Utilizing OMOP-Harmonised Automatically Extracted Hospital Network-Wide EHR Data to Detect Association Between Obesity and Sex-related Survival Difference in Lung Cancer

Alexey Ryzhenkov[1], Salma Rachidi[1], Valtteri Nieminen[1,2,5], Johanna Sanoja[1], Marianna Niemi[2], Juuso Paajanen[3], Johanna Niklander[2], Paula Kauppi[3], Aija Knuuttila[1,3], Ilkka Ilonen[1,3,4], Eric Fey[1,2*], Kimmo Porkka[1*]

* shared authorship

1. iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki, and Helsinki University Hospital Comprehensive Cancer Center, Departments of Oncology and Hematology, Helsinki, Finland
2. HUS IT Management, Helsinki University Hospital, Helsinki, Finland
3. Heart and Lung Center, Pulmonary Department, Helsinki University Hospital, Helsinki, Finland
4. Department of Thoracic Surgery, Heart and Lung Center, Helsinki University Hospital and University of Helsinki
5. Department of Computing, University of Turku, Finland

Pre-Publication Corresponding Author: Alexey Ryzhenkov, alexey.ryzhenkov@helsinki.fi

Post-Publication Corresponding author:
prof. Kimmo Porkka
Helsinki University Hospital Comprehensive Cancer Center
Haartmaninkatu 4, 00290 Helsinki, Finland.
Tel. +358-50-427-0192
Email: kimmo.porkka@helsinki.fi

Keywords: sex difference, obesity
Shorted header: OMOP-Harmonized EHR Data: Obesity, Sex, and Lung Cancer Survival

Abstract

Purpose

Survival discrepancy between male and female patients is a recurring highlight in treatment efficacy evaluation reviews. Previous studies have used different patient cohorts and clinical covariates and have not included obesity, which is associated with longer lung cancer survival. This study aims to evaluate the effect of obesity on survival differences while including all histologies and therapy types in the study utilizing a unique hospital-wide OMOP-harmonized patient cohort.

Methods

A retrospective, real-world cohort of 5,598 patients diagnosed and treated in a university hospital network from 2015 to 2024 was used in the study. All clinical data were harmonized to the OMOP[1] common data model for interoperability and scalability. Patients were stratified by body mass index (BMI) and univariate and multivariate analyses of survival were performed using treatment- and tumor-related covariates.

Results

Higher BMI was associated with smaller sex-related survival differences. In the normal BMI cohort, the two-year overall survival was 43% in females and 24% in males ($p<0.01$). In the high BMI cohort the difference was 47% vs 38% ($p<0.01$). When stratified by histology, the largest effect of high BMI was observed in squamous cell carcinoma. When full multivariate analysis was performed separately for high and normal BMI patients, the effect of male sex on survival was 42% smaller among high BMI patients. Introduction of covariates reduced but did not eliminate the effect compared to univariate analysis.

Conclusions

Higher BMI was associated with reduced survival gap between sexes underlying the need of BMI reporting in future clinical trials, RWE studies and national records.

[1] Observational Medical Outcomes Partnership

1 Introduction

Lung cancer is causing 19% of all cancer deaths in the world with a yearly incidence of 2,5 million cases worldwide [1]. Better survival outcomes in female over male lung cancer patients have been noted already in the end of the 1990s [2]**.** Multiple studies have reported that controlling for clinical covariates such as stage, treatment and age does not remove the association between sex and survival, concluding that sex is an independent prognostic factor in lung cancer [3–6]. However, some studies were not able to support these results stating that controlling for covariates eliminates or significantly reduces the survival difference between male and female patients [7,8,9].
It was further reported that the presence and magnitude of the survival difference between male and female patients depends on the analyzed cohort (for example difference among adenocarcinoma patients being larger than in other histology cohorts [10]). Survival discrepancies may also depend on the geographical location due to different population parameters and treatment practices [11].

In addition to these commonly analyzed factors, obesity was shown to be associated with better survival [12–14] in lung cancer. Because obesity is usually associated with negative effects on health, this phenomenon has been called the 'obesity paradox' [15,16]. Obesity rates are also increasing worldwide[17], urging for a better understanding of the specifics of this patient group. As obesity varies at least by sex, age and geographical location [18,19], its total influence on survival could vary in different patient cohorts. Most of the previous studies on sex-associated survival differences have not used obesity or body mass index (BMI) as a covariate or stratification factor, nor reported the share of obese patients in the study population [5,6,10,11,20–25]. In addition, they used different clinical cohorts [6,20,24], geographic locations [21,25,26] and statistical methods [5,21,25], which limits the comparability of these studies. Finally, many studies on the topic[13,15,27] have focused on NSCLC excluding other lung cancer types commonly present in clinical practice.

To overcome these limitations and describe the effect of obesity on sex-associated survival differences we collected a large retrospective cohort of all cancer-naive lung cancer patients from a single hospital in Finland, which covers an area of ~2 million inhabitants. We quantified the association between sex and survival as well as the effect of covariates like stage, treatment and histology in low BMI (BMI<18.5), normal BMI (BMI 18.5-25) and high BMI patients (BMI>25). In addition, we showcased the value of the Observational Medical Outcomes Partnership (OMOP) Common Data Model in representing a granular view of all clinical data in a large university hospital setting [28].

2 Material and methods

2.1 Data sources, variables and preprocessing

Pseudonymized clinical data from Helsinki University Central Hospital (HUS) was used for analysis in this report. HUS is responsible for secondary/tertiary cancer care (diagnostics, treatment, follow-up) of 2.2 million inhabitants. Structured data including prescribed and administered drugs, medical procedures, laboratory measurements, visits, diagnoses and patient demographics was collected from multiple operative medical information systems and harmonized into the OMOP common data model format by mapping local medical codes (e.g. drug or procedure codes) to standard OMOP vocabularies and concepts[28]. In addition, unstructured data (patient notes), including patient journals, admission notes, briefing reports, questionnaires etc. were collected from information systems and used to derive the stage at diagnosis, smoking status and metastasis status and location (see Supplementary Appendix 1 for the full description). The full derivation of each clinical feature, preprocessing and OMOP definitions are provided in Supplementary Appendix 2. This study was conducted in accordance with the EU General Data Protection Regulation (GDPR), which permits the secondary use of personal health data for scientific research purposes under Article 9(2)(j) and Article 89, and complies with the Finnish Act on the Secondary Use of Health and Social Data (552/2019).

2.2 Patient cohorts and exclusions

First, all patients with lung cancer (ICD10 C34) diagnosed between January 1, 2015 to December 31, 2024 were selected. Patients with non-malignant disorders (ICD10 C34.x7) were excluded from the cohort. Patients who were not registered at HUS at least 1 month before lung cancer diagnosis were further excluded to ensure data availability at the moment of diagnosis, which led to filtering out 488 (6%) of patients (See Supplementary Fig. 1). Patients who had any other cancer diagnosed before lung cancer diagnosis (n=1902) were excluded to reduce the variance in clinical outcomes and interpretation of the results, and to simplify automated data processing of unstructured data.

Body Mass Index (BMI) was used to split patients into cohorts. BMI is defined as patient weight (kg) divided by the square of patient height (m). To derive BMI at diagnosis we took the average value of all BMI measurements 60 days before diagnosis and maximum 60 days after or until the first treatment. Where BMI was not available, we derived it from independent height and weight measurements if they were accessible. Patients were split based on BMI into three sub-cohorts - low BMI (BMI<18.5, n=284), normal BMI (BMI 18.5-25, n=2029) and high BMI (BMI>25, n=2057).

2.3 Survival analysis

Survival was defined as the time from the date of diagnosis to the date of death. Date of diagnosis was defined as the date of the first appearance of lung cancer diagnosis code (any ICD C34.xx excluding non-malignant codes C34.x7). The follow-up time for all patients is until January 21, 2025. The difference of distribution of clinical variables (histology, age, sex, stage, smoking status, blood measurements, and comorbidity index) between cohorts was compared by Pearson's chi-squared test for categorical variables and Mann–Whitney U test for numerical variables. For the comparison of single covariates between cohorts we used a z-test of proportions (categorical values) and a Mann–Whitney U test (numerical values). Bonferroni correction was applied to correct for multiple comparisons.

Cox[29] and RMST (Restricted Mean Survival Time)[30] regression were used for survival analysis. Cox regression, Kaplan–Meier survival curves and statistical testing was performed using Python package 'lifelines' (version 0.30.0)[31]. Proportional hazard assumption was tested by the inspection of Schoenfeld Residuals. Cox regression was stratified by histology using functionality of 'lifelines'.

The restricted mean survival time analysis (RMST) was used to estimate the effect of sex on survival differences [30] [32]. RMST represents the average event-free survival time up to a specified time point (Supplementary Figure 2). RMST is defined as the area under the Kaplan–Meier curve up to a defined time point and provides a robust measure that does not rely on the assumption of proportional hazards [33]. The effect of covariates was evaluated by regression methodology [34]. R package 'survrm2' was used for RMST regression, p-values and confidence intervals. Selected methods and their underlying assumptions allow to generalize results across the lung cancer population in Finland, but require external validation for a wider generalization.

Results

Patient characteristics

In total, 7,988 patients with malignant lung cancer diagnosed between January 1, 2015 and December 31, 2024 were identified in HUS patient records (Supplementary Fig. 2). Out of those, 7,500 were registered at HUS at least one month before lung cancer diagnosis. After filtering out patients with other cancers diagnosed before lung cancer, the full study cohort included 5,598 patients out of which 53% were male patients. Patients with previous cancer diagnosis were excluded to reduce the survival variance caused by other cancers, clarify the interpretation of medical records (see Supplementary Appendix 2), and simplify interpretability of results. Patients were split into three sub-cohorts based on BMI thresholds – low BMI (BMI<18.5, n=284), normal BMI (BMI 18.5-25, n=2029) and high BMI (BMI>25, n=2057) (Table 1). Out of patients with available BMI measurements, patients with high BMI represented 47% among males and 48% among females, which is a lower value than measured for the whole population in Finland in comparable age groups (75% in males and 70% in females)[36].

In addition, 22% of all patients (n=1219) had missing BMI values. As expected, this group had a low share of patients treated with surgery (3% vs 16% in full cohort, p<0.001) and chemotherapy (6% vs 28% in full cohort, p<0.001). This indicates that this group mostly includes patients who either voluntarily declined the treatment or were not eligible for treatment due to fragility or advanced progression of cancer and thus their BMI was not measured.

