FPN: 270P

Machine Learning-Based Prediction of Survival Prognosis for Treated NSCLC Patients: Insights from Finnish Real-World Data

Ville Koistinen*[1], Lisse-Lotte Hermansson[2,4], Heikki Ekroos[3], Sanaz Jamalzadeh[4], Riikka Mattila[4], Olivia Hölsä[4], Aija Knuuttila[3]

[1]Wellbeing services county of Kymenlaakso, Finland [2]Turku University, Faculty of Medicine, Finland, [3]Helsinki University Hospital, Finland, [4]Medaffcon Oy, Finland

**Conflict of interest: Congress travel costs: MSD, Lecture fee: AstraZeneca*

Background: Despite advancements in treatment, non-small cell lung cancer (NSCLC) remains a leading cause of cancer-related mortality worldwide. Understanding the factors that influence survival outcomes is crucial for optimizing treatment strategies and improving patient prognoses.

Methods: We included 1,085 adult NSCLC patients treated at HUS, Helsinki University Hospital between 2018 and 2023. Survival time was calculated as the number of months from the initiation of first-line treatment until death or the end of the study. Smoking status was extracted from the patient texts using a natural language processing approach. A Random Survival Forest model was trained on the data following hyperparameter tuning, and its performance was evaluated using the area under the curve (AUC) at six specific follow-up time points. Variable importance with 95% confidence interval was extracted from the model to rank predictive features and highlight how the importance varies over time. Overall survival was computed by fitting Kaplan Meir curves for the overall treated cohort, stratified by the covariates included in the prediction model.

Results: We developed a machine learning-based survival prediction model using data from 443 treated NSCLC patients with complete data coverage, eliminating the need for data imputation. From this group, 311 patients were used to train the model, and 132 patients were used to evaluate it. The model demonstrated robust performance at 6 months, 24 months, and 30 months of follow-up, achieving an AUC of 79%, 80%, and 83%, respectively. The integrated AUC over six timepoints was 72%. Our analysis highlighted several key factors influencing survival outcomes, including ECOG performance status, C-reactive protein (CRP) levels, blood neutrophil counts, and type of systemic treatment. To address the selection bias from using complete data, we stratified the survival by the included covariates in the full cohort of 1,085 treated patients, confirming their significance also outside the patients with complete data.

Conclusions: We developed a robust machine learning-based survival prediction model using high-quality real-world clinical data. Our findings demonstrate that the model achieves excellent predictive performance at specific timepoints during follow-up as well as identifies the key factors contributing to the prognoses in NSCLC patients treated with pharmacological lung cancer therapies.

DATA SOURCE

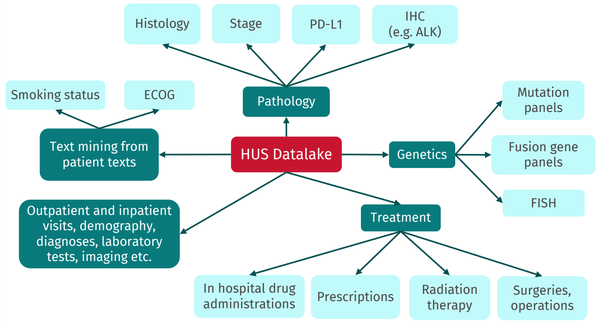

Figure 1. *Integration of raw data from 28 EHR data systems to derive study variables for cohort analysis, utilizing access to deep, comprehensive and granular clinical data. This extensive integration allows for the analysis of critical NSCLC factors such as pathology, genetics, and treatment, enhancing the reliability and applicability of the study findings.*

COHORT

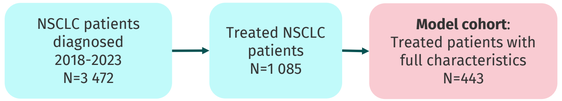

Figure 2. *Cohort formation.*

- **Initial Cohort**: Included all patients diagnosed with NSCLC from 2018 to 2023.
- **Treated Cohort**: Comprised all patients treated with pharmacological lung cancer therapies to assess the results.
- **Prediction Model Cohort**: Included patients receiving pharmacological lung cancer treatment with complete characteristics.
  - **Model Training**: Utilized 70% of the cohort for training the prediction model.
  - **Model Testing**: Used 30% of the cohort for evaluating the performance of the model.

PREDICTION MODEL

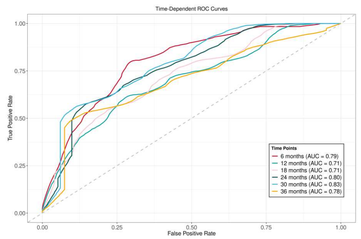

Figure 3. *Model performance.*

- Random forest survival model predicts survival at various time points
- ROC curves measure how well the model can distinguish between the individuals who have died by the time point vs those who are alive at the time point.
- AUC summarizes the area under the ROC curve
- Integrated AUC ~ 0.72

