was defined as the time from treatment initiation to tumor progression or death from any cause, with censoring of patients who were lost to follow-up or discontinue treatments without observation of the progression. Overall survival (OS) was defined as the time from treatment initiation to death, with censoring of patients who were lost to follow-up. Missing values for PD-L1 expression, somatic mutations, ECOG PS, BMI, and laboratory tests were coded as a separate untested category. Variables that were categorized include PS (0–1, >=2), BMI (underweight: BMI < 18.5 kg/m$_2$; normal: 18.5 kg/m$_2$ ≤ BMI < 25 kg/m$_2$; overweight: 25 kg/m$_2$ ≤ BMI < 30 kg/m$_2$; obese: BMI ≥ 30 kg/m$_2$) and laboratory tests (under, normal, above clinical normal range) to facilitate easier clinical interpretation.

Cox proportional hazard regression was used to fit a survival model for progression free survival. Penalized regression with group selection of the multi-level categorical covariates was applied for variable section and coefficient estimation. Group minimax concave penalty (MCP) was used as the penalty term.[18] Compare to group LASSO, which is the most commonly used penalty term, group MCP is more consistent for variable selection and yield estimators with asymptotic normality.[19] Feature selection procedure was combined with leave-one-out cross-validation (LOOCV) to avoid overfitting in the selection of features and achieve an unbiased estimate of model performance. In LOOCV, one sample was left out at a time and used for testing, while the remaining samples were used for training. For each training set, we used 10-fold cross-validation to select the value of the penalty parameter in a way that minimized the model deviance. To measure the performance of prediction model selected by Group MCP, time-dependent area under the receiver operating characteristic (ROC) curve, and the corresponding estimate of the AUC at each month from the 1st to the 12th month was calculated to assess the discrimination for PFS.[20] To construct the prognostics score, Group MCP was applied in the pooled dataset to select the final variables and calculate the coefficients of these variables. A prognostic scoring model incorporating the weighted coefficients of these variables was developed. The pooled population were then grouped into three subgroups according to the developed prognostic score. Survival curves were estimated using the Kaplan-Meier method and compared by log rank test.

Discontinuation of the treatment could be due to disease progression, intolerable adverse effects, lost follow up, move to hospice care, or death.[21] In the primary analysis, we treated the observation at the last visit time with no documented progression as right-censored, which leads to potential informative censoring. In the sensitivity analysis, treatment discontinuation due to adverse effect and hospice care was considered as a disease progression event. Subgroup analysis was conducted for patients who received different types of PD-1/L1 inhibitors to assess the utility of prognostic score. Statistical analyses were conducted using the R software. *P* values less than 0.05 were considered statistically significant.

Results

A total of 412 patients were collected retrospectively with a diagnosis of advanced NSCLC treated with PD-1/L1 inhibitors. 274 patients (67%) had disease progression, and 180 patients (44%) were dead. Among patients who did not observe disease progression before treatment discontinuation, 31 patients were not progressed, 69 patients had severe adverse effects, 15 patients moved to hospice, and 28 patients were lost to follow up. Median follow up was 7.1 (95% CI: 5.8 - 8.4) months, median PFS was 3.6 (95% CI: 2.9 - 4.3) months and median OS was 18.4 (95% CI: 14.2-24.9) months.

Baseline characteristics of the patients were summarized (Table 3.1). 46.1% of patients were female; the median age was 67 years; 71.8% of them were adenocarcinoma. Most of them were smokers (87.1%) and without an autoimmune history (91.5%). 29.9% had an ECOG PS of 2 or higher. 72.8% of patients received nivolumab, 20.9% received pembrolizumab, and 6.3% of them received atezolizumab. Most of the patients received PD-1/L1 inhibitors as second line or above, with a history of surgery, chemotherapy, or radiation therapy. For PD-L1 expression, 39.1% of the patients were tested with 48 patients tested negative (PD-L1 expression <1%), 35 patients tested weak expression (PD-L1 expression between 1% and 49%) and 74 patients tested strong expression (PD-L1 expression ≥50%). 7.5%, 33.0%, and 1.9% of patients were tested positive for *EGFR, KRAS*, and *ALK*, respectively. 10% of patients used prednisone before treatment initiation. The distributions of laboratory variables in this study were summarized in Supplemental Table 3.1.

Table 3.1 Patients' characteristics

Characteristics	Number (%)	Characteristics	Number (%)
Age at treatment initiation	67.3±10.4	Surgery history	
Gender		No	331 (80.3)
Female	190 (46.1)	Yes	81 (19.7)
Male	222 (53.9)	Chemotherapy history	
Type		No	66 (16.0)
Adenocarcinoma	296 (71.8)	Yes	346 (84.0)
Squamous cell	100 (24.3)	Radiation therapy history	
Other	16 (3.9)	No	168 (40.8)
Smoker		Yes	244 (59.2)
Non-smoker	53 (12.9)	PD L1 expression status	
Smoker	359 (87.1)	<1%	48 (11.7)
BMI		1-49%	35 (8.5)
Under	17 (4.1)	>=50%	74 (18.0)
Normal	142 (34.5)	Untested	255 (61.9)
Over	158 (38.3)	EGFR mutation status	
Obese	60 (14.6)	Wild type	300 (72.8)
Untested	35 (8.5)	Mutant	31 (7.5)
ECOG		Untested	81 (19.7)
0-1	229 (55.6)	KRAS mutation status	
>=2	123 (29.9)	Wild type	193 (46.8)
Untested	60 (14.6)	Mutant	136 (33.0)
Drug type		Untested	83 (20.1)
Atezolizumab	26 (6.3)	ALK mutation status	
Nivolumab	300 (72.8)	Wild type	267 (64.8)
Pembrolizumab	86 (20.9)	Mutant	8 (1.9)
First line		Untested	137 (33.3)
No	359 (87.1)	Autoimmune disease history	
Yes	53 (12.9)	No	377 (91.5)
		Yes	35 (8.5)
		Use of prednisone before treatment	
		No	371 (90.0)
		Yes	41 (10.0)

The AUCs from the 1st to the 12th month calculated from penalized Cox proportional hazard model with LOOCV were presented in Figure 3.1. AUCs were higher than that for the model based solely on PD-L1 expression.