In the full cohort, Stage IV was the most frequent stage at diagnosis (41%). In addition, for 34% of patients, the stage was missing (not retrievable from patient charts). However, only 12% of patients with the missing stage received surgery (compared to 46% among patients with known stage). This group mostly includes patients with a progressed disease, where an explicit TNM classification was not necessary for clinical decisions, as well as a small share of patients with lower stage not detected from data due to technical limitations.

Adenocarcinoma was the most frequent histology in the full cohort (n=2,225), followed by squamous cell carcinoma (n=850) and undetermined (n=834), which includes patients for whom the biopsy was not taken for example due to fragility and the diagnosis was recorded only on the basis of CT or MRI scan. Patients in the age group 70-79 were the largest patient group in all three cohorts and, patients in the age of 80+ were underrepresented in the high BMI cohort representing only 11% with corresponding share in the full cohort being 18% (p<0.001).

Treatment lines were derived from the data based on the time intervals and treatment types (See methods). Treatment options were classified into surgery,

chemotherapy, immune checkpoint inhibitors (ICI), radiotherapy, and targeted therapy. Sankey plot of the first three treatment lines is shown in Supplementary Figure 3. Radiotherapy was the most used therapy in the first line treatment, administered either alone (n=1312) or in combination (n=490). In the full cohort, 2324 patients (42%) received only one line of therapy.

Feature	Value	Full cohort	Low BMI	Normal BMI	High BMI	Missing BMI
Stage	I	603 (11%)	16 (6%)	205 (10%)	265 (13%)	117 (10%)
	III	217 (4%)	6 (2%)	66 (3%)	114 (6%)	31 (3%)
	III	585 (10%)	18 (6%)	223 (11%)	251 (12%)	93 (8%)
	IV	2266 (41%)	128 (45%)	909 (45%)	849 (41%)	380 (31%)
	Unknown	1918 (34%)	116 (41%)	626 (31%)	578 (29%)	598 (49%)
Histology	Adenocarcinoma	2225 (40%)	87 (31%)	874 (43%)	958 (47%)	306 (25%)
	NSCLC other	205 (4%)	12 (4%)	93 (5%)	75 (4%)	25 (2%)
	Non-specified, or other	800 (14%)	52 (18%)	255 (13%)	202 (10%)	291 (24%)
	SCLC	675 (12%)	33 (12%)	277 (14%)	307 (12%)	58 (5%)
	Squamous cell carcinoma	850 (15%)	42 (15%)	311 (15%)	344 (17%)	153 (13%)
	Undetermined	834 (15%)	58 (20%)	219 (11%)	171 (8%)	386 (32%)
Smoker	Ever	4416 (79%)	226 (80%)	1637 (81%)	1694 (82%)	859 (70%)
	Never	466 (8%)	9 (3%)	175 (9%)	187 (9%)	95 (8%)
	Unknown	707 (13%)	49 (17%)	221 (11%)	176 (9%)	265 (22%)
Age group	59-	689 (12%)	27 (10%)	267 (13%)	329 (16%)	66 (5%)
	60-69	1584 (28%)	79 (28%)	581 (29%)	656 (32%)	268 (22%)
	70-79	2315 (41%)	133 (47%)	864 (43%)	846 (41%)	472 (39%)
	80+	1001 (18%)	45 (16%)	317 (16%)	226 (11%)	413 (34%)
First line includes:	CI	306 (5%)	13 (5%)	141 (7%)	138 (7%)	14 (1%)
	Chemo	1545 (28%)	65 (23%)	675 (33%)	736 (36%)	79 (6%)
	Surgery	910 (16%)	20 (7%)	362 (18%)	495 (24%)	33 (3%)
	Radiotherapy	1802 (31%)	81 (29%)	624 (31%)	590 (29%)	507 (42%)
	Targeted	247 (4%)	10 (4%)	92 (5%)	95 (5%)	50 (4%)
Sex	Female	2597 (47%)	159 (56%)	930 (46%)	933 (46%)	575 (47%)
	Male	3992 (53%)	125 (44%)	1099 (54%)	1124 (54%)	644 (53%)
Total		5589	284	2029	2057	1219

Table 1. Patient characteristics in the full patient group and BMI-based cohorts: low (BMI<18.5), normal BMI (BMI 18.5-25) and high BMI (BMI >25) Undetermined histology = biopsy was not collected. Abbreviations: NSCLC other (Non small cell lung cancer, other);

NS, or other (Non-specified or other); SCLC Small cell lung cancer; ICI – immune checkpoint inhibitor therapy;

Comparison of clinical features between male and female patients with normal and high BMI

Sex differences at baseline (time of diagnosis) in clinical features for the patients in normal and high BMI cohorts are summarized in Table 2. In the normal BMI cohort female patients had lower stage at diagnosis, as 16% of female patients were diagnosed with stage I-II lung cancer, while for male patients the share was 11% ($p<0.01$). Adenocarcinoma was more common among female patients (53% Female vs. 35% male, $p<0.001$), while squamous cell carcinoma was more common among male patients (11% female - 19% male, $p<0.001$). The proportion of never smokers was larger in female patients (13% Female vs. 5% male, $p<0.001$). Male patients also had a higher Charlson comorbidity index[37] (CCI) (see Supplementary Appendix 1). Female patients received targeted therapy (7% female patients - 3% male patients, $p<0.001$) and surgical treatment (23% female patients - 14% male patients, $p<0.001$) more often, likely due to being diagnosed at an earlier stage (Table 2).

Feature	Value	Male Normal	Female Normal	P-value Normal	Male High BMI	Female High BMI	P-value High BMI
Stage	I	84 (8%)	121 (13%)	0.015	135 (12%)	130 (14%)	0.008
	II	36 (3%)	30 (3%)		48 (4%)	66 (7%)	
	III	119 (11%)	104 (11%)		150 (13%)	101 (11%)	
	IV	499 (45%)	410 (44%)		458 (41%)	391 (42%)	
	Unknown	361 (33%)	265 (28%)		333 (30%)	245 (26%)	
Histology	Adenocarcinoma	385 (35%)	489 (53%)	<0.001	468 (42%)	490 (52%)	<0.001
	NSCLC other	59 (5%)	34 (4%)		52 (5%)	23 (2%)	
	Non-specified, or other	168 (15%)	87 (9%)		107 (10%)	95 (10%)	
	SCLC	145 (13%)	132 (14%)		175 (16%)	132 (14%)	
	Squamous cell carcinoma	207 (19%)	104 (11%)		229 (20%)	115 (12%)	
	Undetermined	135 (12%)	84 (9%)		93 (8%)	78 (8%)	
Smoker	Ever	928 (84%)	709 (76%)	<0.001	971 (86%)	723 (78%)	<0.001
	Never	51 (5%)	124 (13%)		61 (5%)	126 (14%)	
	Unknown	120 (11%)	97 (10%)		92 (8%)	84 (9%)	
First line includes:	CI	77 (7%)	64 (7%)	0.9	82 (7%)	56 (6%)	0.20
	Chemo	375 (34%)	300 (32%)	0.4	427 (33%)	309 (33%)	0.02
	Surgery	148 (14%)	214 (23%)	<0.001	257 (26%)	238 (26%)	0.01

	Radiotherapy	363 (33%)	261 (28%)	0.017	312 (30%)	278 (30%)	0.3
	Targeted	31 (3%)	61 (7%)	<0.001	42 (6%)	53 (6%)	0.04
Comorbidity_score		0.99	0.78	<0.001	1.17	0.92	<0.001

*Table 2. Comparison of clinical covariates between male and female patients in normal BMI and high BMI cohorts. *Continuous values, mean shown. Distributions of categorical features were compared for each category by a chi-squared test, and numerical values by a Mann–Whitney U test. Abbreviations NSCLC other (Non small cell lung cancer, other); NS, or other (Non-specified or other); SCLC Small cell lung cancer; ICI - checkpoint inhibitor therapy;*

Survival difference between males and females is lower in high BMI patients in univariate analysis especially in squamous cell carcinoma

In the full cohort (Figure 1A), survival in females in univariate analysis was significantly longer than in males ($p<0.001$) with a 29% two-year (2y) survival rate in males and 41% in females. Among patients within the normal BMI cohort, 2y survival difference was 18% (41% female and 24% male). The difference was smaller among patients with high BMI - 9% (38% male and 47% female 2y survival rates, $p<0.001$) (Figure 1 D). The result was also validated by a cox-regression with an interaction variable, indicating interaction between high BMI and male sex ($p<0.01$) (Supplementary Figure 4).

As survival difference between males and females was smaller among high BMI patients compared to normal, we wanted to see whether this holds for all histologies. We repeated the analysis separately for each histology (Figure 2, Supplementary Figures 5-9). The effect was the largest in Squamous cell carcinoma (Figure 2) (2y survival rates among normal BMI: males - 28% females 54%, high BMI: males - 50% females 57%). In adenocarcinoma the effect was moderate (2y survival rates among normal BMI: males - 37% females 54%, high BMI: males - 46% females 59%), while in SCLC there was no effect of high BMI on survival difference.

