A collection of multiregistry data on patients at high risk of lung cancer—a Danish retrospective cohort study of nearly 40,000 patients

Høstgaard Bang Henriksen, Margrethe; Hansen, Torben Frøstrup; Jensen, Lars Henrik; Lohman Brasen, Claus; Peimankar, Abdolrahman; Ebrahimi, Ali; Wiil, Uffe Kock; Hilberg, Ole

Published in:
Translational Lung Cancer Research

DOI:
10.21037/tlcr-23-495

Publication date:
2023

Document version:
Final published version

Original Article

A collection of multiregistry data on patients at high risk of lung cancer—a Danish retrospective cohort study of nearly 40,000 patients

Margrethe Bang Henriksen[1]^, Torben Frøstrup Hansen[1]^, Lars Henrik Jensen[1]^, Claus Lohman Brasen[2]^, Abdolrahman Peimankar[3]^, Ali Ebrahimi[3]^, Uffe Kock Wiil[3], Ole Hilberg[4]^

[1]Department of Oncology, Vejle University Hospital, Vejle, Denmark; [2]Department of Biochemistry and Immunology, Vejle University Hospital, Vejle, Denmark; [3]SDU Health Informatics and Technology, Mærsk Mc-Kinney Møller Instituttet, University of Southern Denmark, Odense, Denmark; [4]Department of Internal Medicine, Vejle University Hospital, Vejle, Denmark

Contributions: (I) Conception and design: MB Henriksen, TF Hansen, LH Jensen, CL Brasen, O Hilberg; (II) Administrative support: MB Henriksen, TF Hansen, LH Jensen, CL Brasen, O Hilberg; (III) Provision of study materials or patients: MB Henriksen; (IV) Collection and assembly of data: MB Henriksen; (V) Data analysis and interpretation: MB Henriksen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Margrethe Bang Henriksen, MD. Department of Oncology, Vejle University Hospital, Beridderbakken 4, 7100 Vejle University Hospital, Vejle, Denmark. Email: margrethe.hostgaard.bang.henriksen@rsyd.dk.

Background: Lung cancer (LC) is the leading cause of cancer related deaths, and several countries are implementing screening programs. Risk models have been introduced to refine the LC screening criteria, but the use of real-world data for this task demands a robust data infrastructure and quality. In this retrospective cohort study, we aim to address the different relevant risk factors in terms of data sources, descriptive statistics, completeness and quality.

Methods: Data on comorbidity, prescription medication, smoking history, consultations, symptoms, familial predispositions, exposures, laboratory data among others were collected for all patients examined on a risk of LC over a 10-year period in the Region of Southern Denmark. Data were delivered from the regional data warehouse as well as the Danish Lung Cancer Registry. Associations between LC and non-LC groups were examined through Chi-squared test (categorical variables) and Wilcoxon signed-rank test (continuous variables that were non-parametric). These associations were investigated on both the original datasets and the subset of patients with complete data.

Results: The number of examined individuals increased over the study period and more patients were diagnosed with LC in stage I–II, from 18% in 2009 to 31% in 2018. LC patients were more likely to be older, smoker, with a registered prescription of the included medication. They also exhibited differences in laboratory analysis indicating inflammation and hyponatremia. Weight loss, fatigue and pain were more prevalent in the LC group, while hemoptysis and fever were more common among the non-LC patients. Advanced-stage LC patients experienced a higher rate of symptoms compared to those in the low stages. Within the sub-cohort with complete dataset results, most observed trends persisted, although data on comorbidities were susceptibility to change.

Conclusions: This study provides key insights into LC risk assessment using a robust dataset of patients examined for suspected LC. A consistent positive trend in early-stage LC diagnosis was observed throughout the study period. LC patients exhibited distinct smoking behaviors, medication patterns, variations in lab results, and specific symptoms. These discoveries have the potential to enhance discrimination in machine

^ ORCID: Margrethe Bang Henriksen, 0000-0002-1245-8874; Torben Frøstrup Hansen, 0000-0001-7476-671X; Lars Henrik Jensen, 0000-0002-0020-1537; Claus Lohman Brasen, 0000-0001-8654-2449; Abdolrahman Peimankar, 0000-0001-9779-9442; Ali Ebrahimi, 0000-0002-3332-6205; Ole Hilberg, 0000-0002-3075-3463.

learning-based prediction models, particularly those capable of handling complex distributions. Serving as a detailed account of real-world data collection and processing, the study establishes a foundation for future development of prediction models aimed at facilitating the early referral of LC patients.

Keywords: Lung cancer (LC); early diagnosis; risk prediction, real-world data; data collection

Submitted Sep 05, 2023. Accepted for publication Dec 07, 2023. Published online Dec 22, 2023.
doi: 10.21037/tlcr-23-495
View this article at: https://dx.doi.org/10.21037/tlcr-23-495

Introduction

Lung cancer (LC) accounted for 11.4% of all new cancer diagnoses in 2020, and 18% of all cancer related deaths. It is the second most frequently occurring cancer and the leading cause of cancer death globally (1). The main challenge of LC is late time of diagnosis, since patients with advanced or metastatic disease are not eligible for curative surgical treatment (2,3). The delay is often caused by a lack of symptoms or the presence of uncharacteristic symptoms such as a cough, which frequently occurs in the background population, thereby challenging the task of timely referral (4).

Screening for LC represents a partial solution to the problem and is gradually introduced in several countries based on different screening trials with a rate of improved mortality up to 20% depending on screening methods (5-7). In Denmark a pilot project, currently in the planning stage, will test LC screening on a Danish high-risk population. The experiences from the pilot project will eventually determine whether all heavy smokers will receive a similar offer (8). In the USA annual screening for LC with low-dose computed tomography has been offered since 2013, but screening rates remain low and the screening criteria mainly focus on high-risk individuals (9).

While several prediction models have been introduced to refine the screening criteria, most of them include a limited number of additional factors such as patient demographics, smoking history, chronic obstructive pulmonary disease (COPD), and heredity (10,11). Implementation of such models in clinical practice remains sparse, mainly due to inadequate external validation of models, poor methodological development with limitations in data accessibility, and limited exploration of potential clinical factors (12,13). Another barrier to adoption of clinical risk supportive tools is the lack of integration with electronic health records (EHR), where different clinical data are stored across multiple digital systems. Consequently, even though the amount of developed risk models is substantial, the evidence of successful clinical implementation is rare (13).

The Danish government-funded universal health care system combined with the Danish Civil Registration System as well as the broad spectrum of registries make Denmark ideal for data collection (14). The availability of updated nationwide registries provides the ability to join high quality administrative, health, and clinical data sources on an individual-linked level, enabling lifelong follow-up (15).

In this paper, we exploited the possibility of using regional copies of national registry data as well as free text to create an ideal database holding various information on patients examined on suspicion of LC. We assessed the completeness and validity of these datasets, and by

Highlight box

Key findings

- Lung cancer (LC) patients were associated with higher age, active or former smoking history, active prescription medication, weight loss, fatigue and several differences in laboratory results compared to the non-LC patients.

What is known and what is new?

- Prediction models aiming to refine LC screening are mostly based on small sample sizes or restricted data availability.
- We describe the collection and processing of a broad variety of real-world data on nearly 40,000 patients examined on suspicion of LC. Despite the similarities in risk factors, LC patients differed from non-LC patients in several aspects.

What is the implication, and what should change now?

- Differences in results that are often small and difficult to discern in the everyday clinic, could enhance the performance of prediction models able to flag patients according to their individual risk of LC.
- An accurate LC prediction model could be of use as a supportive tool in general practice, or incorporated into a LC screening program with the aim of increasing survival through early diagnosis.

combining relevant clinical information, we were able to characterize a large risk-cohort. This dataset is the first step towards the creation of individualized prediction models applicable in general practice for early detection of LC.

This study aims to address the following questions:

- ❖ Can we differentiate LC patients from non-LC patients based on clinical and laboratory data?
- ❖ Do we have sufficient data availability to create prediction models, and how do we handle and interpret missing data?
- ❖ Does the cohort with combined data reflect the relevant population at risk?

We present this article in accordance with the STROBE reporting checklist (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-495/rc) (16).

Methods

Data sources and selection criteria

This retrospective cohort study included all patients in the Region of Southern Denmark examined on suspicion of LC in the period 2009-01-01 to 2018-12-31 (*Figure 1*). They were defined using the two classification-codes: AFB26 (indicating the initiation of examinations in the LC fast-track clinic) and/or DZ031b (under observation for LC). The "Z-code" reports that is a tentative "obs pro" (observation for) diagnosis, and that the diagnosis is not yet confirmed (14). Both classifications are provided from the Health Care Classification System [Danish, Sundheds-væsenets Klassifikations System (SKS)], which is used throughout the Danish healthcare system (17). The date of the assigned SKS-code was referred to as the index date and used as reference point. If patients had multiple entries, the first referral date was chosen to be the one of interest.