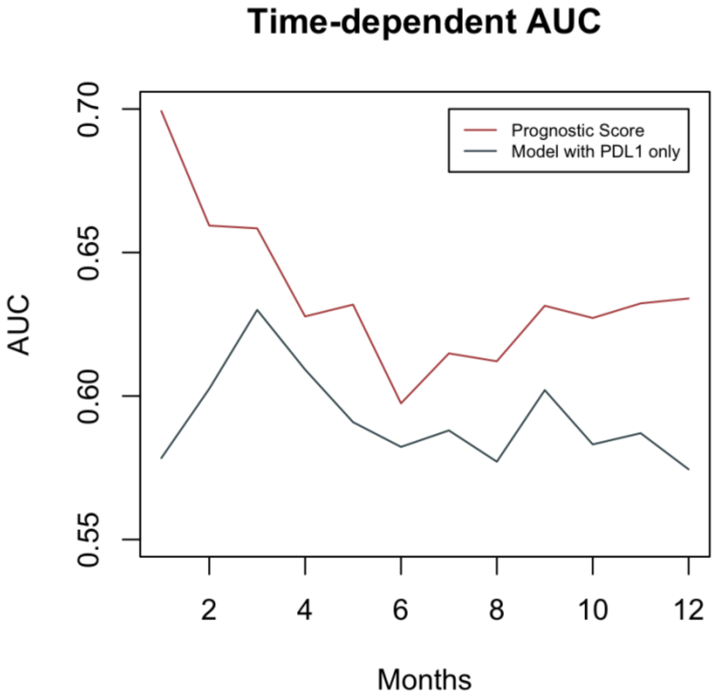


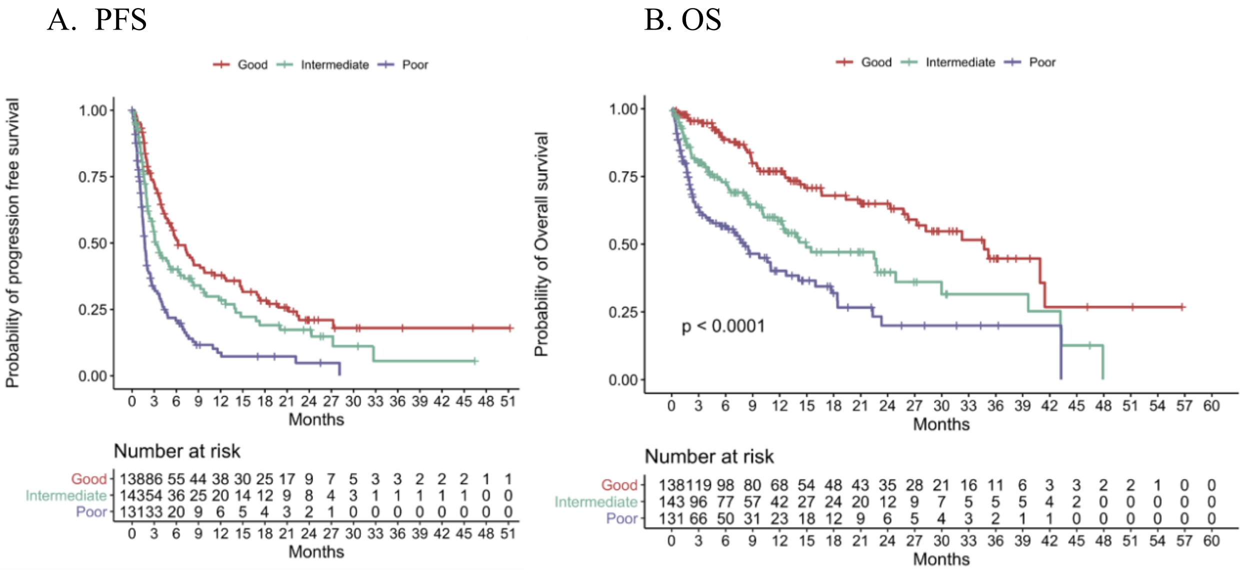
Figure 3.1 Time-dependent AUCs for the prognostic score for every month from the first to the 12th month.

Four variables were selected in the final model by applying group MCP in the pooled dataset, including PD-L1 expression level, EGFR mutation status, ECOG performance status, and albumin ≥ 3.5g/dl. A prognostic scoring model was developed, incorporating the weighted coefficients of these variables (Table 3.2). The prognostic score grouped advanced NSCLC patients into three subgroups having approximately the same sample size based on the tertile distribution: poor group (n= 131) with score > 0.564, intermediate group (n= 143) with score 0.184-0.564 and good group (n=138) with score ≤0.184.

