Real-World Evidence Study of Patients With NSCLC in Finland: Use of Machine Learning Algorithm to Extract Smoking Status From Patient Texts and Analysis of Resource Use and Survival by Smoking Status

Ekroos H[1], Koistinen V[2], Hölsä O[3], Mattila R[3], Knuuttila A[4]

[1]HUS Porvoo Hospital, Porvoo, Finland, [2]Wellbeing services county of Kymenlaakso, Kotka, Finland, [3]Medaffcon Oy, Espoo, Finland, [4]Helsinki University Hospital, Helsinki, Finland

Background

Lung cancer is the second most common cancer in men and fourth most common cancer in women in Finland [1]. Like in many other countries, lung cancer mortality rates in Finland reflect the past smoking habits [2]. While smoking is known to affect the biology of the disease and is a known risk factor shortening the survival in non-small cell lung cancer (NSCLC), it is commonly registered only as unstructured data in medical records, significantly limiting the usability of health data. This study leverages machine learning to extract smoking status from unstructured data, aiming to improve data usability and informed clinical and policy decision-making in NSCLC.

Objectives

- Identify the smoking status of NSCLC patients using a previously developed machine learning algorithm [3]
- Analyze the overall survival and
- Analyze the healthcare resource utilization (HCRU) of NSCLC patients

Methods

- All data was extracted from Helsinki University Hospital (HUS) datalake (HUS 56/2023)
- Adult patients with a diagnosis of NSCLC between January 2013 and August 2023 were included
- Detection of patients with ICD-10: C34.x0-C34.x5 and C34.x9, M-SNOMED codes M81403, M80703, M80123, M85603, or M80103 in lung, pleura, or bronchus, or M80706, M80106, M81406, or M80003 or in any organ for those with no other in any organ
- Collection of exhaustive data at specialized care (Figure 1)
  - The collected raw data was processed to form study variables
- Patients were followed from first diagnosis until death or end of follow-up (31 August 2023)

Smoking status classifier: Previously developed machine learning algorithm was used to identify smoking status of included patients from patient texts (Figure 2) [3].

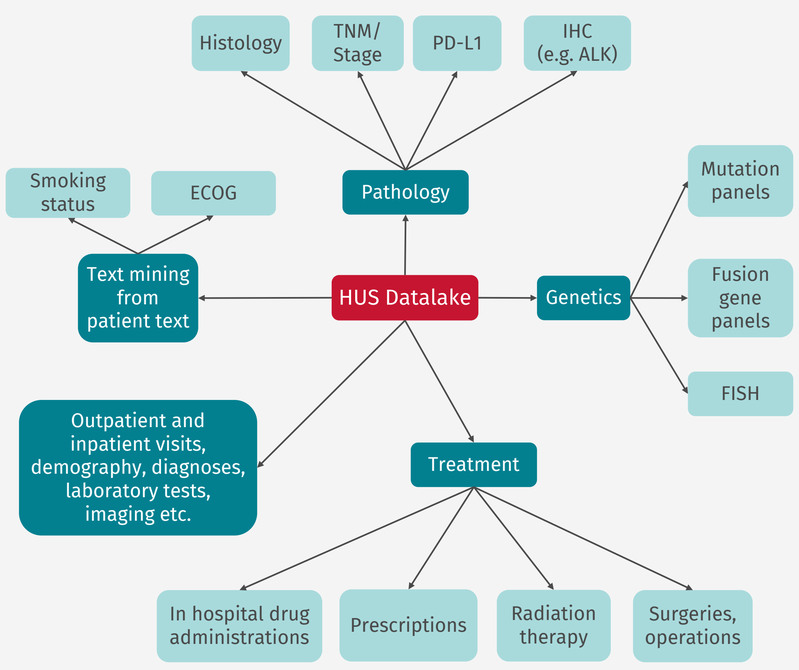

Figure 1. Integration of raw data from 28 EHR data systems to derive study variables for cohort analysis, utilizing access to deep, comprehensive and granular clinical data. This extensive integration allows for the analysis of critical NSCLC factors such as pathology, genetics, and treatment, enhancing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