The primary outcome was the diagnosis of LC, defined by the ICD-10 code C34 (Malignant neoplasm of bronchus and lung) (18). The Danish Lung Cancer Registry delivered data on all patients diagnosed with LC same period and region. The LC cohort was matched with the cohort with available SKS-codes, resulting in an additional 1,646 LC-patients, who had bypassed the LC fast-track clinic. These were added to the study cohort, with the date of the LC-diagnosis used as index date. Fifty-six patients were excluded based on missing information on sex due to temporary civil registry numbers from which sex cannot be derived. A total of 283 patients were excluded based on a prior diagnosis with LC before 2009-01-01. The final cohort consisted of 38,944 patients, i.e., 11,284 LC (29%) and 27,660 non-LC patients (71%). The types of data collected and joined in one common dataset were: details on LC diagnosis, comorbidity (medical diagnosis), prescription medication, smoking history, number of consultations with general practitioner, C-reactive protein (CRP) rapid tests, routine blood sample analysis, the presence 15 common symptoms, relevant exposures and familial predispositions to LC. All types of data were delivered from the regional data warehouse, which holds a copy of all data delivered to the national registries.

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Danish Data Protection Agency (No. 19/30673, Date: 2020-06-12) and the Danish Patient Safety Authority (No. 3-3013-3132/1, Date: 2020-03-30), and individual consent for this retrospective analysis was waived.

Comorbidity data

Registration of relevant comorbidity diagnoses was based on the corresponding ICD-10 codes available from Quan *et al.* (19). These reflected hospital encounters only and does not provide information on diagnoses registered in general practice (20). Consequently, a patient with mild or moderate COPD will most likely not be registered with this diagnosis at the hospital level. History of prescribed COPD-related medication may be used as a proxy of registration of COPD (14), which was the strategy applied in this study to partly overcome the issue of missing registrations of diagnosis from general practice.

ICD-10 codes included in the Charlson comorbidity index (CCI) from the Royal College of Surgeons, was used as reference in this study, since it has proven acceptable for use in registry-based research (21,22). A condition was noted present if registered within two years before the index date. A disease not registered within two years of the index date was considered unlikely to cause clinical impact at the time of diagnosis of LC. The registration of "Any malignancy" included in the CCI was adapted to "Other malignancies" with removal of LC (C34: Malignant neoplasm of bronchus and lung) from the ICD-10 codes in order to avoid bias in the LC-cohort (18). The supplementary tables hold a description of the different disease categories, corresponding ICD-10 codes and assigned weights (Table S1). Based on expert opinions, six additional pulmonary disease categories were included: Pneumonia, pulmonary tuberculosis, sarcoidosis, interstitial

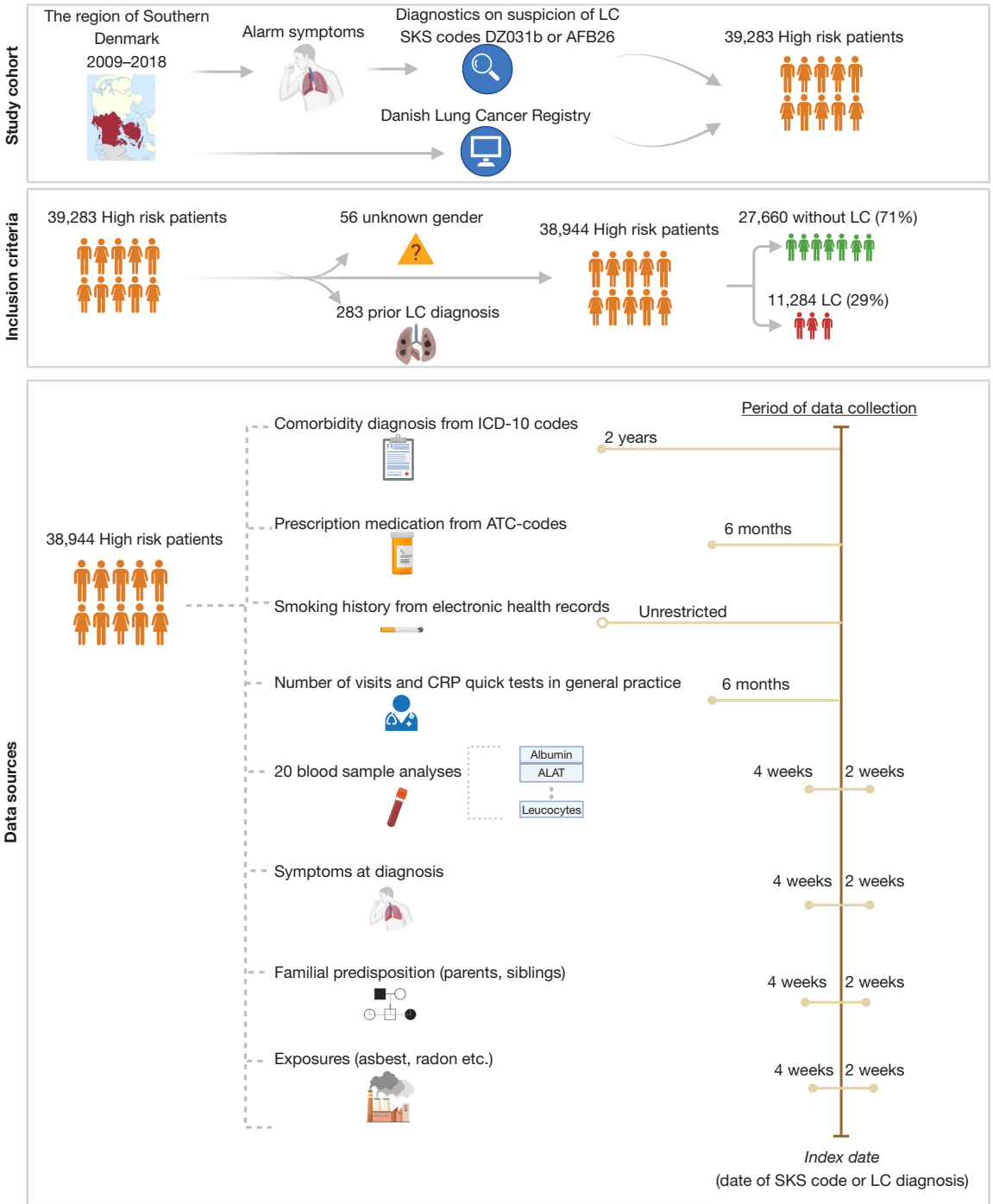

Figure 1 Flowchart of inclusion process of all patients examined on suspicion of LC, over the study period 2009-01-01 to 2018-12-31 in the Region of Southern Denmark. Created with Biorender.com. SKS, Sundheds-væsenets Klassifikations System; LC, lung cancer; AFB26, classification code for the initiation of examinations in the lung cancer fast-track clinic; DZ031b, classification code indicating an ongoing observation for LC; ICD-10, International Classification of Diseases 10th Revision; ATC, Anatomical Therapeutic Chemical classification; CRP, C-reactive protein.

lung disease, abscess and pleural disease. The disease category "Other malignancies" was defined by the ICD-10 codes C00-C97 (Table S2). All comorbidity data were transformed into binary format, with either the presence [1] or absence [0] of a condition within the 2-year period up to the index date.

Collection of prescription medication data

A panel of relevant drugs was included within six months up to the index date in order to reveal patterns of prescriptions in a close time interval preceding the LC diagnosis. Since symptoms of LC often overlap or coexist with symptoms of pneumonia and COPD, the most common antibiotics used to treat pneumonia as well as corticosteroids and inhalation devices used to treat COPD were included. Different antidepressants were also included as a proxy for depression and anxiety. A description of the different types of medication included and corresponding ATC-codes is attached in Table S3. All prescription data were transformed into binary format, with either the presence or absence of a prescription within the 2-year period up to the index date.

Collection of smoking data

Structured registration on smoking habit was only available on the LC patients, as they were registered in the Danish Lung Cancer Registry. Obtaining the information on the non-LC cohort as well required free text annotation of smoking status in the EHR. For comparability, smoking was manually annotated on the entire cohort, and to access validity results were compared with the registrations in the Danish Lung Cancer Registry. All EHR-notes including the sub-header "smoking" or "risk factors" (no limit as to time period) was extracted from the applied systems. Smoking data were labeled into binary categories: never smoker, current/former smoker, since information on pack-years was only available on a small subset of the entire cohort.

Data from general practice

The general practitioners in Denmark have a gate-keeping function and are paid through a combination of fees-for-service and capitation, which is funded by the healthcare region (23). Every consultation (with minor exceptions) is registered using a specific code, which results in an exact fee per consultation (24). Apart from the consultation itself, a large number of actions (e.g., CRP rapid tests) are also charged with specific codes for exact billing, and the quality of the codes is generally considered to be high (14). Patients under suspicion of LC often present with cough or dyspnea, and in these cases a CRP rapid test is often used as a an indicator of inflammation (25). The number of consultations as well as the number of CRP rapid tests registered within 6 months before the index date was included and analyzed both as continuous variables and categorical variables with a cut-off of 0 and 4 registrations.