Table 3.2 Weighted coefficients for prognostic score

	Effect size
PD L1 expression	
>=50%	0
1-49%	0.119
<1%	0.564
Untested	0.128
EGFR	
Wild type	0
Mutant	0.920
Untested	0.511
ECOG	
1-2	0
>=2	0.055
Untested	0.035
Albumin	
>3.5g/dl	0
<=3.5g/dl	0.301
Untested	0.148

The prognostic score was associated with PFS with *P*< 0.001. Median PFS were 1.7 (95% CI: 1.5-2.1) vs 3.2 (95% CI: 2.7-6.3) vs 6.2 (95% CI: 4.9–9.7) months for the poor, intermediate and good groups, respectively. The prognostic score was also associated with OS with *P*< 0.001. Median OS were 8.2 (95% CI: 4.5-14.1) vs 14.9 (95% CI: 11.8-30.0) vs 34.7 (95% CI: 26.2 -NE) months for the poor, intermediate and good groups, respectively.(Figure 3.2) In multivariate analysis of the four covariates in the prognostic score, all of them were significantly associated with both PFS and OS, as demonstrated in Table 3.3.



A. PFS

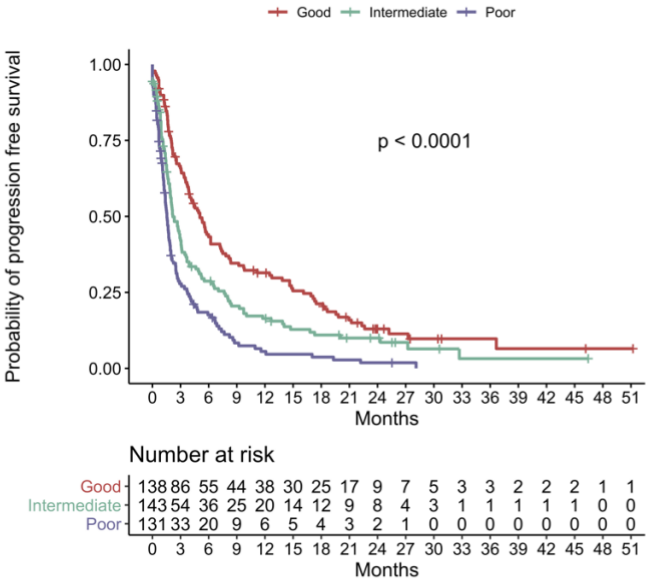
B. OS

Figure 3.2 PFS and OS according to prognostics score groups.

Table 3.3 Multivariate analysis for PFS and OS

		PFS		OS	
		HR (95% CI)	*P* value	HR (95% CI)	*P* value
PD L1 expression					
	>50%	1 (ref)			
	1-49%	1.20 (0.72-1.99)	0.49	2.39 (1.36-4.17)	0.002
	<1%	2.63 (1.70-4.07)	<0.001	0.95 (0.49-1.86)	0.95
	Untested	1.27 (0.90-1.79)	0.17	1.54 (0.99-2.38)	0.06
EGFR					
	Wild type	1 (ref)			
	Mutant	2.65 (1.72-4.08)	<0.001	2.38 (1.36-4.15)	0.002
	Untested	1.67 (1.23-2.28)	<0.001	2.29 (1.59-3.29)	<0.001
ECOG					
	1-2	1 (ref)			
	>=2	1.46 (1.08-1.98)	0.02	2.24 (1.55-3.24)	<0.001
	Untested	1.27 (0.88-1.83)	0.20	1.71 (1.12-2.61)	0.02
Albumin					
	>3.5g/dl	1 (ref)			
	<=3.5g/dl	1.54 (1.14-2.08)	<0.001	2.73 (1.89-3.94)	<0.001
	Untested	1.29 (0.94-1.77)	0.14	1.45 (0.97-2.15)	0.07

In the sensitivity analysis, treating intolerable adverse effects or move to hospice care as progression led to a decrease of median survival time but still showed a significant difference for three groups. (Figure 3.3). Median PFS were 1.6 (95% CI: 1.3-1.9) vs 2.2 (95% CI: 1.9-3.1) vs 5.1 (95% CI: 3.9-7.2) months for the poor, intermediate and good groups, respectively.



	0	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51
Good	138	86	55	44	38	30	25	17	9	7	5	3	3	2	2	2	1	1
Intermediate	143	54	36	25	20	14	12	9	8	4	3	1	1	1	1	1	0	0
Poor	131	33	20	9	6	5	4	3	2	1	0	0	0	0	0	0	0	0

Figure 3.3 PFS according to prognostics score group in the sensitivity analysis.

Subgroup analysis of PFS according to drug type showed that the ability of the prognostic score to stratify patients was maintained for patients treated with nivolumab and pembrolizumab, the association was not seen in the atezolizumab given the small sample size (Figure 3.4).