Table 1. Patient characteristics at index by smoking status

Variable		Smoker	Ex-smoker	Never-smoker	Missing (%)	p-value
N		2320	2759	715		
Age, years, median (IQR)		69 (63-75)	73 (68-79)	73 (66-80)	0	<0.001
Sex, N(%)	Female	1014 (44)	1064 (39)	506 (71)	0	<0.001
	Male	1306 (56)	1695 (61)	209 (29)		
Resectable status	Resectable	500 (22)	644 (23)	204 (29)	0	<0.001
	Unresectable	1820 (78)	2115 (77)	511 (71)		
Histology	Adeno-carcinoma	1067 (46)	1331 (48)	486 (68)	0	<0.001
	Other NSCLC	738 (32)	873 (32)	190 (27)		
	Squamous cell carcinoma	515 (22)	555 (20)	39 (5)		
PD-L1 status	1-49%	243 (33)	270 (30)	76 (30)	67	0.002
	50-100%	201 (27)	209 (23)	46 (18)		
	Negative (0%)	291 (40)	423 (47)	132 (52)		
Metastatic	De novo metastatic	1198 (52)	1347 (49)	368 (51)	0	0.003
	Recurrent metastatic	386 (17)	448 (16)	85 (12)		
	No metastasis	736 (32)	964 (35)	262 (37)		
ECOG performance status	0	148 (16)	175 (17)	63 (23)	61	0.003
	1	354 (38)	417 (39)	126 (46)		
	2	249 (27)	284 (27)	54 (20)		
	3-4	176 (19)	184 (17)	33 (12)		
CCI (Charlson comorbidity index)	0	843 (36)	907 (33)	353 (49)	0	<0.001
	1	770 (33)	920 (33)	244 (34)		
	2	409 (18)	526 (19)	84 (12)		
	3	181 (8)	255 (9)	22 (3)		
	4+	117 (5)	151 (5)	12 (2)		
Length of follow-up, months, median (IQR)		23 (7-68)	27 (8-75)	40 (13-98)	0	<0.001

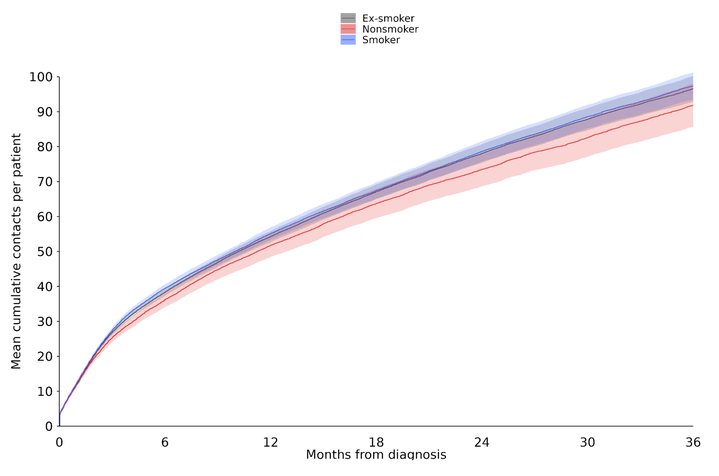

Figure 3. Mean cumulative health care contacts per patients stratified by smoking status.

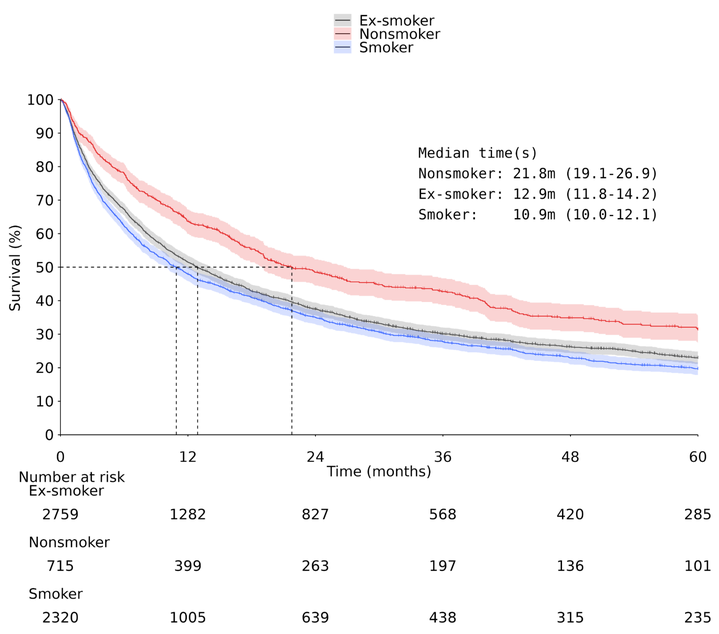

Figure 4. Overall survival from index stratified by smoking status.