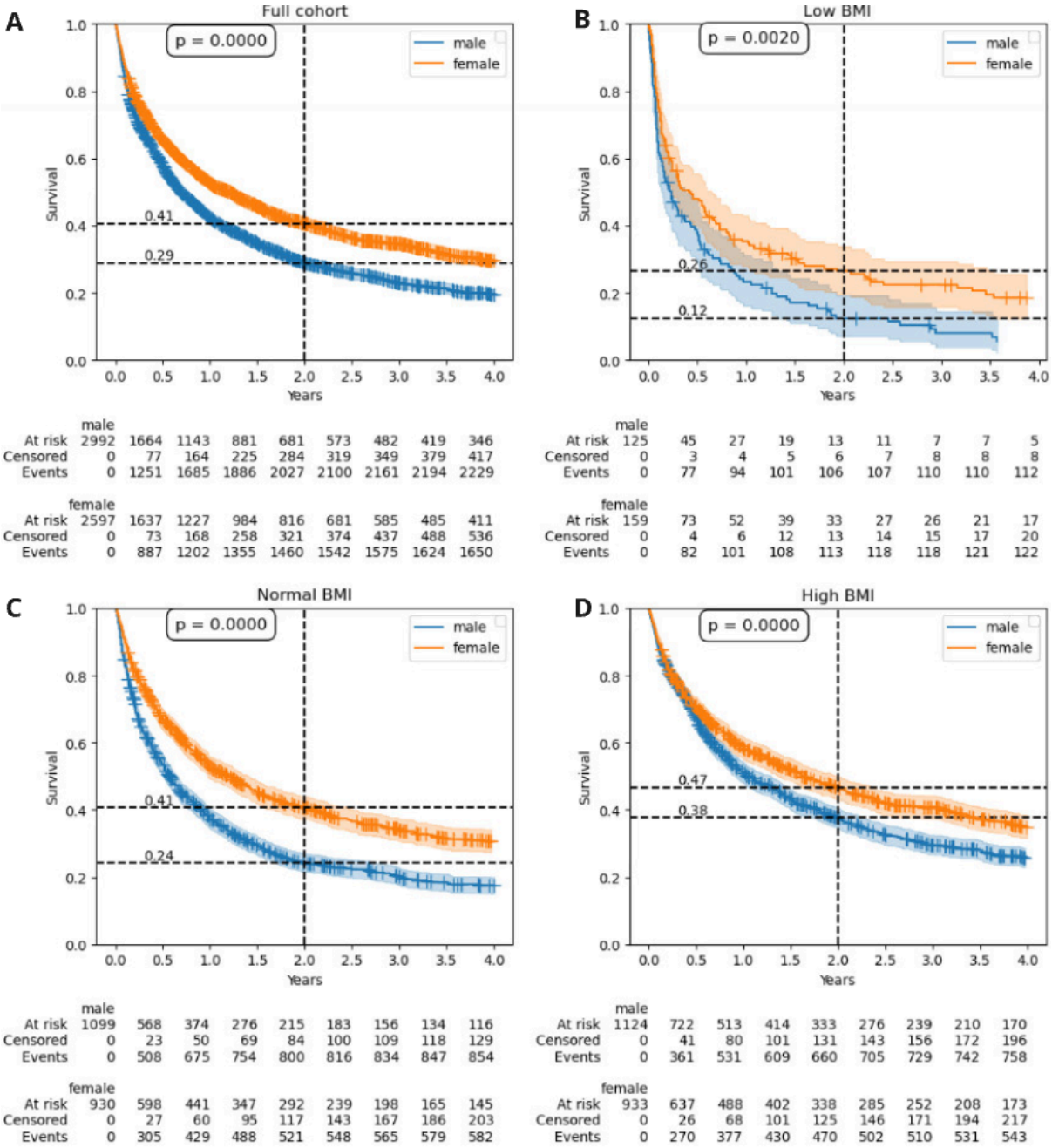


Figure 1. Kaplan-Meier survival curves stratified by gender for full cohort (A), Low BMI (B), Normal BMI (C), and High BMI (D). P values are calculated from a log-rank test. 2-year survivals are visualized by dashed lines.

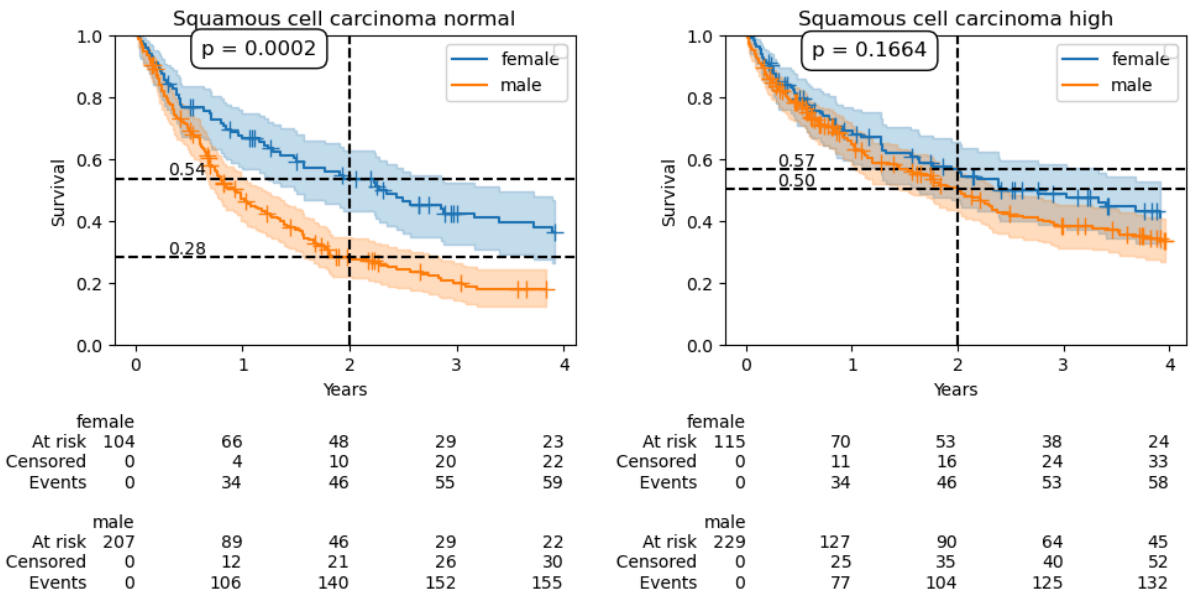


Figure 2 Kaplan-Meier survival plot for patients with squamous cell carcinoma histology stratified by sex , normal BMI (left) and high BMI (right) comparison.

The effect of male sex on survival was 42% smaller among high BMI patients in multivariate analysis

To understand how covariates affect the effect of sex on survival difference we performed multivariate regression analyses in high and normal BMI patients. In order to understand the effects of covariates on all patients before stratified analysis, we ran multivariate Cox and RMST regression analyses for the full patient cohort using sex, age, BMI group, histology, smoking status, stage, first line therapy, and comorbidity index as covariates (Figure 3, Supplementary Figure 8). There were no contradictions between the results obtained by the two methods. As expected, higher stage, age, never-smoking and male sex were associated with lower survival. High BMI was associated with higher survival (RMST 0.93 ICI 0.45-1.41) while low BMI was associated with lower survival (RMST -1.10 ICI -2.09-0.10) than normal BMI, which was used as a baseline cohort.

To understand how BMI influences the effect of sex on survival, multivariate regression was done separately for high BMI (Figure 4A) and normal BMI (Figure 4B) patients using age, sex, group, histology, smoking status, stage, first line therapy, and comorbidity index as covariates. The effect of male sex was reduced by 44% in high BMI patients (RMST -1.06 ICI -1.73 – -0.40, $p<0.001$) compared to normal BMI patients (RMST -1.83 ICI -2.52 – -1.14, $p<0.001$).

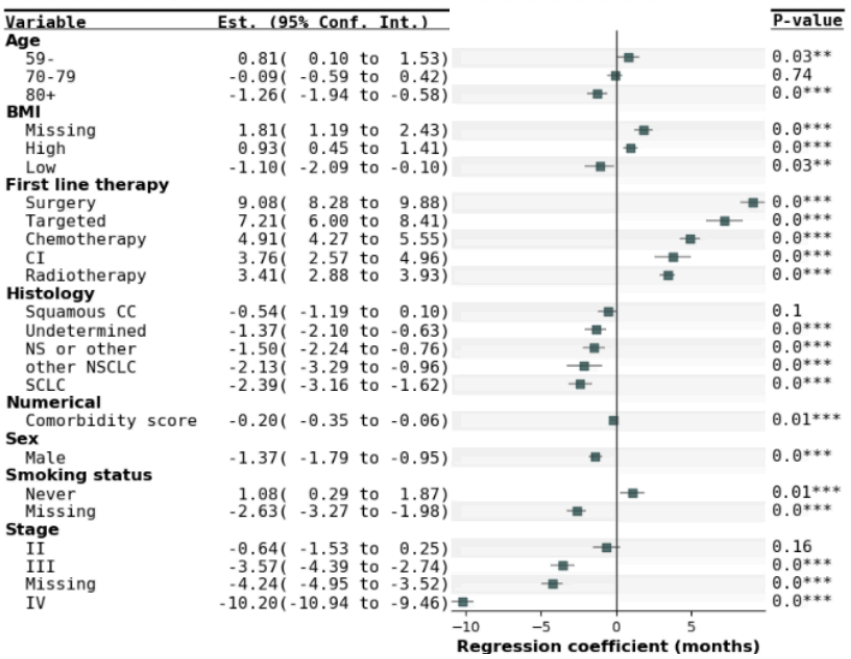


Figure 3. 2y RMST regression for the full cohort. Abbreviations NSCLC other (Non small cell lung cancer, other); NS, or other (Non-specified or other); SCLC Small cell lung cancer; ICI - checkpoint inhibitor therapy; Conf. Int - confidence interval. Baseline categorical covariates are age 60-69, normal BMI, no first line treatment, adenocarcinoma, female sex, smoker, stage I.

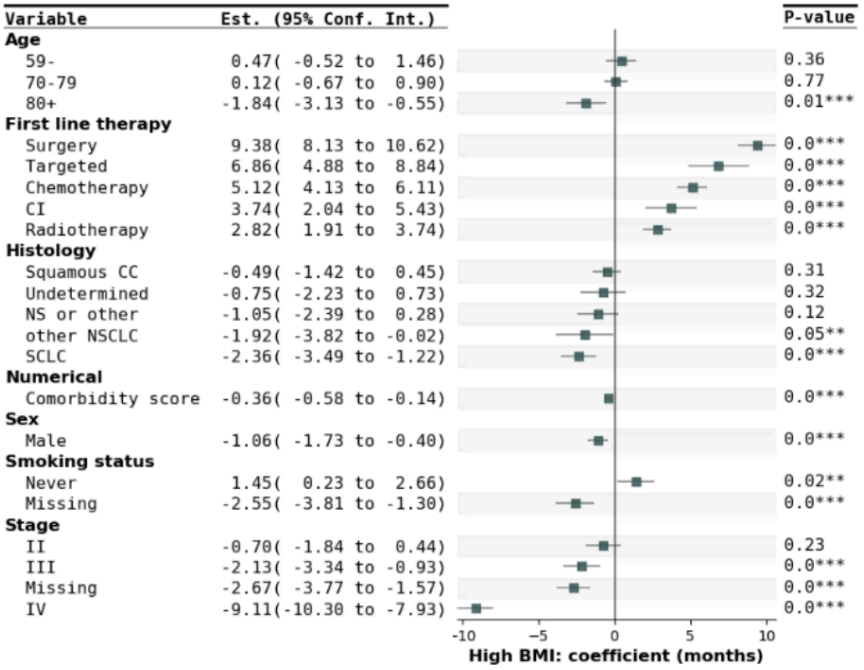


Figure 4A 2y RMST regression for high BMI patients in the full cohort. Abbreviations NSCLC other (Non small cell lung cancer, other); NS, or other (Non-specified or other); SCLC Small cell lung cancer; ICI - checkpoint inhibitor therapy; Conf. Int - confidence interval. Baseline

categorical covariates are age 60-69, normal BMI, no first line treatment, adenocarcinoma, female sex, smoker, stage I.