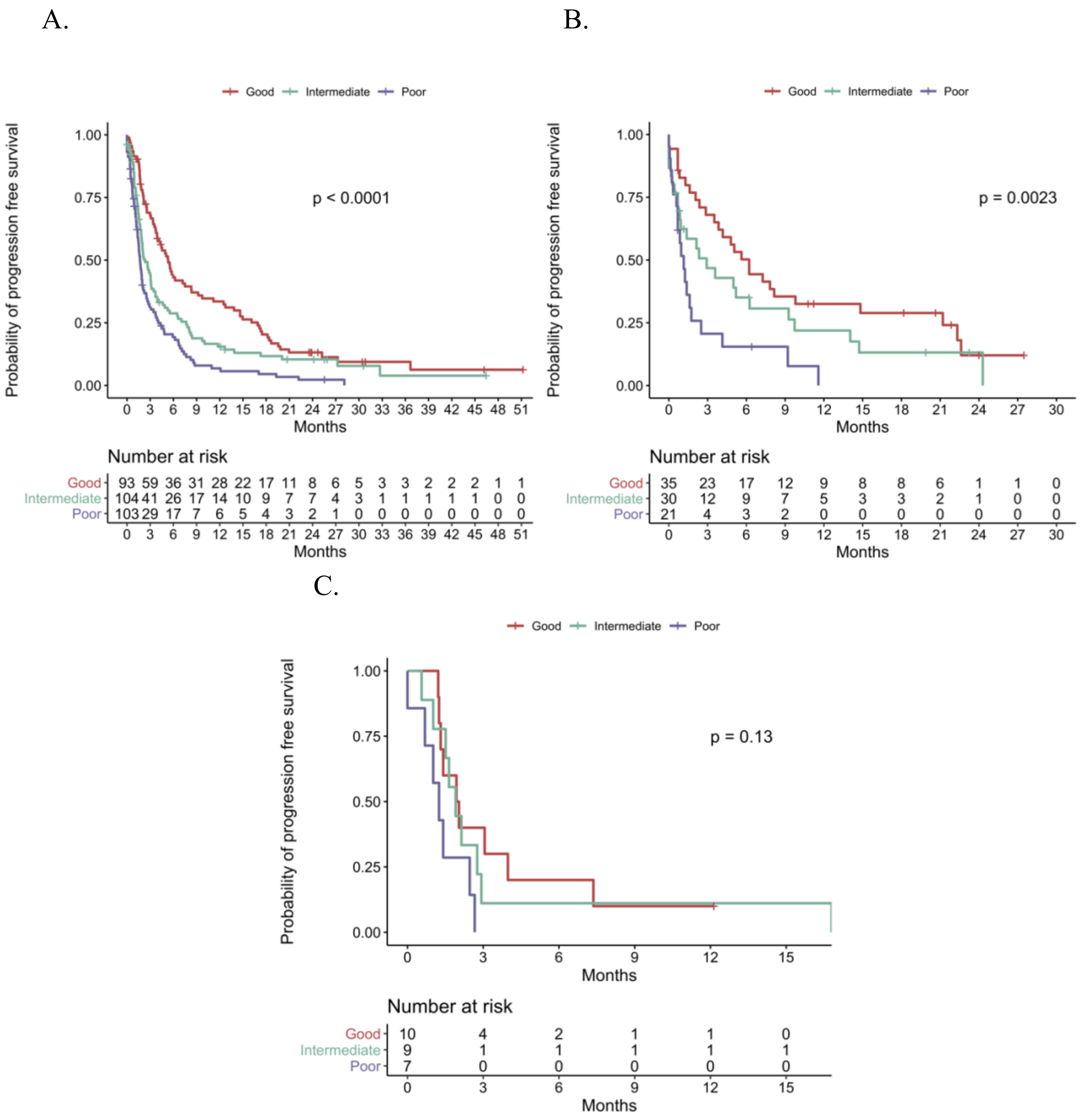


Figure 3.4 Subgroup analysis of PFS according to drug type: (A) Nivolumab, (B) Pembrolizumab, (C) Atezolizumab

Discussion

We proposed a prognostic score that combined PD-L1 expression, EGFR mutation status, ECOG performance status, and albumin level to stratify advanced NSCLC patients treated with PD-1/L1 inhibitors into poor, intermediate, and good groups. In this study, PFS was the primary endpoint, which was a direct measure of clinical benefit not subject to influence from post protocol therapy; OS was the secondary endpoint that demonstrated clinical benefits that were meaningful to patients.[22] Our prognostic score can stratify patients for both PFS and OS. A sensitivity analysis was performed to explore the impact of missingness of progression (adverse effect and move to hospice) and evaluate the robustness of results. Our study has emphasized the negative prognostics role of PD-L1 expression <1%, mutant EGFR, ECOG PS ≥ 2, and albumin< 3.5g/dl, which were consistent with findings in other studies and had high clinical relevance to progression and overall survival in advanced NSCLC patients.[14,23-25] Compared to the published prognostic score for patients received PD-1/L1 inhibitors, our prognostics score is the first one to include PD-L1 expression level and EGFR mutation status. The performance of the model was better than the model using PD-L1 alone.

Patients in this study were from real-world oncology practices, who did not necessarily meet numerous and restrictive eligibility criteria for clinical trials.[26] We included patients with older age, poor performance status, history of autoimmune disease, chronic steroid requirement, and symptomatic brain metastases.[27] Accordingly, the results of our study may be more applicable to select patients who should receive these treatments in real-world oncology practice. Variables were from routine medical care could be easily collected and integrated for clinical utility.

This study has several limitations. First, there were only 157 patients who had PD-L1 expression data available; this is because PD-L1 status testing is not mandatory for patients who were treated in a second line or above setting.[28] We treated these missing data as a separate category. As a result, this prognostic score could be applied to everyone treated with PD-1/L1 inhibitors. Second, not every patient's death was captured as patients may leave the healthcare system and loss to follow up, which inflated the estimates of median survival time. The median OS in our study is 18.4, which is longer than the median OS reported by other studies (10-12 months). Third, other markers such as tumor mutation burden, microsatellite instability status is warranted as they become available.[29,30] However, test of tumor mutation burden is still an expensive technique that is presently unavailable in wide clinical practice. Finally, data were retrospectively collected from MGH and BWH, prospective validation of our model another population would be desirable.