Results

Smoking status recognizable from patient text

- Total patients with NSCLC: 6 248
- Reported smoking status: 93%
  - 2320 (40%) smokers
  - 2759 (48%) ex-smokers
  - 715 (12%) never-smokers
- Smokers were younger, more likely to have metastatic disease, and had more comorbidities (Table 1)

Similar HCRU in patients with different smoking status

- Patients in all smoking-status groups had near equal number of specialized healthcare contacts during the first-year follow-up (Figure 3)
  - Smokers: 55 (CI 95% 53-57)
  - Ex-smokers: 54 (CI 95% 53-56)
  - Never-smokers: 52 (CI 95% 48-55)
- Corresponding costs for first-year follow-up were similar regardless of smoking status
  - Smokers: 26 221€ (CI 95% 25 320-27 121)
  - Ex-smokers: 25 858€ (CI 95% 25 028-26 687)
  - Never-smokers: 25 189€ (CI95% 23 241-27 136)

Smoking Significantly Reduces Overall Survival (Figure 4)

- Smokers : 10.8 months (CI95% 10.0-12.1)
  - Squamous cell histology: 12.5 months (95% CI: 10.3-16.8)
  - Non-squamous cell histology: 10.3 months (95% CI: 9.0-11.7)
- Ex-smokers: 12.9 months (CI95% 11.8-14.2)
  - Squamous cell histology: 14.0 months (95% CI: 11.8-16.4)
  - Non-squamous cell histology: 12.6 months (95% CI: 11.3-14.1)
- Never-smokers: 21.8 months (CI95% 19.1-26.9)

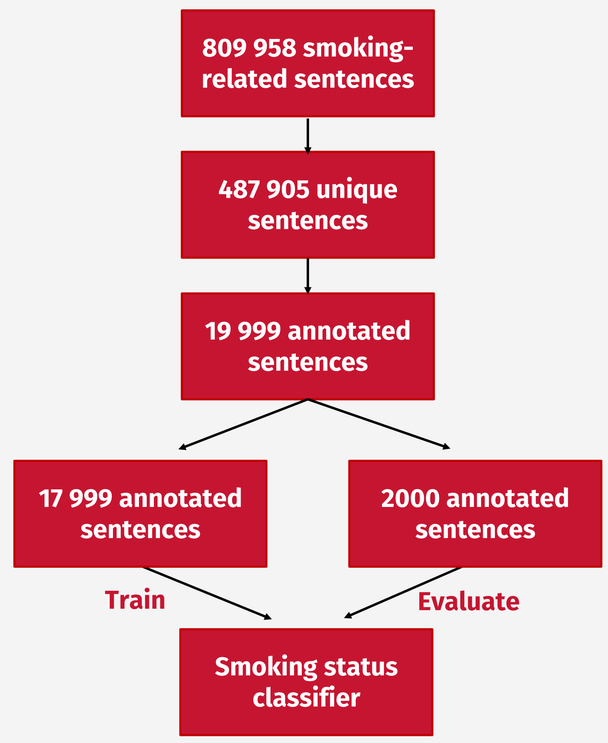

Figure 2. Different phases of the development in the smoking status NLP-model.

Conclusions

- Smoking status is comprehensively available in patients' medical records and can be reliably extracted with a machine learning –based algorithm, demonstrating the potential for automated data extraction to enhance research efficiency.
- Text mining unstructured medical data to complement patients' health records can enable the use of new research variables in epidemiological analyses.
- Smoking status did not affect the overall first-year follow-up costs in NSCLC patients
- Expectedly, smoking had a negative effect to the overall survival of NSCLC patients

References

[1] Lung cancer. Current Care Guidelines. The Finnish Medical Society Duodecim, 2017

[2] Pitkäniemi J, Malila N, Virtanen A, Degerlund H, Heikkinen S, Seppä K. Syöpä 2018. Suomen Syöpäyhdistyksen julkaisuja nro 93. Suomen Syöpäyhdistys, Helsinki 2020.

[3] Gräsbeck HL, Reito ARP, Ekroos HJ, Aakko, JA, Hölsä O, Vasankari TM. BJS Open, 2023, 7(2).