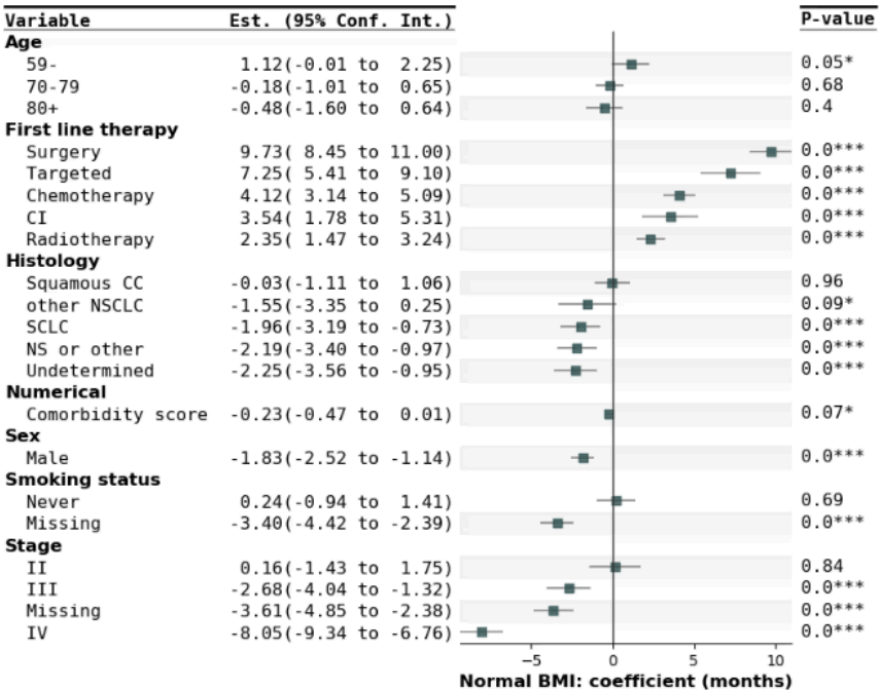


Figure 4B 2y RMST regression for normal BMI patients in the full cohort. Abbreviations NSCLC other (Non small cell lung cancer, other); NS, or other (Non-specified or other); SCLC Small cell lung cancer; ICI - checkpoint inhibitor therapy; Conf. Int - confidence interval

Discussion

In this study, we showcased the clinical value of large-scale cancer data collection and harmonization using OMOP methodologies and standards. In case of a larger adoption of these practices, similar studies could be replicated among data providers. Compared to traditional registry studies this approach provides access to more granular patient data enabling the analysis of complex research questions like the one addressed in this study.

We identified that the survival difference between males and females is smaller among patients with higher BMI as high BMI is associated with a larger survival benefit in males than in females. We have also shown the variation of the effect by histology, underlying the need for further research on less frequently occurring histologies such as squamous cell carcinoma. We also confirmed that the rate of patients with high BMI is lower among lung cancer patients than in the whole Finnish population.

Based on the findings we propose that future RWE studies report BMI distributions of patient populations as well as use BMI as a covariate in the multivariate survival analysis. We state a hypothesis that differences in overall survival and sex-related survival discrepancies between previous studies can be partially explained by different BMI distributions of patient populations in these studies. We also propose that the reports of lung cancer registries[26,38] would include BMI stratification as they can partly explain observed survival difference between sexes. Finally, we also suggest that any further lung cancer survival and therapy response models and scores would include BMI as an input feature.

From a clinical perspective interaction between male sex and high BMI may indicate that there are biological mechanisms related to higher BMI which are stronger in males. Earlier it has been stated that female patients have a stronger immune response to lung cancer[39]. As high BMI is also associated with immune system activity[40] it may have an impact on the poorer immune response in male patients. In addition, BMI may act as a proxy for another biological feature. Previous studies[16,41] have tried to differentiate between high muscle mass and high fat percentage patients to understand which of them are driving better survival rates.

This study has similar limitations as other RWE studies. It relies on the assumption that the data created in clinical work is accurate and does not have systematic biases. In addition, the data features extraction pipeline from free-form medical reports does not have 100% accuracy as addressed in the methodology. However, as ~90% of all extractions are done correctly we state this does not have a significant impact on the main conclusions of the study. Finally, smoking status is self-reported by patients and may include intentional or unintentional errors.

In conclusion, survival comparison between sexes in lung cancer should include obesity assessments and stratification. In addition, based on this study we would like to encourage stratification of patients by BMI during clinical trials and reporting BMI distribution of trial participants. In addition we conclude that the difference is associated with histology and we encourage further studies on squamous cell carcinoma to understand the effect of higher BMI on survival.

References:

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clin. 2024;74(3):229–63.
2. Ferguson MK, Wang J, Hoffman PC, Haraf DJ, Olak J, Masters GA, et al. Sex-associated differences in survival of patients undergoing resection for lung cancer. Ann Thorac Surg. 2000;69(1):245–9.
3. Mincuzzi A, Carone S, Galluzzo C, Tanzarella M, Lagravinese GM, Bruni A, et al. Gender differences, environmental pressures, tumor characteristics, and death rate in a lung cancer cohort: a seven-years Bayesian survival analysis using cancer registry data from a contaminated area in Italy. Front Public Heal. 2024;11:1278416.
4. Guerreiro T, Forjaz G, Antunes L, Bastos J, Mayer A, Aguiar P, et al. Lung cancer survival and sex-specific patterns in Portugal: A population-based analysis. Pulmonology. 2023;29:S70–9.
5. Visbal AL, Williams BA, Nichols FC, Marks RS, Jett JR, Aubry MC, et al. Gender differences in non–small-cell lung cancer survival: an analysis of 4,618 patients diagnosed between 1997 and 2002. Ann Thorac Surg. 2004;78(1):209–15.
6. Cerfolio RJ, Bryant AS, Scott E, Sharma M, Robert F, Spencer SA, et al. Women With Pathologic Stage I, II, and III Non-small Cell Lung Cancer Have Better Survival Than Men. Chest. 2006;130(6):1796–802.
7. Lærum D, Strand TE, Brustugun OT, Gallefoss F, Falk R, Durheim MT, et al. Evaluation of sex inequity in lung-cancer-specific survival. Acta Oncol. 2024;63:27572.
8. Bugge A, Kongerud J, Brunborg C, Solberg S, Lund MB. Gender-specific survival after surgical resection for early stage non-small cell lung cancer. Acta Oncol. 2017;56(3):448–54.
9. Yu XQ, Yap ML, Cheng ES, Ngo PJ, Vaneckova P, Karikios D, et al. Evaluating Prognostic Factors for Sex Differences in Lung Cancer Survival: Findings From a Large Australian Cohort. J Thorac Oncol. 2022;17(5):688–99.
10. Salmerón D, Chirlaque MD, Izarzugaza MI, Sánchez MJ, Marcos-Gragera R, Ardanaz E, et al. Lung cancer prognosis in Spain: The role of histology, age and sex. Respir Med. 2012;106(9):1301–8.
11. Araghi M, Fidler-Benaoudia M, Arnold M, Rutherford M, Bardot A, Ferlay J, et al. International differences in lung cancer survival by sex, histological type and stage at diagnosis: an ICBP SURVMARK-2 Study. Thorax. 2022;77(4):378–90.
12. Yang R, Cheung MC, Pedroso FE, Byrne MM, Koniaris LG, Zimmers TA. Obesity and Weight Loss at Presentation of Lung Cancer are Associated with Opposite Effects on Survival. J Surg Res. 2011;170(1):e75–83.
13. Wei L, Liu C, Wang S, Zhang H, Ruan G, Xie H, et al. Sex differences in the obesity paradox of body compositions in non-small cell lung cancer. Nutrition. 2025;132:112690.
14. Li S, Wang Z, Huang J, Fan J, Du H, Liu L, et al. Systematic review of prognostic roles of body mass index for patients undergoing lung cancer surgery:

does the ‘obesity paradox’ really exist? Eur J Cardio-Thorac Surg. 2017;51(5):817–28.
15. Zhang X, Liu Y, Shao H, Zheng X. Obesity Paradox in Lung Cancer Prognosis: Evolving Biological Insights and Clinical Implications. J Thorac Oncol. 2017;12(10):1478–88.
16. Barbi J, Patnaik SK, Pabla S, Zollo R, Smith RJ, Sass SN, et al. Visceral Obesity Promotes Lung Cancer Progression—Toward Resolution of the Obesity Paradox in Lung Cancer. J Thorac Oncol. 2021;16(8):1333–48.
17. Phelps NH, Singleton RK, Zhou B, Heap RA, Mishra A, Bennett JE, et al. Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. Lancet. 2024;403(10431):1027–50.
18. Ahrens W, Pigeot I, Pohlabeln H, Henauw SD, Lissner L, Molnár D, et al. Prevalence of overweight and obesity in European children below the age of 10. Int J Obes. 2014;38(Suppl 2):S99–107.
19. Collaborators G 2021 UOF, Ng M, Dai X, Cogen RM, Abdelmasseh M, Abdollahi A, et al. National-level and state-level prevalence of overweight and obesity among children, adolescents, and adults in the USA, 1990–2021, and forecasts up to 2050. Lancet. 2024;404(10469):2278–98.
20. Ferguson MK, Wang J, Hoffman PC, Haraf DJ, Olak J, Masters GA, et al. Sex-associated differences in survival of patients undergoing resection for lung cancer. Ann Thorac Surg. 2000;69(1):245–9.
21. Mincuzzi A, Carone S, Galluzzo C, Tanzarella M, Lagravinese GM, Bruni A, et al. Gender differences, environmental pressures, tumor characteristics, and death rate in a lung cancer cohort: a seven-years Bayesian survival analysis using cancer registry data from a contaminated area in Italy. Front Public Heal. 2024;11:1278416.
22. Guerreiro T, Forjaz G, Antunes L, Bastos J, Mayer A, Aguiar P, et al. Lung cancer survival and sex-specific patterns in Portugal: A population-based analysis. Pulmonology. 2023;29(sup4):S70–9.
23. Lærum D, Strand TE, Brustugun OT, Gallefoss F, Falk R, Durheim MT, et al. Evaluation of sex inequity in lung-cancer-specific survival. Acta Oncol. 2024;63:27572.
24. Bugge A, Kongerud J, Brunborg C, Solberg S, Lund MB. Gender-specific survival after surgical resection for early stage non-small cell lung cancer. Acta Oncol. 2017;56(3):448–54.
25. Yu XQ, Yap ML, Cheng ES, Ngo PJ, Vaneckova P, Karikios D, et al. Evaluating Prognostic Factors for Sex Differences in Lung Cancer Survival: Findings From a Large Australian Cohort. J Thorac Oncol. 2022;17(5):688–99.
26. Børø S, Thoresen S, Helland Å. Improvements in survival for patients with stage IV adenocarcinoma in the lung, diagnosed between 2010 – 2020 - A population-based registry study from Norway. Front Oncol. 2022;12:1017902.
27. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics [Internet]. [cited 2025 Apr 13]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/