Conclusions

A prognostics score combining PD-L1 expression level, EGFR mutation status, ECOG performance status, and albumin level was associated with poor outcomes of PD-1/L1 inhibitors. This prognostic score may help to identify advanced NSCLC patients who likely benefit from PD-1/L1 inhibitors.

Reference:

1. Reck M, Rodriguez-Abreu D, Robinson AG, et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. N Engl J Med 2016;375:1823-33.

2. Herbst RS, Baas P, Kim DW, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet 2016;387:1540-50.

3. Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. N Engl J Med 2015;373:123-35.

4. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer. N Engl J Med 2015;373:1627-39.

5. Rebuzzi SE, Leonetti A, Tiseo M, Facchinetti F. Advances in the prediction of long-term effectiveness of immune checkpoint blockers for non-small-cell lung cancer. Immunotherapy 2019;11:993-1003.

6. Rittmeyer A, Barlesi F, Waterkamp D, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Lancet 2017;389:255-65.

7. Yan X, Zhang S, Deng Y, Wang P, Hou Q, Xu H. Prognostic Factors for Checkpoint Inhibitor Based Immunotherapy: An Update With New Evidences. Front Pharmacol 2018;9:1050.

8. Liu X, Guo CY, Tou FF, et al. Association of PD-L1 expression status with the efficacy of PD-1/PD-L1 inhibitors and overall survival in solid tumours: A systematic review and meta-analysis. Int J Cancer 2019.

9. Puzanov I, Diab A, Abdallah K, et al. Managing toxicities associated with immune checkpoint inhibitors: consensus recommendations from the Society for Immunotherapy of Cancer (SITC) Toxicity Management Working Group. J Immunother Cancer 2017;5:95.

10. Minami S, Ihara S, Ikuta S, Komuta K. Gustave Roussy Immune Score and Royal Marsden Hospital Prognostic Score Are Biomarkers of Immune-Checkpoint Inhibitor for Non-Small Cell Lung Cancer. World J Oncol 2019;10:90-100.

11. Mezquita L, Auclin E, Ferrara R, et al. Association of the Lung Immune Prognostic Index With Immune Checkpoint Inhibitor Outcomes in Patients With Advanced Non-Small Cell Lung Cancer. JAMA Oncol 2018;4:351-7.

12. Prelaj A, Ferrara R, Rebuzzi SE, et al. EPSILoN: A Prognostic Score for Immunotherapy in Advanced Non-Small-Cell Lung Cancer: A Validation Cohort. Cancers (Basel) 2019;11.

13. Shiroyama T, Suzuki H, Tamiya M, et al. Pretreatment advanced lung cancer inflammation index (ALI) for predicting early progression in nivolumab-treated patients with advanced non-small cell lung cancer. Cancer Med 2018;7:13-20.

14. Park W, Mezquita L, Okabe N, et al. Association of the prognostic model iSEND with PD-1/L1 monotherapy outcome in non-small-cell lung cancer. Br J Cancer 2020;122:340-7.

15. Liu J, Li S, Zhang S, et al. Systemic immune-inflammation index, neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio can predict clinical outcomes in patients with metastatic non-small-cell lung cancer treated with nivolumab. J Clin Lab Anal 2019;33:e22964.

16. Soo RA, Lim SM, Syn NL, et al. Immune checkpoint inhibitors in epidermal growth factor receptor mutant non-small cell lung cancer: Current controversies and future directions. Lung Cancer 2018;115:12-20.

17. Arbour KC, Mezquita L, Long N, et al. Impact of Baseline Steroids on Efficacy of Programmed Cell Death-1 and Programmed Death-Ligand 1 Blockade in Patients With Non-Small-Cell Lung Cancer. J Clin Oncol 2018;36:2872-8.

18. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics 2010;38:894-942.

19. Huang J, Breheny P, Ma S. A Selective Review of Group Selection in High-Dimensional Models. Statist Sci 2012;27:481-99.

20. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000;56:337-44.

21. Sridhara R, Mandrekar SJ, Dodd LE. Missing data and measurement variability in assessing progression-free survival endpoint in randomized clinical trials. AACR; 2013.

22. Korn RL, Crowley JJ. Overview: progression-free survival as an endpoint in clinical trials with solid tumors. AACR; 2013.

23. Patel SP, Kurzrock R. PD-L1 expression as a predictive biomarker in cancer immunotherapy. Molecular cancer therapeutics 2015;14:847-56.

24. Yu S, Liu D, Shen B, Shi M, Feng J. Immunotherapy strategy of EGFR mutant lung cancer. Am J Cancer Res 2018;8:2106-15.

25. Ibrahimi S, Mukherjee S, Roman D, King C, Machiorlatti M, Aljumaily R. Effect of body mass index and albumin level on outcomes of patients receiving anti PD-1/PD-L1 therapy. Journal of Clinical Oncology 2018;36:213-.

26. Rashdan S, Gerber DE. Immunotherapy for non-small cell lung cancer: from clinical trials to real-world practice. Translational Lung Cancer Research 2018;8:202-7.

27. Rashdan S, Gerber DE. Immunotherapy for non-small cell lung cancer: from clinical trials to real-world practice. Transl Lung Cancer Res 2019;8:202-7.

28. Ancevski Hunter K, Socinski MA, Villaruz LC. PD-L1 Testing in Guiding Patient Selection for PD-1/PD-L1 Inhibitor Therapy in Lung Cancer. Mol Diagn Ther 2018;22:1-10.