Collection of blood sample data

Laboratory results were collected both from the current system in use (BCC, 2011–2018) and former system in use (LABKA, 2008–2011). The analyses performed in the LC fast-track clinics varied among departments and have changed over time. The combination of analyses used in the Diagnostics Department, Vejle Hospital, was chosen as relevant for further investigation, i.e., hemoglobin, sodium, potassium, lactate dehydrogenase (LDH), alanine transaminase (ALAT), CRP, creatinine, international normalized ratio (INR), calcium-total, albumin, amylase, bilirubin-total, alkaline phosphatase, counts of basophils, neutrophils, leucocytes, monocytes, lymphocytes, eosinophils and platelets. The results of all 20 blood sample analyses were requested on the entire study cohort for a period of 180 days before the index date until 14 days after. In addition, the neutrophil to lymphocyte ratio was also calculated. To capture the sample taken at the LC fast-track clinic, data were reduced to a maximum of 28 days before the index date and 14 days after, and only samples ordered by one of the four diagnostic departments were included. If case of multiple samples, the sample with the highest number of included analyses was chosen for examination. All laboratory results were analyzed as continuous variables.

Collection of symptoms, familial predispositions and exposures

Due to the unavailability of structured data concerning symptoms, familial predispositions, and relevant exposures, we included free text from the regional data warehouse. Due to the time-consuming task of manual annotation, this process was carried out on the subset of patients with complete data from all of the above-mentioned datasets. We collected all free text within a period of four weeks before to two weeks after the index date, and extracted only relevant notes obtained from the LC fast-track clinics.

Two medical students, under the supervision of a medical doctor, manually annotated the outpatient records. The annotated symptoms or conditions included hemoptysis, pneumonia, cough, dyspnea, fever, weight loss, fatigue, hot flashes, hoarseness, back pain, other pain, angina, headache, dizziness, and edema. We noted these present if they referred to the examination period and absent if referring to prior periods.

Familial predispositions for LC were noted present if the patient reported to have a sibling or parent with LC. Exposures were marked present if the patient had a history of working with radon, asbestos, nickel, chromium, aromatic hydrocarbons, and welding throughout their lifetime, regardless of the duration of exposure. All symptoms, familial predispositions and exposures were analyzed as binary variables with either the presence or absence of the specific condition.

Statistical analyses

The distribution of patient and data variables in relation to LC are presented as percentages (categorical variables) or median and interquartile range (IQR) (continuous variables). Associations between groups were examined through Chi-squared test (categorical variables) and Wilcoxon signed-rank test (continuous variables that were non-parametric). Bar-charts were used to display distributions in several categorical data, and significant difference between groups were indicated by a star and colored according to the group with the highest rate. All statistical tests were two-sided with a level of statistical significance set at P<0.01. The proportion of missing data was assessed independently for each variable, and evaluated if it was missing at random or systematically, potentially introducing selection bias. Subanalyses compared associations between groups in the cohort with complete data. All statistical analyses were performed in the Stata version 17.0.

Results

LC incidence and stage distribution

Of the 38,944 patients examined on suspicion of LC, 11,284 were diagnosed with the disease. The patients in the LC cohort were significantly older than those in the non-LC cohort [70 (IQR, 63–77) *vs.* 67 (IQR, 56–75) years, respectively, P<0.001] and included a larger proportion of females (48% *vs.* 45%, P<0.001). The LC incidence increased slightly over the study period from 1,093 in 2009 to 1,169 in 2018 (*Figure 2A*). Despite the increase in LC incidence, the proportion of LC patients decreased from 39% in 2009 to 25% in 2018, due to a higher number of patients examined over the study period. Twenty-six percent LC patients were diagnosed in stage I–II, 68% in stage III–IV, and 6% did not have available information on disease stage. The distribution of low-stage LC (stage I–II) increased from 18% in 2009 to 31% in 2018. The proportion of unknown or unregistered stages decrease from 13% in 2009 to only 1% in 2018 (*Figure 2B*).

Comorbidity

Exploration of missing results revealed that 45% of the cohort were not registered with any of the comorbidity-codes included in the 2-year interval. The initial cohort of 38,944 patients was defined by the presence of either the SKS code AFB26 or DZ031b. Consequently, the absence of the comorbidities was interpreted as absence of the specific diseases and not as missing observations. *Figure 3* shows the rate of comorbidities in both the LC and the non-LC cohort. The most common condition were other malignancies presented in a higher proportion of the non-LC cohort than the LC cohort (14.0 *vs.* 11.7, P<0.001). Other common conditions were COPD and pneumonia, with no statistical different between groups. The LC cohort presented a higher proportion of cerebrovascular disease, peripheral vascular disease and metastatic solid tumor compared to the non-LC cohort (P<0.01 for all). A CCI of 0 was registered for 62% of the LC cohort and 65% of the non-LC cohort (P<0.001). There was no significant difference in the proportion of late-stage LC (stage III–IV) among patients with comorbidities compared to patients without comorbidities (72% *vs.* 73%, P=0.11).

Figure 4 depicts the distribution of other malignancies in the total cohort. The most common malignancies were colorectal and breast cancer and significantly more frequent in the non-LC cohort than the LC cohort (2.3% *vs.* 1.4% and 2.0% *vs.* 1.4%, respectively, P<0.001 for all). In contrast, head-neck cancers, brain and esophagus-stomach cancers were more common among the LC group: 1.0 *vs.* 0.7 (P<0.01), 0.9 *vs.* 0.1 (P<0.01) and 0.3 *vs.* 0.2 (P=0.02). Investigation of a subset of patients with LC and brain-cancer showed that many patients were primarily diagnosed with brain cancer, but a subsequent analysis revealed a primary LC with brain metastasis. Analyzing a minor proportion of the esophagus-ventricle cancer patients did

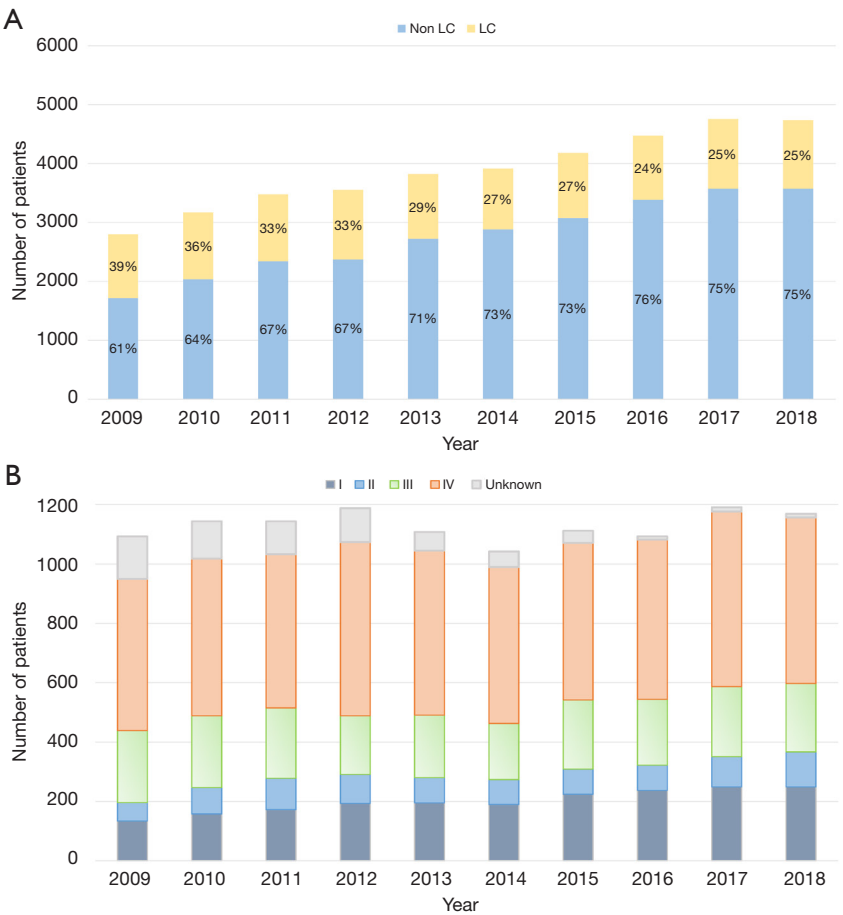

Figure 2 Lung cancer incidence (A) as well as distribution of LC stages I–IV during the study period (B). LC, lung cancer.