28. Cox D. Regression Models and Life-Tables. Cox, David R. 1972 May 10;
29. Han K, Jung I. Restricted Mean Survival Time for Survival Analysis: A Quick Guide for Clinical Researchers. Korean J Radiol. 2022;23(5):495–9.
30. Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw. 2019;4(40):1317.
31. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Méd Res Methodol. 2013;13(1):152.
32. Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? JAMA. 2020;323(14):1401–2.
33. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics. 2014;15(2):222–33.
34. PATERNOSTER R, BRAME R, MAZEROLLE P, PIQUERO A. USING THE CORRECT STATISTICAL TEST FOR THE EQUALITY OF REGRESSION COEFFICIENTS. Criminology. 1998;36(4):859–66.
35. Koponen P, Borodulin K, Lundqvist A, Sääksjärvi K, Koskinen S. Terveys, toimintakyky ja hyvinvointi Suomessa FinTerveys 2017-tutkimus. Terveyden ja hyvinvoinnin laitos. 2018 Feb 8;
36. Glasheen WP, Cordier T, Gumpina R, Haugh G, Davis J, Renda A. Charlson Comorbidity Index: ICD-9 Update and ICD-10 Translation. American Health & Drug Benefits. 2019 Jun 13;

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clin. 2024;74(3):229–63.
2. Ferguson MK, Wang J, Hoffman PC, Haraf DJ, Olak J, Masters GA, et al. Sex-associated differences in survival of patients undergoing resection for lung cancer. Ann Thorac Surg. 2000;69(1):245–9.
3. Mincuzzi A, Carone S, Galluzzo C, Tanzarella M, Lagravinese GM, Bruni A, et al. Gender differences, environmental pressures, tumor characteristics, and death rate in a lung cancer cohort: a seven-years Bayesian survival analysis using cancer registry data from a contaminated area in Italy. Front Public Heal. 2024;11:1278416.
4. Guerreiro T, Forjaz G, Antunes L, Bastos J, Mayer A, Aguiar P, et al. Lung cancer survival and sex-specific patterns in Portugal: A population-based analysis. Pulmonology. 2023;29:S70–9.
5. Visbal AL, Williams BA, Nichols FC, Marks RS, Jett JR, Aubry MC, et al. Gender differences in non–small-cell lung cancer survival: an analysis of 4,618 patients diagnosed between 1997 and 2002. Ann Thorac Surg. 2004;78(1):209–15.
6. Cerfolio RJ, Bryant AS, Scott E, Sharma M, Robert F, Spencer SA, et al. Women With Pathologic Stage I, II, and III Non-small Cell Lung Cancer Have Better Survival Than Men. Chest. 2006;130(6):1796–802.
7. Lærum D, Strand TE, Brustugun OT, Gallefoss F, Falk R, Durheim MT, et al. Evaluation of sex inequity in lung-cancer-specific survival. Acta Oncol. 2024;63:27572.

8. Bugge A, Kongerud J, Brunborg C, Solberg S, Lund MB. Gender-specific survival after surgical resection for early stage non-small cell lung cancer. Acta Oncol. 2017;56(3):448–54.
9. Yu XQ, Yap ML, Cheng ES, Ngo PJ, Vaneckova P, Karikios D, et al. Evaluating Prognostic Factors for Sex Differences in Lung Cancer Survival: Findings From a Large Australian Cohort. J Thorac Oncol. 2022;17(5):688–99.
10. Salmerón D, Chirlaque MD, Izarzugaza MI, Sánchez MJ, Marcos-Gragera R, Ardanaz E, et al. Lung cancer prognosis in Spain: The role of histology, age and sex. Respir Med. 2012;106(9):1301–8.
11. Araghi M, Fidler-Benaoudia M, Arnold M, Rutherford M, Bardot A, Ferlay J, et al. International differences in lung cancer survival by sex, histological type and stage at diagnosis: an ICBP SURVMARK-2 Study. Thorax. 2022;77(4):378–90.
12. Yang R, Cheung MC, Pedroso FE, Byrne MM, Koniaris LG, Zimmers TA. Obesity and Weight Loss at Presentation of Lung Cancer are Associated with Opposite Effects on Survival. J Surg Res. 2011;170(1):e75–83.
13. Wei L, Liu C, Wang S, Zhang H, Ruan G, Xie H, et al. Sex differences in the obesity paradox of body compositions in non-small cell lung cancer. Nutrition. 2025;132:112690.
14. Li S, Wang Z, Huang J, Fan J, Du H, Liu L, et al. Systematic review of prognostic roles of body mass index for patients undergoing lung cancer surgery: does the 'obesity paradox' really exist? Eur J Cardio-Thorac Surg. 2017;51(5):817–28.
15. Zhang X, Liu Y, Shao H, Zheng X. Obesity Paradox in Lung Cancer Prognosis: Evolving Biological Insights and Clinical Implications. J Thorac Oncol. 2017;12(10):1478–88.
16. Barbi J, Patnaik SK, Pabla S, Zollo R, Smith RJ, Sass SN, et al. Visceral Obesity Promotes Lung Cancer Progression—Toward Resolution of the Obesity Paradox in Lung Cancer. J Thorac Oncol. 2021;16(8):1333–48.
17. Phelps NH, Singleton RK, Zhou B, Heap RA, Mishra A, Bennett JE, et al. Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. Lancet. 2024;403(10431):1027–50.
18. Ahrens W, Pigeot I, Pohlabeln H, Henauw SD, Lissner L, Molnár D, et al. Prevalence of overweight and obesity in European children below the age of 10. Int J Obes. 2014;38(Suppl 2):S99–107.
19. Collaborators G 2021 UOF, Ng M, Dai X, Cogen RM, Abdelmasseh M, Abdollahi A, et al. National-level and state-level prevalence of overweight and obesity among children, adolescents, and adults in the USA, 1990–2021, and forecasts up to 2050. Lancet. 2024;404(10469):2278–98.
20. Ferguson MK, Wang J, Hoffman PC, Haraf DJ, Olak J, Masters GA, et al. Sex-associated differences in survival of patients undergoing resection for lung cancer. Ann Thorac Surg. 2000;69(1):245–9.
21. Mincuzzi A, Carone S, Galluzzo C, Tanzarella M, Lagravinese GM, Bruni A, et al. Gender differences, environmental pressures, tumor characteristics, and death

rate in a lung cancer cohort: a seven-years Bayesian survival analysis using cancer registry data from a contaminated area in Italy. Front Public Heal. 2024;11:1278416.
22. Guerreiro T, Forjaz G, Antunes L, Bastos J, Mayer A, Aguiar P, et al. Lung cancer survival and sex-specific patterns in Portugal: A population-based analysis. Pulmonology. 2023;29(sup4):S70–9.
23. Lærum D, Strand TE, Brustugun OT, Gallefoss F, Falk R, Durheim MT, et al. Evaluation of sex inequity in lung-cancer-specific survival. Acta Oncol. 2024;63:27572.
24. Bugge A, Kongerud J, Brunborg C, Solberg S, Lund MB. Gender-specific survival after surgical resection for early stage non-small cell lung cancer. Acta Oncol. 2017;56(3):448–54.
25. Yu XQ, Yap ML, Cheng ES, Ngo PJ, Vaneckova P, Karikios D, et al. Evaluating Prognostic Factors for Sex Differences in Lung Cancer Survival: Findings From a Large Australian Cohort. J Thorac Oncol. 2022;17(5):688–99.
26. Børø S, Thoresen S, Helland Å. Improvements in survival for patients with stage IV adenocarcinoma in the lung, diagnosed between 2010 – 2020 - A population-based registry study from Norway. Front Oncol. 2022;12:1017902.
27. Nie W, Lu J, Qian J, Wang SY, Cheng L, Zheng L, et al. Obesity and survival in advanced non-small cell lung cancer patients treated with chemotherapy, immunotherapy, or chemoimmunotherapy: a multicenter cohort study. BMC Med. 2024;22(1):463.
28. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics [Internet]. [cited 2025 Apr 13]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/
29. Cox D. Regression Models and Life-Tables. Cox, David R. 1972 May 10;
30. Han K, Jung I. Restricted Mean Survival Time for Survival Analysis: A Quick Guide for Clinical Researchers. Korean J Radiol. 2022;23(5):495–9.
31. Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw. 2019;4(40):1317.
32. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Méd Res Methodol. 2013;13(1):152.
33. Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? JAMA. 2020;323(14):1401–2.
34. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics. 2014;15(2):222–33.
35. PATERNOSTER R, BRAME R, MAZEROLLE P, PIQUERO A. USING THE CORRECT STATISTICAL TEST FOR THE EQUALITY OF REGRESSION COEFFICIENTS. Criminology. 1998;36(4):859–66.
36. Koponen P, Borodulin K, Lundqvist A, Sääksjärvi K, Koskinen S. Terveys, toimintakyky ja hyvinvointi Suomessa FinTerveys 2017-tutkimus. Terveyden ja hyvinvoinnin laitos. 2018 Feb 8;
37. Glasheen WP, Cordier T, Gumpina R, Haugh G, Davis J, Renda A. Charlson Comorbidity Index: ICD-9 Update and ICD-10 Translation. American Health & Drug

Benefits. 2019 June 13;
38. Gouliaev A, Rasmussen TR, Malila N, Fjellbirkeland L, Löfling L, Jakobsen E, et al. Lung cancer registries in Denmark, Finland, Norway and Sweden: a comparison and proposal for harmonization. Acta Oncol. 2023;62(1):1–7.
39. Behrens C, Rocha P, Parra ER, Feng L, Rodriguez-Canales J, Solis LM, et al. Female Gender Predicts Augmented Immune Infiltration in Lung Adenocarcinoma. Clin Lung Cancer. 2021;22(3):e415–24.
40. Hotamisligil GS. Inflammation and metabolic disorders. Nature. 2006;444(7121):860–7.
41. Fu L, Ding H, Mo L, Pan X, Feng L, Wen S, et al. The association between body composition and overall survival in patients with advanced non-small cell lung cancer. Sci Rep. 2025;15(1):3109.