CHARACTERISTICS

Table 1. *Patient characteristics.*

		Model cohort	Whole treated cohort*	
Variable		**N (%)**	**N (%)**	**Missing N (%)**
mNSCLC	M0	180 (41)	477 (44)	0 (0)
	M1	263 (59)	608 (56)	
ECOG	0	82 (19)	150 (20)	330 (30)
	1	241 (54)	403 (53)	
	2	103 (23)	173 (23)	
	3	17 (4)	29 (4)	
CCI	0	176 (40)	448 (41)	0 (0)
	1	158 (36)	381 (35)	
	2	73 (16)	170 (16)	
	3+	36 (8)	86 (8)	
CRP	<10 mg/L	161 (36)	416 (39)	13 (1)
	≥10 mg/L	282 (64)	656 (61)	
Eosinophiles	<0.4 10^9/L	353 (80)	678 (82)	255 (24)
	≥0.4 10^9/L	90 (20)	152 (18)	
Histology	Adenocarcinoma	297 (67)	689 (64)	0 (0)
	Other/unknown	50 (11)	173 (16)	
	SqC	96 (22)	223 (21)	
Neutrophiles	<7 10^9/L	297 (67)	713 (68)	43 (4)
	≥7 10^9/L	146 (33)	329 (32)	
Trombocytes	<360 10^9/L	270 (61)	664 (61)	3 (0)
	≥360 10^9/L	173 (39)	418 (39)	
Age at index	<65 years	155 (35)	352 (32)	0 (0)
	≥65 years	288 (65)	733 (68)	
PD-L1 status	<1%	152 (34)	283 (37)	311 (29)
	1-49%	144 (33)	260 (34)	
	50-100%	147 (33)	231 (30)	
Resectable status	Resectable	54 (12)	147 (14)	0 (0)
	Unresectable	389 (88)	938 (86)	
Sex	Female	213 (48)	512 (47)	0
	Male	230 (52)	573 (53)	
Smoking status	Smoker	148 (33)	340 (32)	19 (2)
	Ex-smoker	238 (54)	573 (54)	
	Never smoker	57 (13)	153 (14)	
Stage	Stage I	10 (2)	42 (4)	0 (0)
	Stage I-III	55 (12)	165 (15)	
	Stage II	29 (7)	65 (6)	
	Stage III	86 (19)	233 (21)	
	Stage IV	263 (59)	580 (53)	
Treatment group	Chemotherapy	205 (46)	588 (54)	0 (0)
	IO (±chemo)	187 (42)	318 (29)	
	TKI(±chemo)	51 (12)	179 (16)	

- Patient characteristics used in the model are presented in **Table 1.**
- Only patients with complete data coverage were included in the model cohort, resulting in 0% missing data.

mNSCLC: first record of metastasis at treatment start, CCI: Charlson's comorbidity index, CRP: C-reactive protein

VARIABLE IMPORTANCE

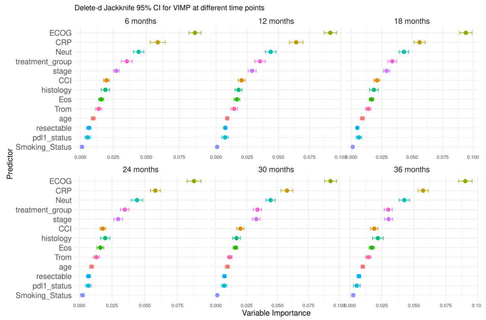

Figure 4. *Variable importance in predicting OS in different time points*

- **Key Predictors**: The most important factors influencing NSCLC survival are ECOG, C-reactive protein (CRP), blood neutrophil counts, and treatment group (Figure 4).

OVERALL SURVIVAL

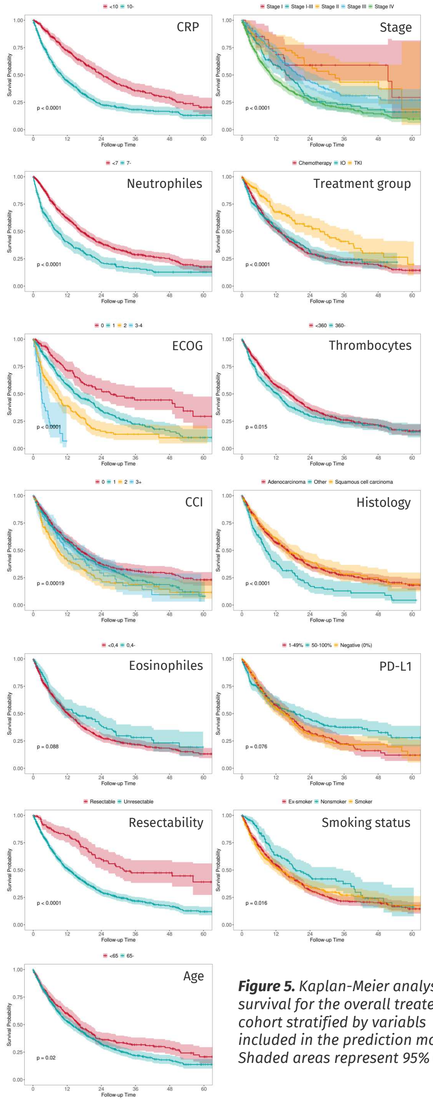

Figure 5. *Kaplan-Meier analysis of survival for the overall treated cohort stratified by variabls included in the prediction model. Shaded areas represent 95% CI.*

- All variables show a significant impact on survival in this population, except eosinophils and PD-L1 status (Figure 5).

Contact: riikka.mattila@medaffcon.fi