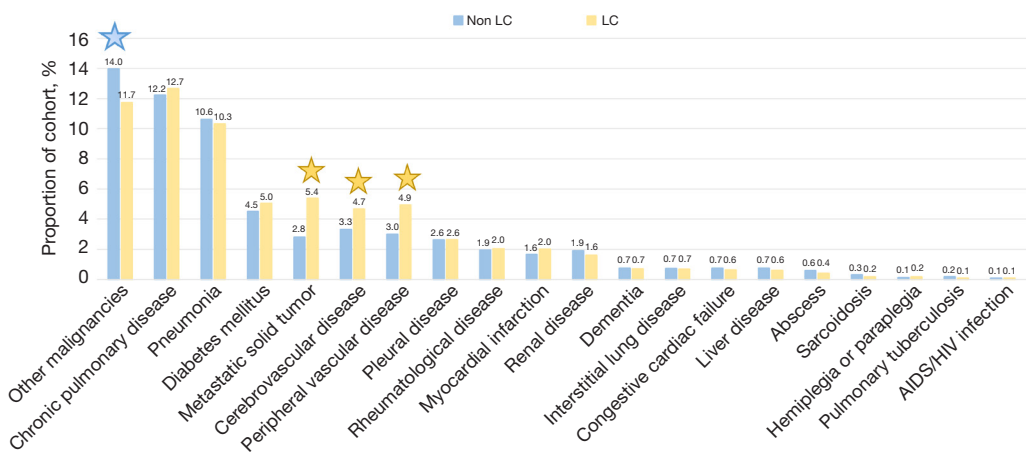

Figure 3 The distribution of comorbidities (%) in the lung cancer and non-lung cancer cohort. Significant difference between groups were indicated by a star and colored according to the group with the highest rate. All statistical tests were two-sided with a level of statistical significance set at P<0.01. LC, lung cancer; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus.

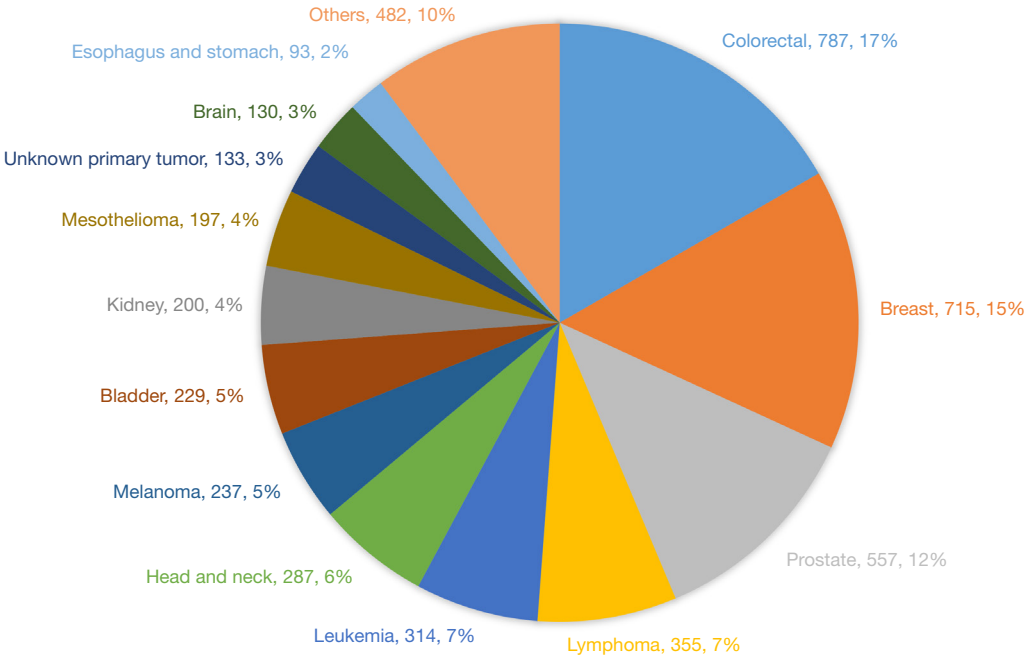

Figure 4 The distribution of other malignancies (%) in both lung cancer and non-lung cancer patients.

not show the same systematic bias.

Prescription medication

Exploration of missing data revealed that 27% of the total study cohort were missing in the dataset. The patients of this group were distributed over the total study period and were not exclusively from a specific region. They were generally younger and with a lower proportion of LC. Considering the high validity of this national registered data, together with the reasonable distribution of missing results, missing cases were included and labelled as absence of the specific drugs. Results were merged into binary variables with presence or absence of a prescription, and the fraction of patients with prescriptions in the LC and non-LC groups was compared (*Figure 5*). There was a significantly higher rate of LC patients with any prescription compared to non-LC patients (77.4% *vs.* 72.0%, P<0.001). Except for antibiotics, all drugs were prescribed to a higher proportion of LC than non-LC patients. The non-significant difference in antibiotics (P=0.93) correlates with the equal number of registrations of pneumonias in the two groups.

Smoking status

Notes containing the two relevant sub-headers ("smoking" and "risk factors") were available on 23,006 of the patients (60%). Of the total population 40% did not have available text-notes, either because no registrations were made containing the two keywords, or no text records were available. This group was largely represented in the first years of the study period, corresponding to the use of the EHR system, which was implemented in 2009 in the Region of Southern Denmark. Since no time limit was set on the data import, patients included in the latter years have a higher probability of accumulating EHR-notes than patients included close to the implementation of the system. Hence, the missing information on smoking status was concluded to be explained by the gradual implementation of the EHR-system. There was no logical way to impute missing variables since the remaining variables were not directly related to smoking status.

All patients were classified as never, former or current smokers. Duplicates were removed and only the exact note that corresponded to the assigned label was kept. Of the 23,006 patients with available text material, a higher rate of non-LC patients were found to be never smokers compared to the LC cohort (31.2 *vs.* 8.9, P<0.001). The rate of both former and current smokers was higher among LC patients compared to non-LC patients (58.8 *vs.* 43.0 and 34.2 *vs.* 26.0, respectively, P<0.001 for all). The proportion of patients with high-stage LC (III–IV) was not significantly

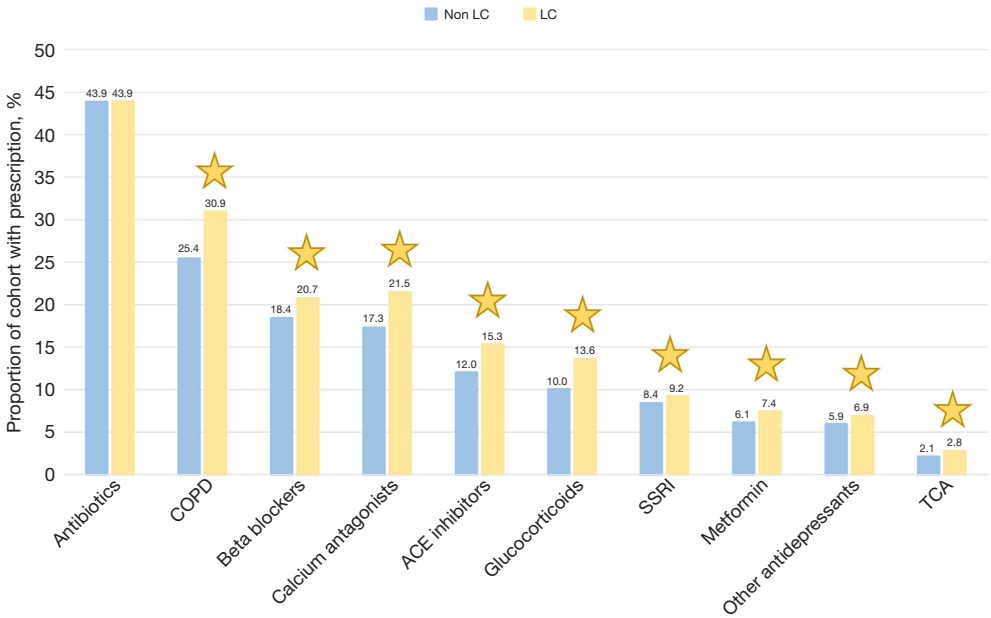

Figure 5 The distribution of prescription medication (%) in the lung cancer and non-lung cancer cohort. Significant difference between groups were indicated by a star and colored according to the group with the highest rate. All statistical tests were two-sided with a level of statistical significance set at P<0.01. LC, lung cancer; COPD, chronic obstructive pulmonary disease; ACE, angiotensin-converting enzyme; SSRI, selective serotonin reuptake inhibitors; TCA, tricyclic antidepressants.

different between the two cohorts, with 58% high-stage patients in the non-smoking cohort and 59% in the former/current cohort (P=0.45).

To validate the annotation of smoking status from the EHR, the distributions were compared with the Danish Lung Cancer Registry. Of all 11,284 LC patients, information on pack-years from the Danish Lung Cancer Registry was available on 83% (N=9,399). The two annotations were equivalent in 83% of the non-smoking cases, and 97% of the current/former smoking cases, which was overall considered to be an acceptable validity of this manual annotation from free-text.