Supplementary Appendix 1

Full description of all clinical features

OMOP concepts are further marked in italic font with their related concept id eg. *antineoplastic and immunomodulating agents ("21601386")*.

Drugs

All *antineoplastic and immunomodulating agents ("21601386")* were selected. Based on the compound name they were manually sorted into chemotherapy, checkpoint-inhibitor therapy and targeted therapy (Appendix 1)

Radiotherapy

Includes following procedures: *Radiotherapy ("1242725")*, *Preoperative course of radiotherapy ("4059384")*, *Postoperative course of radiotherapy ("4058775")*, *Palliative course of radiotherapy ("4057754")*

Surgery

All procedures which are *Lung excision ("4000882")*, except *Excision of lesion of lung ("4300860").* The first category includes actual lung resections, while the latter category includes biopsy sampling procedures.

First line treatments

For each patient, radiotherapy, surgery, chemotherapy, checkpoint-inhibitor therapy and targeted therapy were pulled together and ordered in time. Two treatments of the same type were considered to be the same line of treatment if time duration between them was less than 100 days. Two treatments of different types were considered to be a combination treatment if the time duration between the initiation of treatments was less than 30 days. After labeling all treatments with treatment lines, five binary variables (Therapy_chemo, Therapy_CI, Therapy_Radio, Therapy_Targeted, Therapy_Surgery) were created based on the treatment used during the first line for each patient. If combination therapy was used, patient has multiple binary variable marked as 1 (e.g. Therapy_CI and Therapy_Chemo).

Date of diagnosis

The first appearance of any diagnosis within *Neoplasm of lower respiratory tract ("4112735").* These include all lung, trachea and bronchus neoplasms.

Age

Age at diagnosis is derived from diagnosis date and birth date

Histology

Histology was derived from diagnosis codes, which are based on ICD-10 system.

Stage

First, TNM was extracted using regular expressions from patient notes and converted into stage using AJCC/UICC 8th edition convention (23). In addition, notes mentioning metastasis occurrence were selected using regex and marked as Stage IV in accordance with the convention. Post-therapy

TNM notes “yp”-prefix were excluded. Pathological TNM was used as a primary source for the stage, however if it was not available clinical TNMs were used (see limitations). Patient notes created later than 365 days after the date of diagnosis or earlier than 30 days before were excluded If no stage data was available, the stage was marked as “Unknown”. If multiple stages were detected during the period, the one closest to the date of diagnosis was selected.

Smoking status

Structured smoking data was collected from patient questionnaires. In addition patient notes mentioning smoking were parsed with regular expressions. All datapoints were labeled as “Never smoker”, “Former smoker” or “Current smoker”. Patient notes created later than 100 days after the date of diagnosis or earlier than 365 days before were excluded. If no relevant patient notes were available, smoking status was classified as “unknown”. In case of more than one smoking status was identified, the most frequent status was selected.

Comorbidity index:

Comorbidity index has been constructed by replicating Charlson Comorbidity Index (CCI) based on the ICD codes of patient diagnosis before lung cancer as described in (28)

Supplementary Appendix 2

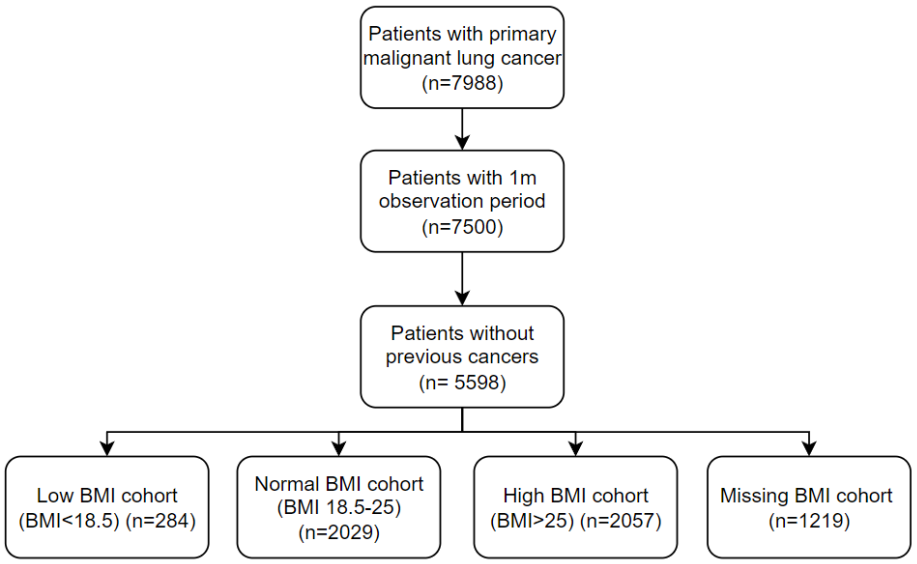
Deriving stage at diagnosis from patient data

First we collected patient notes and related timestamps including consilium meeting notes, imaging statements (CT,MRI,X-Ray, etc.), anamnesis statements and patient journal notes. Two algorithms were used to extract the data. First regex-based algorithm, detected text in the form “[prefix]T[number]N[number]” and optionally “M[number]”. If in a single patient note T and N were present, but M was not available it was assumed to be 0. If multiple notes mentioned the same TNM (including prefix) only the first occurrence was left, and the rest filtered out. TNMs were then mapped to stage using AJCC 8 standards.

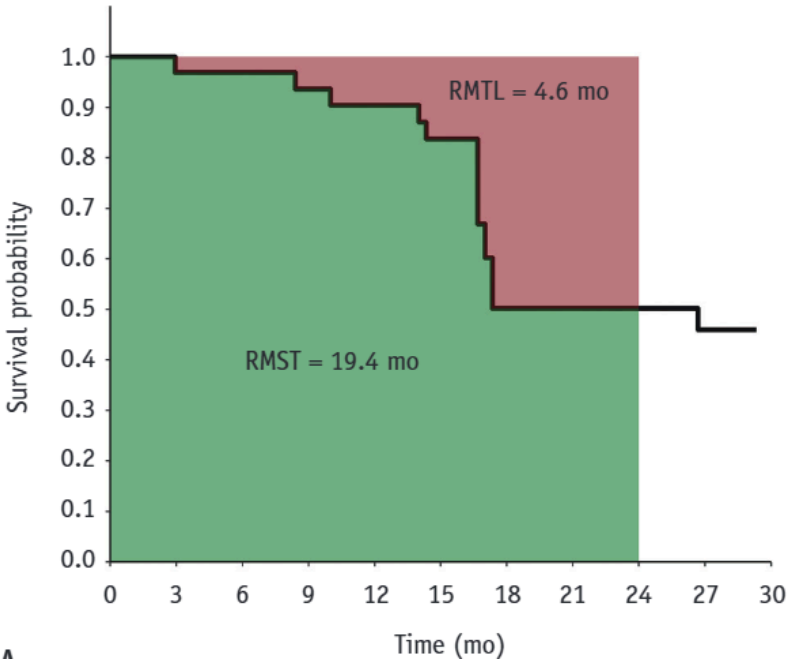
Another algorithm was used to identify notes, which state that the patient has metastasis, but does not have a TNM form. To detect that, we filtered all the senses which include the word metastasis and one of four common distant metastasis locations (‘bone’,’liver’,’brain’,’adrenal gland’). In addition the sentence was not marked as metastatic if it included any of the words

(‘no’,’maybe’,’potential’,’suspect’). Note that the notes and parsing were in Finnish. If a patient had multiple notes mentioning the same metastatic location, only the first was left.

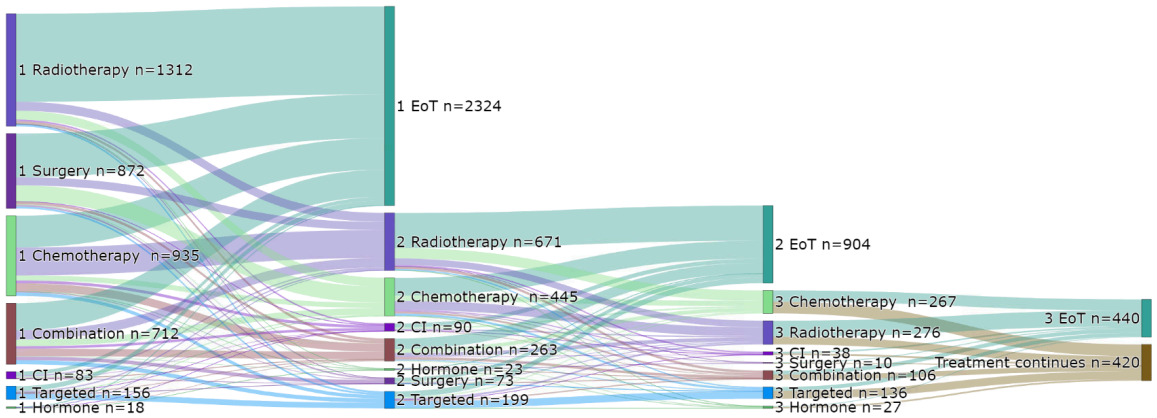
For a sample of 1850 patients, structured Staging information was available. We calculated a binary quality metric based on detection of Stage IV patients. Precision was 0.89 and recall: 0.81.