29. Chan TA, Yarchoan M, Jaffee E, et al. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. Ann Oncol 2019;30:44-56.

30. Chang L, Chang M, Chang HM, Chang F. Microsatellite Instability: A Predictive Biomarker for Cancer Immunotherapy. Appl Immunohistochem Mol Morphol 2018;26:e15-e21.

Supplementary materials

Supplemental Table 1.1 Discrepancies between EMR diagnosis date and random samples/BLCS diagnosis date

Absolute discrepancy	Random samples from LC mart (chart review)	Boston Lung Cancer Study
> 90 days	10.4% (7/67)	12.4%
>180 days	9.0% (6/67)	8.8%
> One year	7.5% (5/67)	6.4%

Supplemental Table 1.2 Discrepancies between first ICD time and random samples/BLCS diagnosis date

Absolute discrepancy	Random samples from LC mart (chart review)	Boston Lung Cancer Study
> 90 days	13.4% (9/67)	12.5%
>180 days	11.9% (8/67)	8.9%
> One year	9.0% (6/67)	6.4%

Supplemental Table 1.3 Comparison of histological type curated from BLCS cohort versus histological type extracted from EMR

	BLCS			
EMR	Adeno	Squamous	NSCLC unspecified	Small cell
Adeno	3386	64	185	3
Squamous	44	951	46	0
NSCLC unspecified	95	41	296	2
Small cell	8	16	12	477
Total	3533	1072	539	482
Accuracy	0.96	0.89	0.55	0.99

Supplemental Table 1.4 Comparison of histological type curated from chart review versus histological type extracted from EMR

	Chart Review			
EMR	Adeno	Squamous	NSCLC unspecified	Small cell
Adeno	38	0	0	0
Squamous	0	8	0	0
NSCLC unspecified	2	1	4	0
Small cell	0	0	0	8
Total	40	9	4	8
Accuracy	0.95	0.89	1.00	1.00

Note: Comparison performed using 67 patients using chart review and EMR extracted histological type

Supplemental Table 1.5 Comparison of stage curated from BLCS cohort versus stage extracted from EMR

	BLCS					
EMR	Stage1	Stage 2	Stage 3	Stage 4	Extensive	Limited
Stage 1	1287	89	37	49	0	1
Stage 2	160	308	24	19	0	0
Stage 3	84	47	955	63	1	0
Stage 4	145	45	94	1302	1	0
Extensive	2	0	2	12	207	11
Limited	25	9	15	8	14	174
Total	1703	498	1127	1453	223	186
Accuracy	0.76	0.62	0.85	0.90	0.93	0.94

Note: Comparison performed using 5190 patients with histology type from both BLCS and EMR data

Supplemental Table 1.6 Comparison of stage from chart review versus stage extracted from EMR

	Chart Review					
EMR	Stage1	Stage 2	Stage 3	Stage 4	Extensive	Limited
Stage 1	9	1	0	0	0	0
Stage 2	3	6	1	1	0	0
Stage 3	1	0	13	1	0	0
Stage 4	1	1	0	14	0	0
Extensive	0	0	0	0	2	1
Limited	0	0	0	0	0	2
Total	14	8	14	16	2	3
Accuracy	0.64	0.75	0.93	0.88	1.00	0.67

Note: Comparison performed using 57 patients with histology type from both BLCS and EMR data

Supplemental Table 2.1 Values of laboratory variables in the training and testing sets.

Patient characteristic		Training set N=8793 (%)	Testing set N=2931 (%)	p-value
Albumin	<=3.5 g/dl	529 (18.0)	1550 (17.6)	0.874
	>3.5 g/dl	1907 (65.1)	5754 (65.4)	
	Missing	495 (16.9)	1489 (16.9)	
ALKP	<=140 IU/L	2382 (81.3)	7106 (80.8)	0.102
	>140 IU/L	193 (6.6)	677 (7.7)	
	Missing	356 (12.1)	1010 (11.5)	
ALT	<=7 IU/L	96 (3.3)	309 (3.5)	0.824
	>=56 IU/L	127 (4.3)	384 (4.4)	
	7-56 IU/L	2340 (79.8)	7043 (80.1)	
	Missing	368 (12.6)	1057 (12.0)	
AST	<=10 IU/L	51 (1.7)	142 (1.6)	0.889
	>=40 IU/L	258 (8.8)	748 (8.5)	
	10-40 IU/L	2251 (76.8)	6809 (77.4)	
	Missing	371 (12.7)	1094 (12.4)	
BUN	<=7 mg/dL	116 (4.0)	335 (3.8)	0.886
	>=20 mg/dL	774 (26.4)	2307 (26.2)	
	7-20 mg/dL	1942 (66.3)	5829 (66.3)	
	Missing	99 (3.4)	322 (3.7)	
Calcium	<=8.5 mg/dl	466 (15.9)	1413 (16.1)	0.204
	>=10.5 mg/dl	68 (2.3)	270 (3.1)	
	8.5-10.5 mg/dl	2229 (76.0)	6604 (75.1)	
	Missing	168 (5.7)	506 (5.8)	
Chloride	<=96 mEq/L	247 (8.4)	850 (9.7)	0.192
	>=106 mEq/L	428 (14.6)	1322 (15.0)	
	96-106 mEq/L	2125 (72.5)	6229 (70.8)	
	Missing	131 (4.5)	392 (4.5)	
Creatinine	above	504 (17.2)	1477 (16.8)	0.434
	normal	2219 (75.7)	6705 (76.3)	
	under	119 (4.1)	312 (3.5)	
	Missing	89 (3.0)	299 (3.4)	
Glucose	<=100 mg/dL	1100 (37.5)	3286 (37.4)	0.984
	>=125 mg/dL	903 (30.8)	2730 (31.0)	
	100-125 mg/dL	779 (26.6)	2343 (26.6)	
	Missing	149 (5.1)	434 (4.9)	
HBG	above	37 (1.3)	119 (1.4)	0.986
	normal	1649 (56.3)	4949 (56.3)	
	under	1149 (39.2)	3437 (39.1)	
	Missing	96 (3.3)	288 (3.3)	