Consultations and CRP rapid tests at the general practitioner

Exploration of missing data revealed that 10% of the total cohort did not have any registration of either consultation or CRP rapid test within the 6-month interval before the index date. Due to the high validity of the registrations in this dataset, missing data were included and labelled as being absent. *Figure 6A* depicts the distribution of consultations in general practice within 6 months up to the index date in the LC and non-LC cohort, respectively. The LC group had a significantly higher number of visits than the non-LC group, even though both had a median of 4 visits [4 (IQR, 2–6) *vs.* 4 (IQR, 2–7), P<0.001]. A higher rate of non-LC patients were absent from consultations (11.2% *vs.* 8.3%), or had only 1–4 consultations (47.2% *vs.* 46.8%) compared to the LC group. Conversely, a higher rate of LC patient had >4 consultations compared to the non-LC cohort (45.0% *vs.* 41.6%, P<0.001 for all).

Figure 6B depicts the distribution of CRP rapid tests in general practice within 6 months up to the index date in the LC and non-LC cohort, respectively. In both cohorts, 52% did not undergo a rapid test, and no clear difference was seen in the number of rapid tests performed between the LC and non-LC group (P=0.074). The proportion of high-stage LC (stage III–IV) patients was significantly higher in the group with CRP rapid tests performed (75% and 70% with and without rapid test, P<0.01).

Blood sample analyses

A total of 34,129 patients were represented to some extent within the 180 days before and 14 days after the index date (*Figure 7*). Of these patients, 18,462 had results within the 28 to 14 days around index date, ordered by one of the four

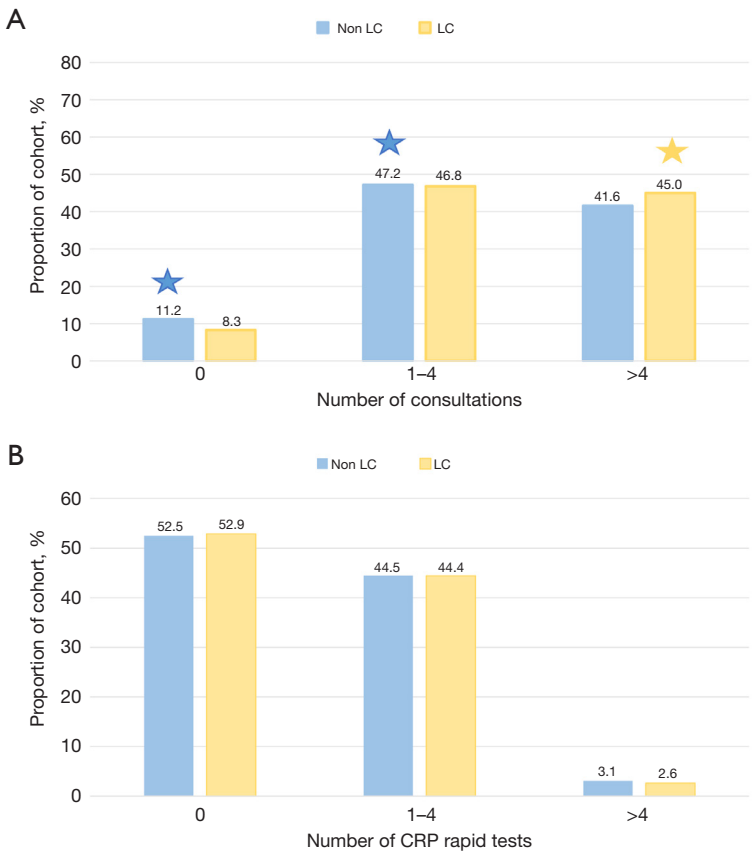

Figure 6 Distribution of consultations (A) and CRP rapid tests in general practice 6 months up to the index date (B). Significant difference between groups were indicated by a star and colored according to the group with the highest rate. All statistical tests were two-sided with a level of statistical significance set at P<0.01. CRP, C-reactive protein; LC, lung cancer.

diagnostic departments, 14,957 had ≥17 analyses present. The proportion of LC patients (30%) and non-LC patients (70%) did not change much compared to the initial cohort (29% and 71%, respectively). *Table 1* provides summary statistics of all 21 analyses on the cohort with ≥17 analyses available (14,957 patients). Minor differences were found between the two groups, even though median values were inside the reference intervals. Most of the white blood cells (leucocytes, neutrophils, monocytes) as well as platelets, calcium, CRP, LDH and alkaline phosphatase were significantly elevated among the LC patients compared to non-LC patients. Conversely, hemoglobin, eosinophils, lymphocytes, albumin, ALAT, creatinine and sodium were decreased among the LC patients compared to the non-LC patients. The neutrophil-to-lymphocyte ratio was significantly higher in the LC group compared to the non-LC group (3.4 and 2.6, P<0.001).

Symptoms, familial predispositions and expositions

Out of the 9,940 patients with complete data across the mentioned datasets, outpatient records from LC fast-track clinics were accessible for 5,587 individuals. The distribution of symptoms, familial predispositions, and relevant exposures in both LC and non-LC cohorts is illustrated in *Figure 8*. No symptoms were reported for 10% of the LC cohort, compared to 13% of the non-LC cohort (P=0.002). Predominant symptoms across groups included cough (53.4%), dyspnea (36.3%), weight loss (25.2%), fatigue (19.9%), and hemoptysis (16.2%). LC patients exhibited higher prevalence of weight loss, fatigue, back pain, and other pains, while hemoptysis and fever were more prevalent in the non-LC cohort (P<0.001 for all). Familial predispositions to LC were observed in 9.0% of LC patients compared to 6.8% in the non-LC cohort

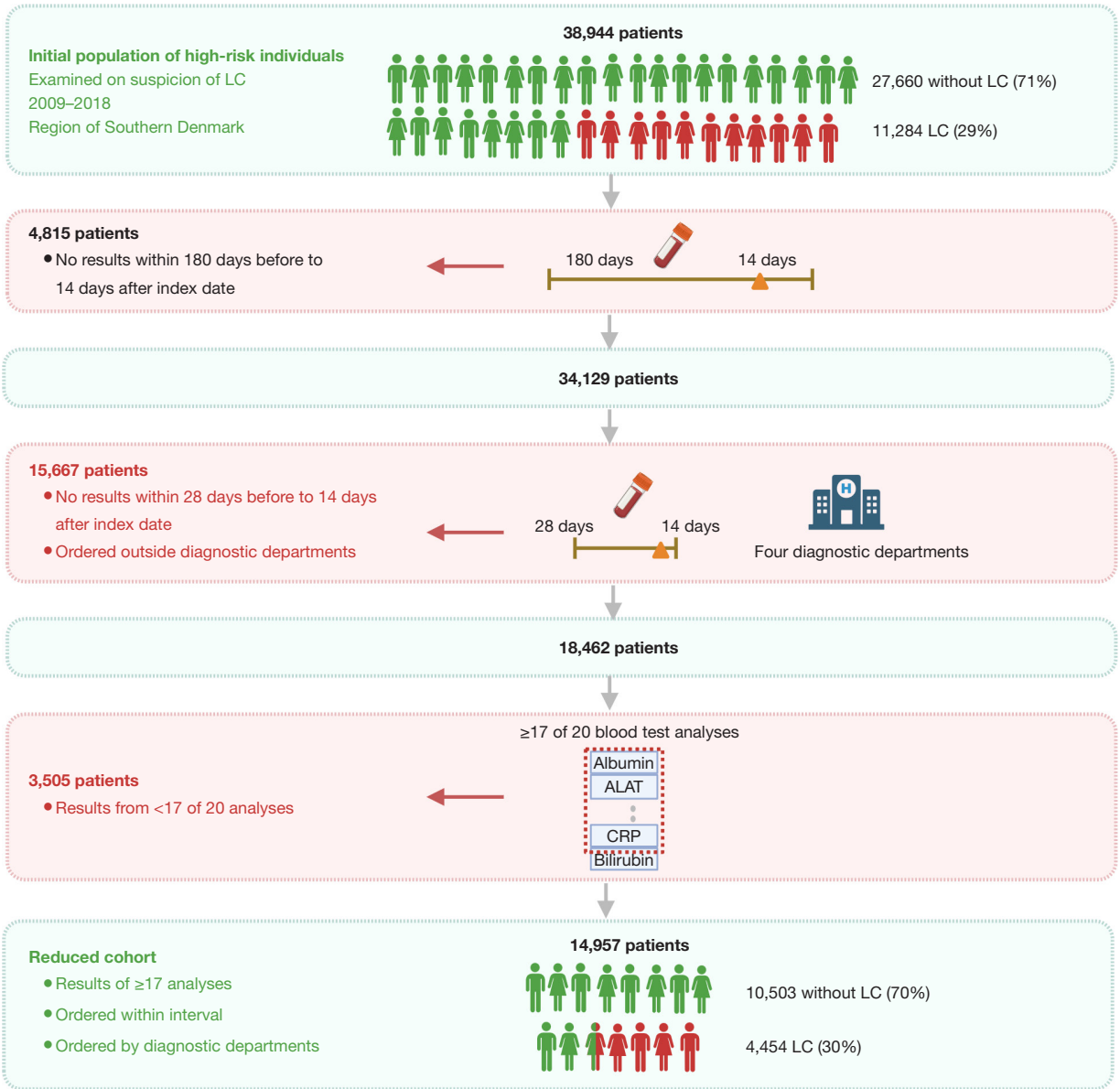

Figure 7 Population with relevant blood sample analyses, reduced due to relevant filtering. Created with Biorender.com. LC, lung cancer; ALAT, alanine transaminase; CRP, C-reactive protein.