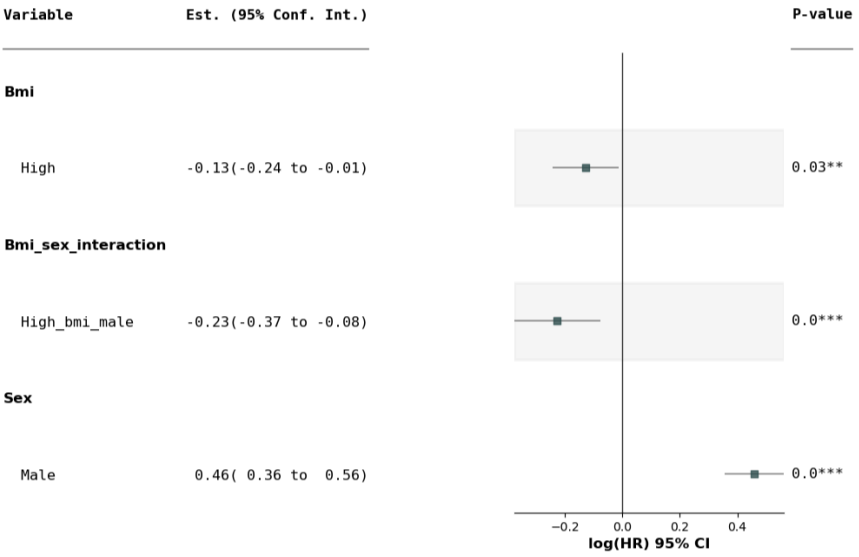
Supplementary Figure 1 Consort flow diagram of the inclusion criteria. all patients with lung cancer (ICD10 C34) diagnosed between January 1, 2015 to December 31, 2024 were selected. Patients with non-malignant disorders (ICD10 C34.x7) were excluded from the cohort. Patients who were not registered at HUS at least 1 year before lung cancer diagnosis (n=567, 10%) were further excluded to ensure data availability before diagnosis (). Patients who had any other cancer diagnosed before lung cancer diagnosis were excluded.



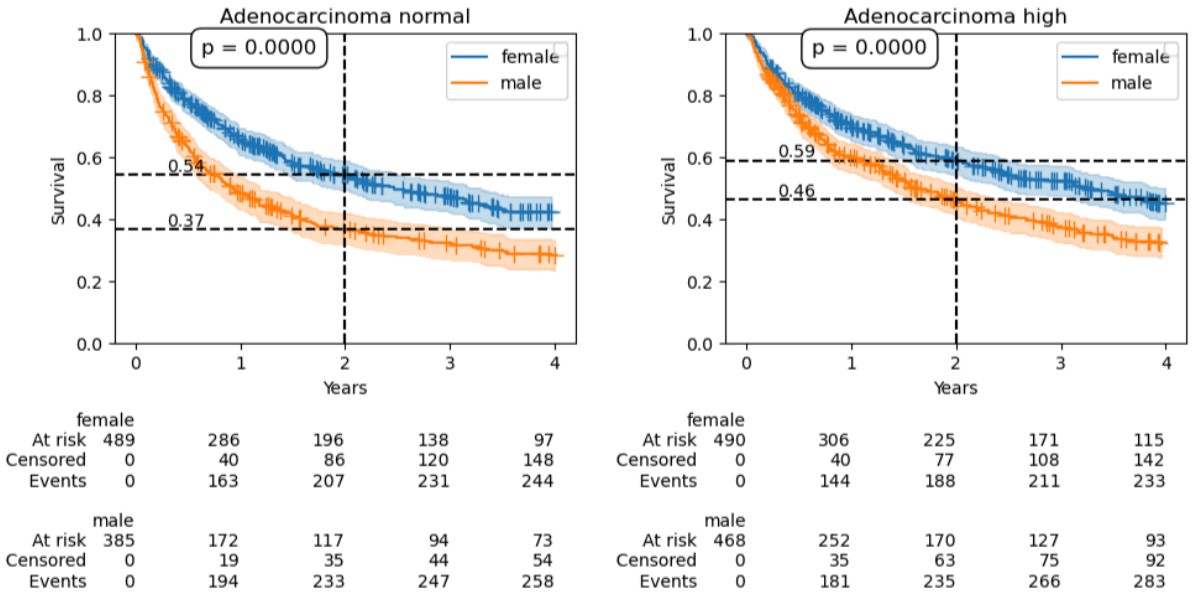
Supplementary Figure 2. Illustrative example of restricted mean survival time (RMST). In this example the threshold T=24. RMST is shown by the green area and is 19.4 months . Figure adapted from (25)



Supplementary Figure 3 Sankey diagram on evolution of different lines of treatment and therapy types for all lung cancer patients within a cohort. "Combination" includes 2 or more therapies of different types. Patients who are treated beyond line 4 are in the group "Treatment continues". The patients who do not continue to the next line of treatment and never receive further treatment are marked as "end of treatment" (EoT).



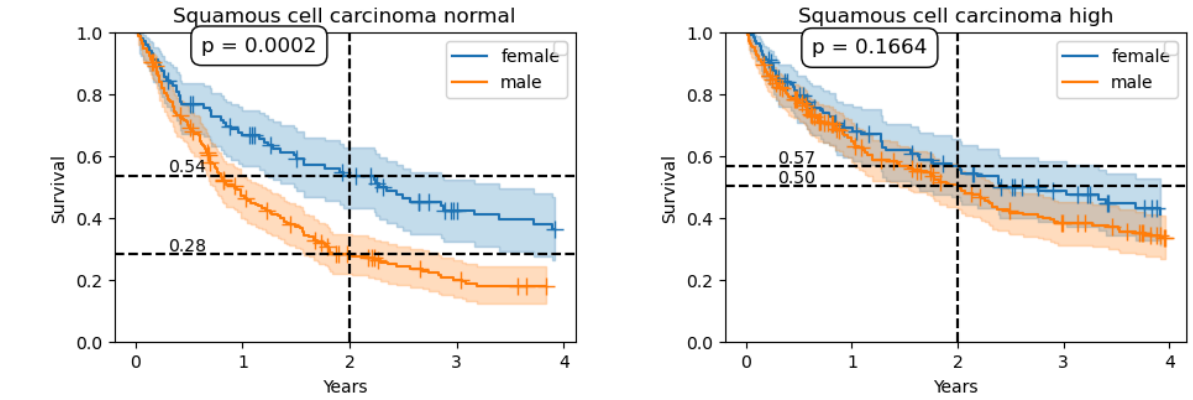
Supplementary Figure 4: Cox-regression with an interaction term. Only patients with high or normal BMI are used. Analysis indicates that there is an interaction between male sex and high BMI. Interaction also depends on histology (Sup Figures 5-9).



female					
At risk	489	286	196	138	97
Censored	0	40	86	120	148
Events	0	163	207	231	244
male					
At risk	385	172	117	94	73
Censored	0	19	35	44	54
Events	0	194	233	247	258

female					
At risk	490	306	225	171	115
Censored	0	40	77	108	142
Events	0	144	188	211	233
male					
At risk	468	252	170	127	93
Censored	0	35	63	75	92
Events	0	181	235	266	283

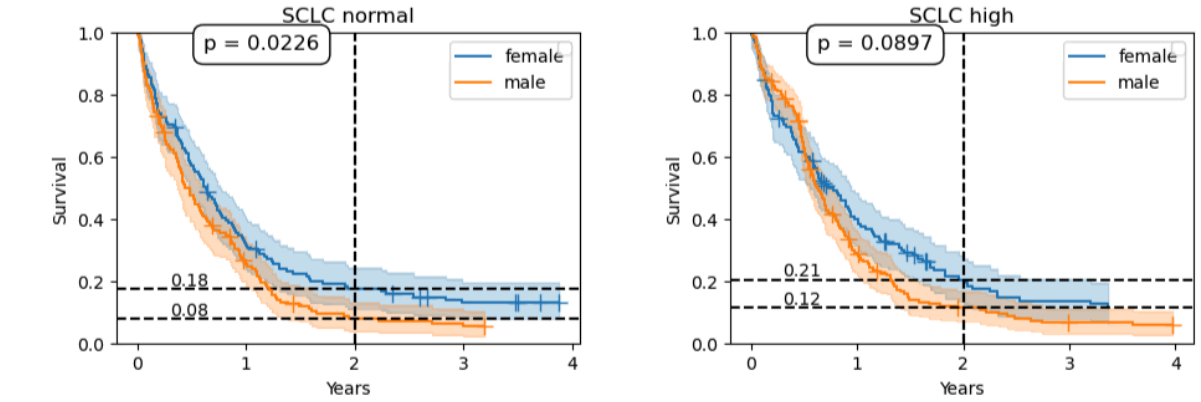
Supplementary Figure 5: Kaplan-Meier survival plot for patients with undetermined histology stratified by sex, normal BMI (left) and high BMI (right) comparison.



female					
At risk	104	66	48	29	23
Censored	0	4	10	20	22
Events	0	34	46	55	59
male					
At risk	207	89	46	29	22
Censored	0	12	21	26	30
Events	0	106	140	152	155

female					
At risk	115	70	53	38	24
Censored	0	11	16	24	33
Events	0	34	46	53	58
male					
At risk	229	127	90	64	45
Censored	0	25	35	40	52
Events	0	77	104	125	132

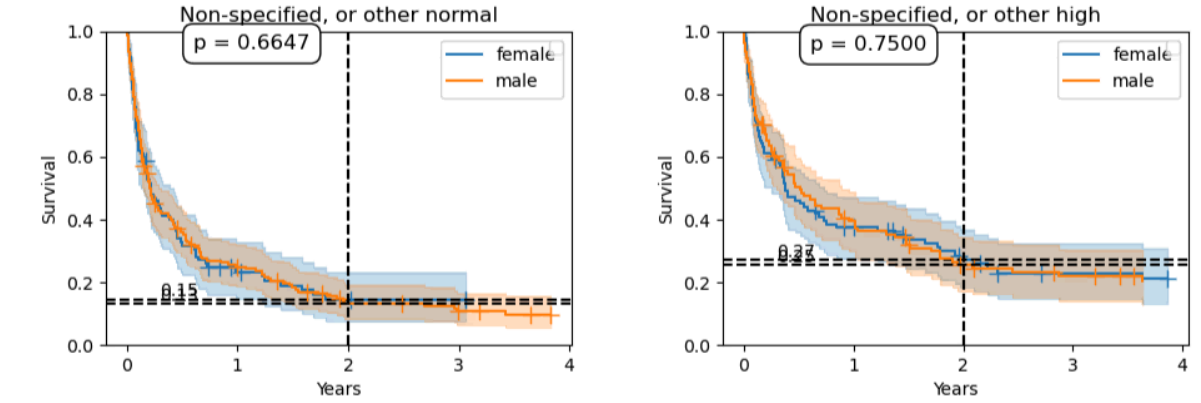
Supplementary Figure 6 Kaplan-Meier survival plot for patients with squamous cell carcinoma histology stratified by sex , normal BMI (left) and high BMI (right) comparison.