Supplemental Table 2.1 (Continued)

MCH	<=27 pg/cell	249 (8.5)	775 (8.8)	0.819
	>=33 pg/cell	234 (8.0)	660 (7.5)	
	27-33 pg/cell	2352 (80.2)	7067 (80.4)	
	Missing	96 (3.3)	291 (3.3)	
MCHC	<=31 g/dL	60 (2.0)	169 (1.9)	0.949
	>=37 g/dL	7 (0.2)	25 (0.3)	
	31-37 g/dL	2769 (94.5)	8310 (94.5)	
	Missing	95 (3.2)	289 (3.3)	
MCV	<=80 femtoliters/cell	159 (5.4)	490 (5.6)	0.986
	>=96 femtoliters/cell	317 (10.8)	937 (10.7)	
	80-96 femtoliters/cell	2359 (80.5)	7076 (80.5)	
	Missing	96 (3.3)	290 (3.3)	
PLT	<=150000/ml	129 (4.4)	392 (4.5)	0.68
	>=450000/ml	286 (9.8)	797 (9.1)	
	150000-450000/ml	2421 (82.6)	7300 (83.0)	
	Missing	95 (3.2)	304 (3.5)	
Potassium	<=3.5 mEq/L	274 (9.3)	902 (10.3)	0.181
	>=5.0 mEq/L	97 (3.3)	331 (3.8)	
	3.5-5.0 mEq/L	2406 (82.1)	7056 (80.2)	
	Missing	154 (5.3)	504 (5.7)	
RBC	normal	1274 (43.5)	3749 (42.6)	0.717
	under	1560 (53.2)	4756 (54.1)	
	MISSING	97 (3.3)	288 (3.3)	
RDW	<=14.5%	2183 (74.5)	6508 (74.0)	0.88
	>14.5%	651 (22.2)	1986 (22.6)	
	Missing	97 (3.3)	299 (3.4)	
Sodium	<=135 mEq/L	443 (15.1)	1376 (15.6)	0.817
	>=145 mEq/L	66 (2.3)	211 (2.4)	
	135-145 mEq/L	2296 (78.3)	6814 (77.5)	
	Missing	126 (4.3)	392 (4.5)	
Bilirubin	<=0.2 mg/dL	332 (11.3)	990 (11.3)	0.625
	>1.2 mg/dL	49 (1.7)	177 (2.0)	
	0.2-1.2 mg/dL	2195 (74.9)	6602 (75.1)	
	Missing	355 (12.1)	1024 (11.6)	

Supplemental Table 2.1 (Continued)

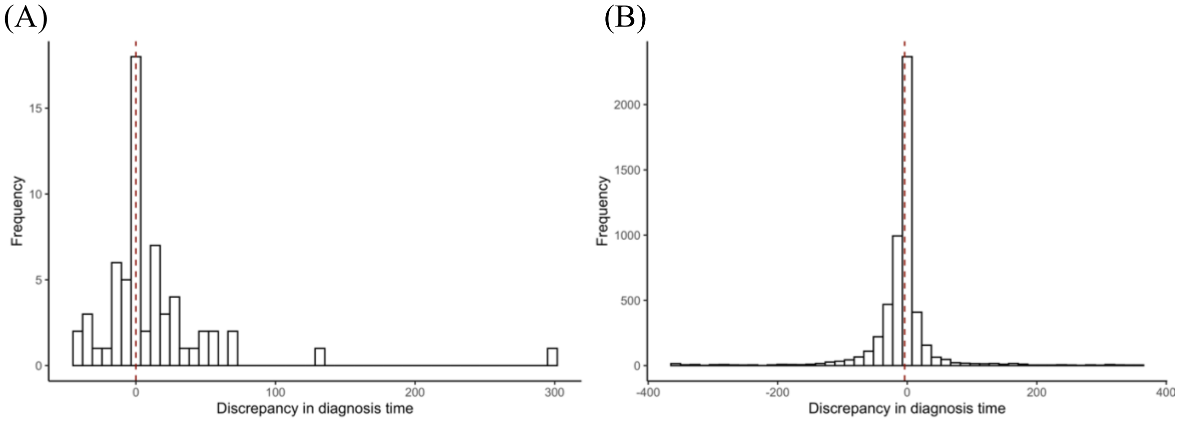
WBC	>=11*10^9/L	2139 (73.0)	6350 (72.2)	0.678
	4.5-11*10^9/L	697 (23.8)	2162 (24.6)	
	Missing	95 (3.2)	281 (3.2)	
Lymphocytes	<=20%	1403 (47.9)	4081 (46.4)	0.545
	>=40%	74 (2.5)	228 (2.6)	
	20-40%	997 (34.0)	3107 (35.3)	
	Missing	457 (15.6)	1377 (15.7)	
Neutrophils	<=40%	22 (0.8)	84 (1.0)	0.449
	>=60%	2113 (72.1)	6252 (71.1)	
	40-60%	336 (11.5)	1083 (12.3)	
	Missing	460 (15.7)	1374 (15.6)	
Monocytes	<=2%	106 (3.6)	351 (4.0)	0.802
	>=8%	513 (17.5)	1560 (17.7)	
	2-8%	1853 (63.2)	5505 (62.6)	
	Missing	459 (15.7)	1377 (15.7)	
Eosinophils	<=1%	945 (32.2)	2738 (31.1)	0.689
	>=4%	403 (13.7)	1209 (13.7)	
	1-4%	1123 (38.3)	3461 (39.4)	
	Missing	460 (15.7)	1385 (15.8)	
Basophils	<=0.5%	1559 (53.2)	4734 (53.8)	0.88
	>=1%	520 (17.7)	1505 (17.1)	
	0.5-1%	385 (13.1)	1153 (13.1)	
	Missing	467 (15.9)	1401 (15.9)	
NLR	<=4	1294 (44.1)	3968 (45.1)	0.602
	>4	1177 (40.2)	3444 (39.2)	
	Missing	460 (15.7)	1381 (15.7)	