(P=0.003). Exposures to LC were present in 20.4% of all patients, with no significant difference between the two groups (P=0.091).

Within the LC patient subset, 17.4% of the low-stage LC patients exhibited no symptoms, compared to 5.4% of the high-stage LC patients (P<0.01). Cough, dyspnea, weight loss, fatigue, hoarseness, back pain, other pain and angina were more commonly observed among high-stage LC patients (P<0.01 for all).

Combined data availability and comparison of cohorts

Figure 9 depicts the reduction from the total cohort to the final cohort after merging of datasets. Missing registration of patients was included and indicated as 0 for data on prescription medication, general practice and comorbidities. In the blood sample datasets only patients with ≥17 analyses present were included, and imputation based on the median was allowed for the remaining three analyses. Hence, the dataset was reduced to 14,957 patients. Merging this cohort

Table 1 Summary statistics of the 21 blood sample analysis in the LC and non-LC cohort

Variable	Reference interval	LC (n=4,454), median [IQR]	Non-LC (n=10,503), median [IQR]	P value
B-hemoglobin, mmol/L	Male: 8.3–10.5, female: 7.3–9.5	8.40 [7.7–9.0]	8.7 [8.0–9.3]	<0.001
B-leucocytes, 10^9/L	3.5–8.8	9.12 [7.43–11.20]	7.64 [6.20–9.46]	<0.001
B-neutrophils, 10^9/L	1.5–7.5	6.10 [4.71–7.95]	4.70 [3.58–6.20]	<0.001
B-lymphocytes, 10^9/L	1.0–4.0	1.74 [1.30–2.27]	1.81 [1.39–2.33]	<0.001
NLR	1–2	3.4 [2.4–5.2]	2.6 [1.8–3.8]	<0.001
B-monocytes, 10^9/L	0.2–0.8	0.76 [0.59–0.97]	0.65 [0.51–0.84]	<0.001
B-basophils, 10^9/L	<0.02	0.04 [0.02–0.06]	0.04 [0.02–0.06]	<0.001
B-eosinophils, 10^9/L	<0.05	0.14 [0.07–0.25]	0.17 [0.10–0.27]	<0.001
B-platelets, 10^9/L	Male: 145–350, Female: 165–390	311 [250–391]	272 [223–334]	<0.001
P-albumin, g/L	34–45	42 [39–44]	43 [41–45]	<0.001
Total calcium, mmol/L	2.15–2.51	2.36 [2.29–2.43]	2.34 [2.27–2.41]	<0.001
P-CRP, mg/L	<6	9.9 [3.0–32.0]	3.7 [1.4–10.0]	<0.001
P-ALAT, U/L	Male: 10–70, female: 10–45	18 [13–26]	22 [16–31]	<0.001
P-LDH, U/L	115–255	214 [182–257]	192 [169–221]	<0.001
P-alkaline phosphatase, U/L	35–105	83 [68–102]	75 [62–92]	<0.001
P-bilirubin-total, µmol/L	5–25	7 [5–9]	7 [6–10]	<0.001
P-amylase (pancreatic), U/L	10–65	25 [18–34]	25 [19–34]	0.79
P-INR	<1.2	1.0 [0.9–1.1]	1.0 [0.9–1.1]	<0.001
P-creatinine, mmol/L	Male: 60–105, female: 45–90	72 [60–87]	76 [64–89]	<0.001
P-sodium, mmol/L	137–145	139 [136–141]	140 [138–142]	<0.001
P-potassium, mmol/L	3.5–4.4	4.0 [3.8–4.3]	4.0 [3.8–4.3]	0.08

The number of digits reported on the blood test results reflects the number of digits provided by the laboratory. LC, lung cancer; IQR, interquartile range; P-, Plasma; B-, Blood; ALAT, alanine aminotransferase; CRP, C-reactive protein; INR, international normalized ratio; LDH, lactate dehydrogenase; NLR, neutrophil to lymphocyte ratio.

with the smoking data reduced the sample size further to 9,940 patients with blood sample results as well as smoking information. Free text from the electronical health records registered at the LC fast-track clinics were accessible on a subset of patients. Consequently, the reduced cohort with complete information consisted of 5,587 individuals of whom 1,854 had LC (33%) and 3,733 did not (67%).

Table S4 illustrates the variable distribution across LC status for the 5,587 patients with complete data. The proportion of LC patients increased from 29% in the initial cohort to 33% in the reduced cohort. Like the initial cohort, LC patients in the reduced cohort were significantly older and had a higher proportion of females compared to the non-LC cohort (P<0.001). In the reduced cohort, patients exhibited a lower degree of comorbidity in the LC group, with a CCI of 0 in 72%, as opposed to 62% in the initial cohort. While the overall frequencies remained similar between the initial dataset and the reduced dataset, several conditions remained rare, and their significance between the LC and non-LC group diminished when

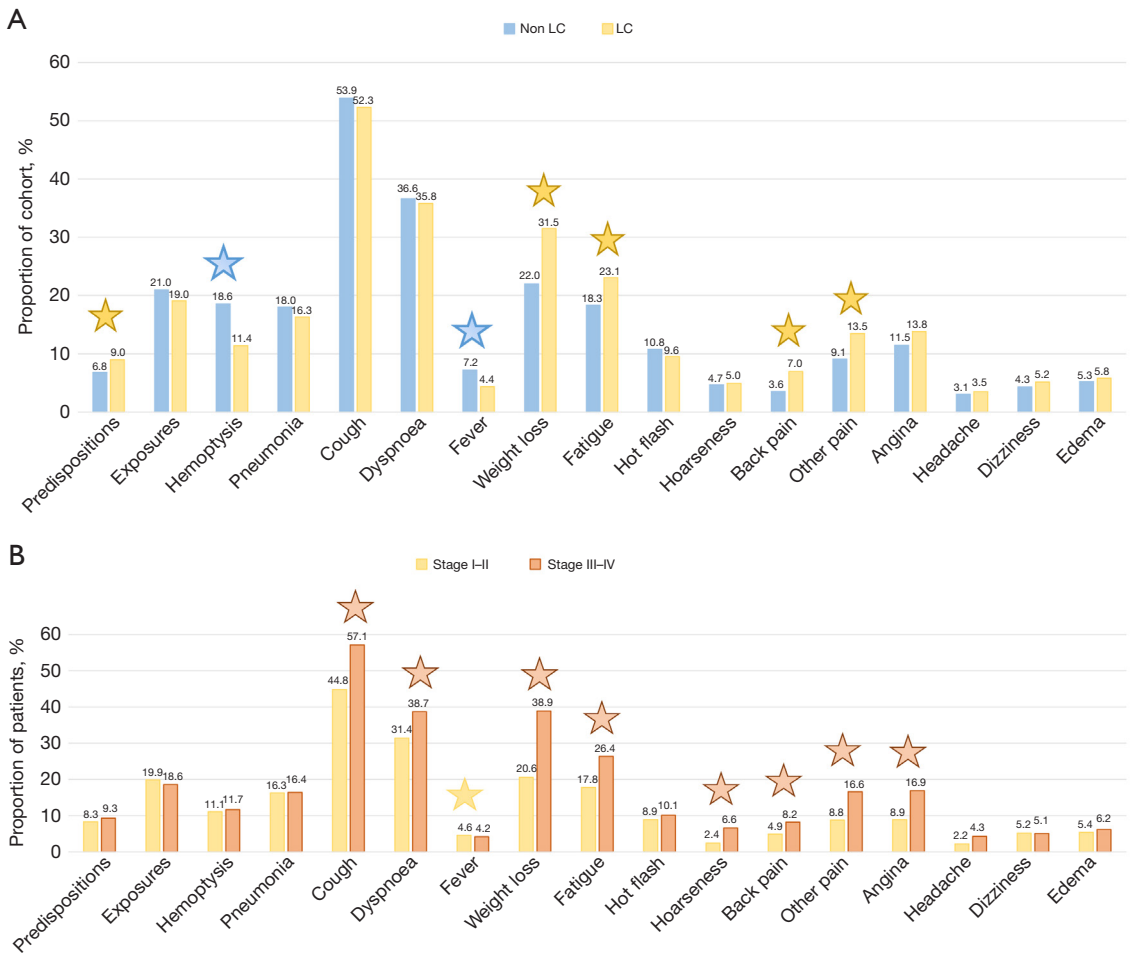

Figure 8 Symptoms, familial predispositions and exposure patterns among individuals based on lung cancer status and stage. The distribution of symptoms, familial predispositions and exposures in the lung cancer and non-lung cancer cohort (A). Analysis of stage I–II *vs.* III–IV lung cancer patients within each category (%) (B). Significant difference between groups were indicated by a star and colored according to the group with the highest rate. All statistical tests were two-sided with a level of statistical significance set at P<0.01. LC, lung cancer.