female					
At risk	132	41	22	14	10
Censored	0	2	3	6	10
Events	0	89	107	112	112
male					
At risk	145	34	10	7	6
Censored	0	5	6	6	7
Events	0	106	129	132	132

female					
At risk	132	49	21	14	12
Censored	0	6	12	12	12
Events	0	77	99	106	108
male					
At risk	175	48	17	9	7
Censored	0	7	10	11	12
Events	0	120	148	155	156

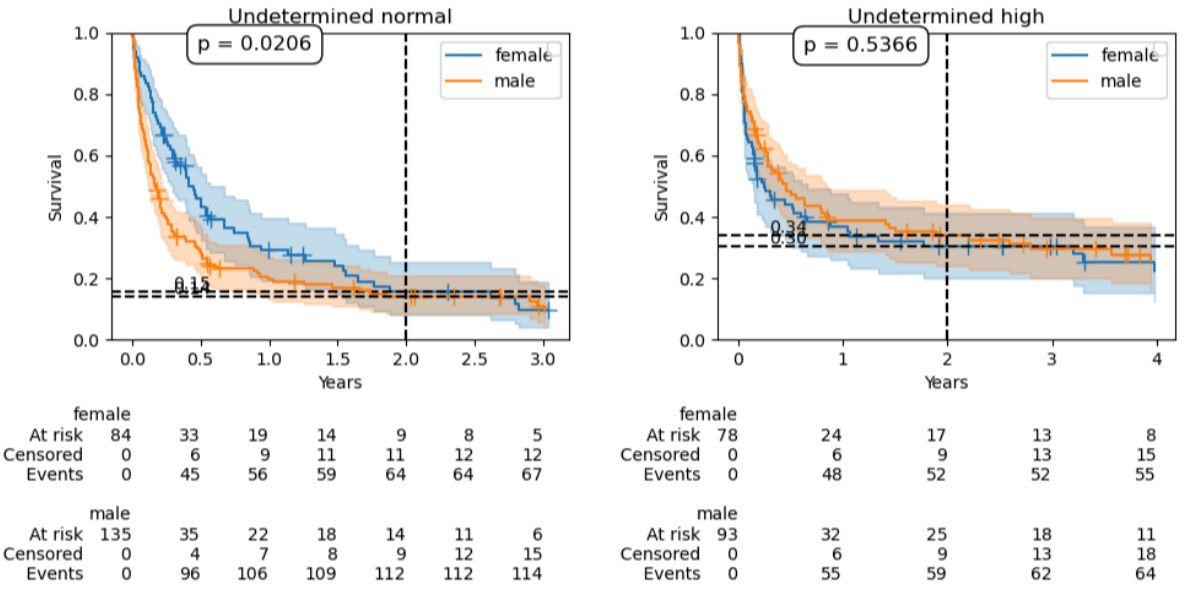
Supplementary Figure 7 Kaplan-Meier survival plot for patients with non specified or other histology stratified by sex , normal BMI (left) and high BMI (right) comparison.



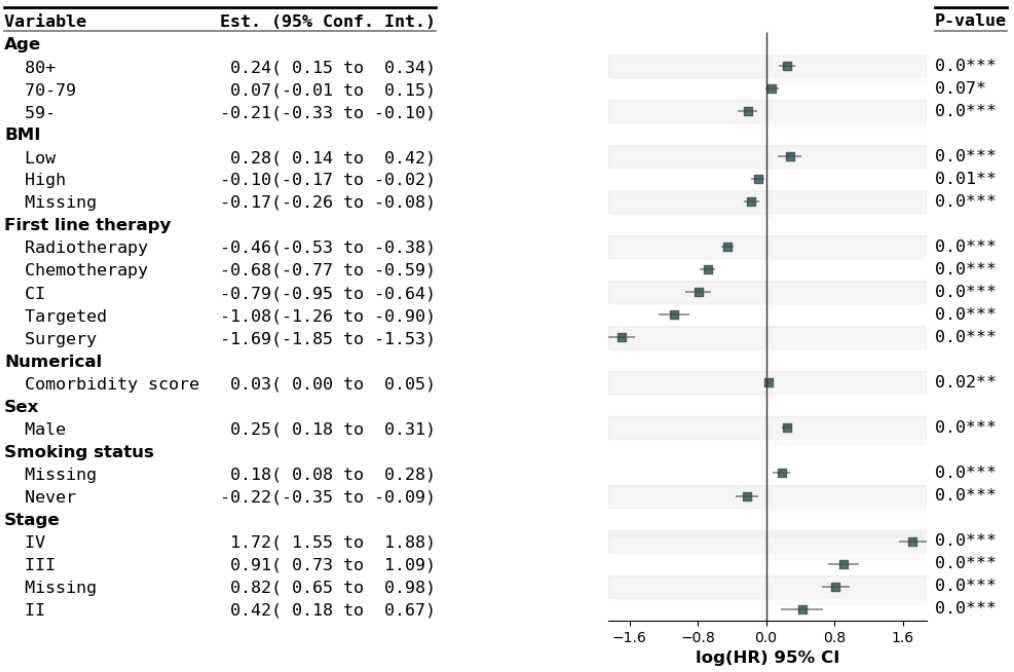
female					
At risk	87	17	10	9	8
Censored	0	4	5	6	7
Events	0	66	72	72	72
male					
At risk	168	38	17	12	8
Censored	0	6	10	12	15
Events	0	124	141	144	145

female					
At risk	95	33	20	14	12
Censored	0	3	8	11	12
Events	0	59	67	70	71
male					
At risk	107	38	22	17	13
Censored	0	6	9	11	14
Events	0	63	76	79	80

Supplementary Figure 8 Kaplan-Meier survival plot for patients with small cell lung cancer stratified by sex , normal BMI (left) and high BMI (right) comparison.

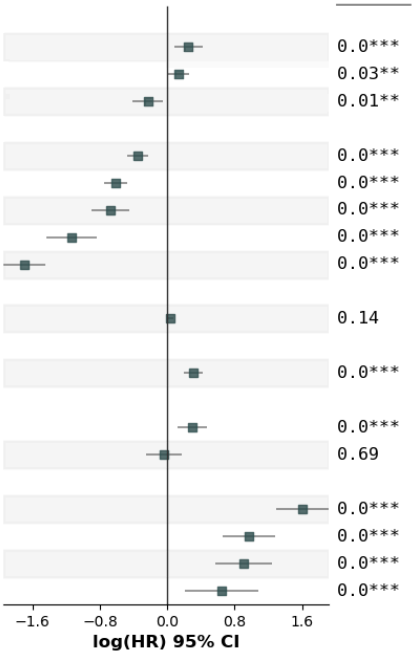


Supplementary Figure 9 Kaplan-Meier survival plot for patients with undetermined histology lung cancer stratified by sex , normal BMI (left) and high BMI (right) comparison.



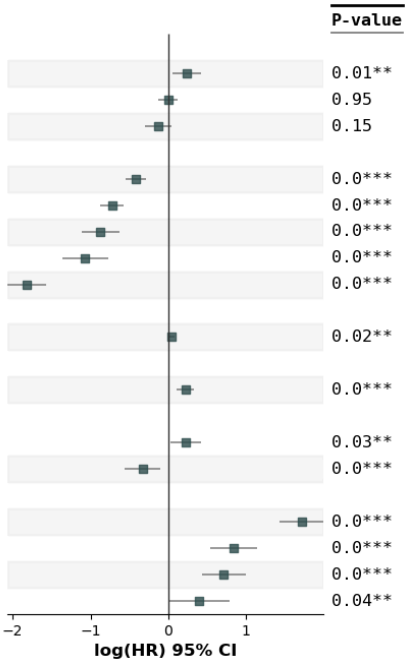
Supplementary Figure 10. Cox regression for the full cohort. Abbreviations NSCLC other (Non small cell lung cancer, other); NS, or other (Non-specified or other); SCLC Small cell lung cancer; CI - checkpoint inhibitor therapy; Conf. Int - confidence interval

Variable	Est. (95% Conf. Int.)	P-value
Age		
80+	0.25(0.08 to 0.42)	0.0***
70-79	0.14(0.01 to 0.26)	0.03**
59-	-0.23(-0.42 to -0.05)	0.01**
First line therapy		
Radiotherapy	-0.35(-0.47 to -0.23)	0.0***
Chemotherapy	-0.61(-0.75 to -0.48)	0.0***
CI	-0.68(-0.90 to -0.45)	0.0***
Targeted	-1.14(-1.44 to -0.84)	0.0***
Surgery	-1.70(-1.94 to -1.45)	0.0***
Numerical		
Comorbidity score	0.03(-0.01 to 0.07)	0.14
Sex		
Male	0.31(0.20 to 0.41)	0.0***
Smoking status		
Missing	0.30(0.12 to 0.47)	0.0***
Never	-0.04(-0.25 to 0.17)	0.69
Stage		
IV	1.60(1.29 to 1.92)	0.0***
Missing	0.97(0.65 to 1.28)	0.0***
III	0.90(0.57 to 1.24)	0.0***
II	0.64(0.20 to 1.08)	0.0***



Supplementary Figure 11. Cox regression for the Normal BMI cohort. Abbreviations CI - checkpoint inhibitor therapy; Conf. Int - confidence interval

Variable	Est. (95% Conf. Int.)	P-value
Age		
80+	0.24(0.05 to 0.42)	0.01**
70-79	-0.00(-0.13 to 0.12)	0.95
59-	-0.12(-0.29 to 0.05)	0.15
First line therapy		
Radiotherapy	-0.42(-0.55 to -0.29)	0.0***
Chemotherapy	-0.72(-0.87 to -0.57)	0.0***
CI	-0.87(-1.11 to -0.63)	0.0***
Targeted	-1.07(-1.36 to -0.77)	0.0***
Surgery	-1.82(-2.07 to -1.57)	0.0***
Numerical		
Comorbidity score	0.04(0.01 to 0.08)	0.02**
Sex		
Male	0.22(0.11 to 0.33)	0.0***
Smoking status		
Missing	0.23(0.03 to 0.43)	0.03**
Never	-0.33(-0.56 to -0.10)	0.0***
Stage		
IV	1.71(1.43 to 2.00)	0.0***
III	0.84(0.54 to 1.14)	0.0***
Missing	0.71(0.43 to 0.99)	0.0***
II	0.40(0.01 to 0.78)	0.04**



Supplementary Figure 12. Cox regression for the High BMI cohort. Abbreviations CI - checkpoint inhibitor therapy; Conf. Int - confidence interval