Supplemental Table 3.1 Distributions of laboratory variables

		Overall
Lab Test	Level	412
Albumin (%)	<=3.5g/dl	127 (30.8)
	>3.5g/dl	194 (47.1)
	Untested	91 (22.1)
ALKP (%)	<=140iu/l	260 (63.1)
	>140iu/l	61 (14.8)
	Untested	91 (22.1)
ALT (%)	<=7iu/l	26 (6.3)
	>=56iu/l	15 (3.6)
	7-56iu/l	280 (68.0)
	Untested	91 (22.1)
AST (%)	<=10iu/l	11 (2.7)
	>=40iu/l	30 (7.3)
	10-40iu/l	280 (68.0)
	Untested	91 (22.1)
BUN (%)	<=7mg/dl	20 (4.9)
	>=20mg/dl	92 (22.3)
	7-20mg/dl	215 (52.2)
	Untested	85 (20.6)
Calcium (%)	<=8.5mg/dl	39 (9.5)
	>=10.5mg/dl	7 (1.7)
	8.5-10.5mg/dl	281 (68.2)
	Untested	85 (20.6)
Chloride (%)	<=96meq/L	76 (18.4)
	>=106meq/L	9 (2.2)
	96-106meq/L	242 (58.7)
	Untested	85 (20.6)
Creatinine (%)	Above	55 (13.3)
	Normal	238 (57.8)
	Under	34 (8.3)
	Untested	85 (20.6)
Glucose (%)	<=100mg/dl	96 (23.3)
	>=125mg/dl	109 (26.5)
	100-125mg/dl	122 (29.6)
	Untested	85 (20.6)
HGB (%)	Above	1 (0.2)
	Normal	69 (16.7)
	Under	256 (62.1)
	Untested	86 (20.9)
MCH (%)	<=27pg/cell	50 (12.1)
	>=33pg/cell	35 (8.5)
	27-33pg/cell	241 (58.5)
	Untested	86 (20.9)

Supplemental Table 3.1 (Continued)

MCHC (%)	<=31g/dl	32 (7.8)
	31-37g/dl	294 (71.4)
	Untested	86 (20.9)
MCV (%)	<=80femtoliters/cell	12 (2.9)
	>=96femtoliters/cell	71 (17.2)
	80-96femtoliters/cell	243 (59.0)
	Untested	86 (20.9)
PLT (%)	<=150000/ml	36 (8.7)
	>=450000/ml	32 (7.8)
	150000-450000/ml	257 (62.4)
	Untested	87 (21.1)
Potassium (%)	<=3.5meq/L	25 (6.1)
	>=5.0meq/L	12 (2.9)
	3.5-5.0meq/L	286 (69.4)
	Untested	89 (21.6)
RBC (%)	Normal	51 (12.4)
	Under	275 (66.7)
	Untested	86 (20.9)
RDW (%)	<=14.5%	114 (27.7)
	>14.5%	212 (51.5)
	Untested	86 (20.9)
Sodium (%)	<=135meq/L	82 (19.9)
	>=145meq/L	2 (0.5)
	135-145meq/L	243 (59.0)
	Untested	85 (20.6)
Bilirubin (%)	>0.4mg/dl	94 (22.8)
	0-0.4mg/dl	227 (55.1)
	Untested	91 (22.1)
Protein (%)	<=6g/dl	38 (9.2)
	>=8.3g/dl	4 (1.0)
	6-8.3g/dl	281 (68.2)
	Untested	89 (21.6)
WBC (%)	>=11*10^9/l	78 (18.9)
	4.5-11*10^9/l	248 (60.2)
	Untested	86 (20.9)
Lymphocytes (%)	<=20%	246 (59.7)
	>=40%	3 (0.7)
	20-40%	75 (18.2)
	Untested	88 (21.4)
Neutrophils (%)	<=40%	2 (0.5)
	>=60%	283 (68.7)
	40-60%	39 (9.5)
	Untested	88 (21.4)
NLR (%)	<=5	147 (35.7)
	>5	176 (42.7)
	Untested	89 (21.6)



Supplemental Figure 1.1 Histogram of date discrepancies for extracted diagnosis date from EMR, as compared to chart review(A)/BLCS cohort (B) diagnosis date. Positive deflections represent an extracted date that is later than the chart review/BLCS date. Outliers beyond +/- 365 days (one year) are not shown.