examined in the reduced dataset. The only exception was other malignancies, which exhibited a significant change like that in the initial cohort. Likewise, frequencies in prescribed medication were also similar between the two datasets, with 75.5% of the LC patients with any prescription compared to 69.5% of the non-LC patients (P<0.001). However, when comparing specific drugs, changes were only significant for COPD-related prescriptions and calcium antagonists. Like the initial dataset, a higher rate of LC patients had >4 consultations in general practice, and a higher rate of non-LC patient did not have any consultations. The fractions of patients with CRP rapid tests remained similar in the reduced cohort. Consistent with the prior findings, LC patients in the reduced cohort were more frequently current and former smokers compared to the non-LC cohort (P<0.001). The significance pattern and distributions in laboratory results remained overall consistent in the reduced cohort.

In summary, when comparing the reduced cohort to the initial cohorts for each dataset, the overall patterns between the LC and non-LC groups were similar in terms of age, sex, smoking status, consultations, CRP rapid tests and laboratory values. However, specific patterns in comorbidity data or prescription medication were susceptible to changes as the cohort was reduced to the 5,585 patients with complete results.

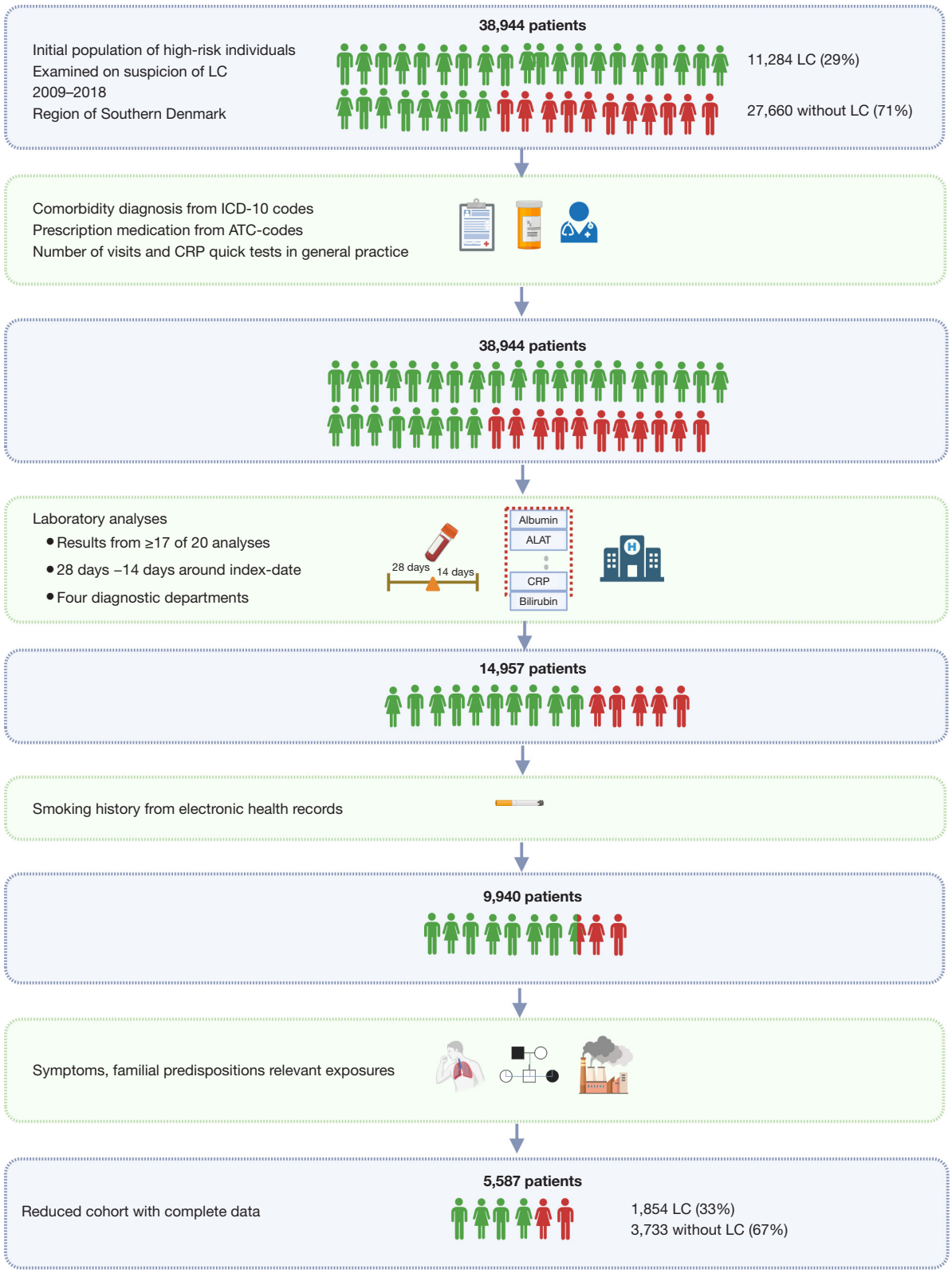

Figure 9 Data completeness after combination of datasets. Created with Biorender.com. LC, lung cancer; ICD-10, International Classification of Diseases 10th Revision; ATC, Anatomical Therapeutic Chemical classification; ALAT, alanine transaminase; CRP, C-reactive protein.

Discussion

Summary of findings

In this study, we have presented an extensive array of clinical and laboratory data encompassing all patients evaluated for the risk of LC over a 10-year period in the Region of Southern Denmark. We described datasets and compared variables across LC status. After consolidating the different datasets, we compared distributions in the reduced dataset to those in the initial datasets.

Throughout the study period, 39,283 individuals underwent examination on suspicion of LC, with 29% receiving a LC diagnosis and 71% not diagnosed with LC. The number of individuals examined increased over the study period, and the proportion of early-stage LC rose from 18% (stages I–II) in 2009 to 31% in 2018.

Common comorbidities in both the LC and non-LC groups included other malignancies, COPD, and pneumonia, but the vast majority were not diagnosed with any of the included diseases at a hospital level. LC patients were more likely to be smokers, and to be prescribed medication, specifically COPD-related medications and calcium antagonists when considering the reduced cohort.

Laboratory results revealed discrete differences, with significantly higher values for most white blood cells, platelets, calcium, CRP, LDH, and alkaline phosphatase in LC patients. Conversely, LC patients exhibited lower values of hemoglobin, albumin, ALAT, and creatinine, among others.

The majority of both cohorts displayed symptoms at the time of examination, with cough, dyspnea, weight loss, and fatigue being the most common. While weight loss, fatigue, and pain were more prevalent in the LC group, hemoptysis and fever were more common among the non-LC group. Advanced-stage LC patients experienced a higher rate of symptoms compared to those in the low stages.

Comparison with similar research

In this study, we found an increase in low-stage LC diagnosis over the study period, aligning with a gradual increase at a national level (26). This increase is, in part, due to the increased number of patients examined in the LC fast-track clinics, which surged by more than one third from 2012 to 2018 (27) on a national scale. This increase in referrals is most likely caused by an increased focus on cancer as an acute condition after introduction of the Danish Cancer Patient Pathways in 2007, which reduced the delay in diagnoses (28). Additionally, in 2007, there was an extension allowing general practitioners to refer patients directly for CT scans in cases of vague symptoms suggestive of cancer. The rise in referrals for LC fast-track clinics and the direct path to CT scans from general practice contribute to an increased frequency of CT scans, a factor that has been demonstrated to lead to a stage shift toward early-stage LC (29,30).

As opposed to the literature, we found no difference in COPD registrations between the two groups. One reason could be potential bias in the registration, since the results only account for hospital-based diagnoses. COPD related medication was used as a proxy indicator of mild-moderate COPD, and these drugs were more frequently prescribed to the LC cohort compared to the non-LC cohort. Another consideration is the initial selection of the study cohort, which has an overall high prevalence of COPD (≈12%). This is more than double the prevalence of the background population, where 5% of the 50–80 years old were diagnosed with COPD in 2018 (31). If the LC cohort were compared to healthy controls matched on age and sex, the prevalence of COPD would most likely be lower, potentially revealing a difference among the groups. Since this cohort aimed to include relevant patients at risk of LC, they already shared some of the same comorbidities.

In this study, the rate of patients with pneumonias were similar between the two groups. Previous studies have demonstrated an increased risk of LC in patients with immunosuppression or chronic inflammation and have also suggested bacterial infections to be an independent risk factor (32). In contrast, an Italian study of 2,100 LC patients and 2,120 controls found that 85% of the examined cohort did not have a history of pneumonias. Actually, it demonstrated a decreased risk of LC to be associated with the number of pneumonias (33). Hence, this field remains complex and warrants further research.

No comorbidities included in the CCI were observed in 62% of the LC cohort and 65% of the non-LC cohort (and even 72% and 65% in the reduced cohort). A Danish study examining 461 patients referred for diagnostics on suspicion of LC revealed that 42% scored 0 in the CCI (34). Tammemagi *et al.* extracted data on 56 comorbidities from 1,155 LC patients and found that in 11.7% of all patients no comorbidity was observed in the months adjacent the examination of LC (35). Even though these comorbidities were extensive compared to the present study, we still have a surprisingly high rate of patients without registered comorbidities. This may imply that risk population does not

only include the "usual suspects" such as patients registered with COPD, cardio-vascular disease or diabetes as expected from the literature, but also a proportion of patients without previous comorbidity registered at the hospital level. If we were able to include codes from International Classification of Primary Care (ICPC), we would assume the registration of milder conditions such as diabetes and COPD to increase.

We found a tendency towards a higher degree of consultations the general practitioner within the 6 months up to index date. A British study found that LC patients were more likely to have had three or more consultations with the general practitioner before referral (36). Comparison across countries, however, can easily be biased due to differences in systematic registrations as well as organization of general practice and referral patterns.

The differences in laboratory values were mostly small, and is doubtfully considered to be clinically significant. However, these differences might contribute to a better discrimination and performance in a predictive model. Many of these minor changes are known from the literature: the elevation in white blood cells and CRP corresponds with several studies linking elevated inflammatory parameters to an increased risk of LC (37,38). Elevated levels of LDH corresponds with increased cell metabolism and is common in most cancer types (39), whereas increased levels of calcium and alkaline phosphatase are linked to bone and liver involvement in metastasized cancer (40). Around 38% of LC patients present with anemia (41). Low albumin indicates malnutrition and has been associated with poor survival in several cancer types (42). Low sodium is the most frequent electrolyte imbalances seen in patients with LC, especially those with small-cell disease (41,43). A recent study that used blood samples to prediction LC, and also found that elevation of leucocytes were among the most important risk factors of LC, while increased platelets and calcium were further down the list (44). Furthermore, the neutrophil-to-lymphocyte ratio was pathologically high among the LC patients. This is usually associated with poor prognosis in LC patients, but have also been associated with increased risk of LC, suggesting that a systemic immune response may be an important pre-clinical element in the development of LC (45-47).

The most common symptoms observed across both groups were caugh and dyspnø. Ruano-Raviña *et al.* conducted a nationwide study examining symptoms at LC diagnosis in Spain and found the same two symptoms at the top with 34% LC patient with cough and 27% reporting dyspnea (48). They found that 59% of patients in LC stage I presented no symptoms, compared to 21% in our current study. Among stage IV patients, 28% had no symptoms in their study, compared to only 5% in our study. Despite difference in results, this emphasizes the need to consider other criteria than the most common symptoms for LC examinations, since a substantial amount of early-stage patients displays no symptoms at diagnosis.

Methodological considerations

This study displays an extensive dataset that includes clinical and laboratory data from all patients evaluated for the risk of LC over a 10-year period in the Region of Southern Denmark, providing a comprehensive overview. It involves comparison of variables across LC status, offering insights into the differences and trends within the studied population. It combines multiple data sources from registry-based data, laboratory data to data obtained from manual annotation of free text (smoking and symptoms, etc.). The careful consideration of handling missing data, combination of datasets and description of the reduced dataset with complete analyses, provides insight to distributions within this subset of patients, and the overall robustness of the patterns observed in the initial datasets.

However, this study also has several limitations. Using the two SKS-codes as inclusion criteria might be associated with certain bias, since inter-site differences exists both in the patterns of referral as well as in the registration of the two SKS-codes. At some hospitals patients are referred directly to the LC fast-track clinic without a prior CT-scan, while other hospitals use the department of radiology as a filtering function and hence only make the referral in case of an abnormal CT-scan (49). Consequently, the fraction of LC-patients varies among hospitals depending on the referral patterns. Using a more uniform inclusion criteria would ease comparison among hospitals; however, this was not an option at the time of inclusion.

Data on comorbidity was drawn from the ICD-10 codes registered at a hospital level. A database of diagnoses from general practice was established for a brief period in 2007, but it was closed due to uncertainty about legislative controversies (20). In future research, it would be ideal to include data on, e.g., comorbidity and symptoms from the ICPC-codes in general practice.

A major limitation to this study is the lack of detailed

smoking information on the broad cohort. Information on pack-years was only available on the LC population, and not registered on the non-LC population. The manual annotation of smoking status based on free-text limited the cohort to 23,006 patients with available smoking status in the level of detail former/current/never smoker. To obtain detailed information on smoking status on the general population we need large questionnaires on a national level. This was part of the recruitment methods for the NLST and NELSON screening trial (50,51).

An examination of procedure codes registering the number of X-rays and CT scans was also performed. We evaluated information on the referring entity and the exact text of the referral. Unfortunately, manual evaluation of a subset of patients revealed that this information contained multiple biases and lack of registrations. In addition, differences in referral patterns between hospital units also challenges comparison of such procedures, and this data source was ultimately excluded.

It was an initial interest to provide data on the socio-economic status on the study population. Such data are stored with Statistics Denmark, and they are available through external electronic access only under certain conditions (52), and they are not easily exportable. Such problematic accessibility would challenge the use-case in general practice where a decision support tool should be able to calculate the risk of LC based on easily obtainable data.

Conclusions

This study provides key insights into LC risk assessment using a robust dataset of patients examined for suspected LC. The rising trend in early-stage diagnoses (18% to 31%) reflects a positive development. Common comorbidities were identified, mostly lacking hospital-level diagnoses. LC patients showed distinct medication patterns, and significant differences in lab results were noted. Symptomatology revealed distinct features, with weight loss, fatigue, and pain more prevalent in the LC group, while hemoptysis and fever were more common in the non-LC group. These findings may overall contribute to improved discrimination and predictive model performance. Serving as a stepwise account of real-world data collection and processing, this study lays the groundwork for future development of prediction models utilizing similar data sources.

Acknowledgments

Funding: The study was supported by research grants from the Region of Southern Denmark, the University of Southern Denmark, the Danish Cancer Society, the Dagmar Marshall Foundation, the Beckett Foundation, the Lilly and Herbert Hansen Foundation, and the Hede Nielsen Family Foundation. The funding bodies did not have any influence on the design of the study, collection, analysis, interpretation of data or writing of manuscript.

Footnote

Reporting Checklist: The authors have completed the STROBE reporting checklist. Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-495/rc

Data Sharing Statement: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-495/dss

Peer Review File: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-495/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-495/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Danish Data Protection Agency (No. 19/30673, Date: 2020-06-12) and the Danish Patient Safety Authority (No. 3-3013-3132/1, Date: 2020-03-30), and individual consent for this retrospective analysis was waived.

References

1. The Global Cancer Observatory. Lung Cancer Fact Sheet. 2020 [cited 2023 Apr 23]. Available online: https://gco.iarc.fr/today/data/factsheets/cancers/15-Lung-fact-sheet.pdf
2. Walters S, Maringe C, Coleman MP, et al. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007. Thorax 2013;68:551-64.
3. Ganti AK, Klein AB, Cotarla I, et al. Update of Incidence, Prevalence, Survival, and Initial Treatment in Patients With Non-Small Cell Lung Cancer in the US. JAMA Oncol 2021;7:1824-32.
4. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. Br J Cancer 2009;101 Suppl 2:S80-6.
5. Smith RA, Andrews KS, Brooks D, et al. Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. CA Cancer J Clin 2019;69:184-210.
6. National Lung Screening Trial Research Team; Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395-409.
7. Dawson Q. NELSON trial: reduced lung-cancer mortality with volume CT screening. Lancet Respir Med 2020;8:236.
8. The Danish Cancer Society. Lung Cancer Screening. 2023. Available online: https://www.cancer.dk/forebyg/screening/lungekraeft/
9. Pham D, Bhandari S, Oechsli M, et al. Lung cancer screening rates: data from the lung cancer screening registry. J Clin Oncol 2018;36:6504.
10. Tang W, Peng Q, Lyu Y, et al. Risk prediction models for lung cancer: Perspectives and dissemination. Chin J Cancer Res 2019;31:316-28.
11. Marcus MW, Raji OY, Field JK. Lung cancer screening: identifying the high risk cohort. J Thorac Dis 2015;7:S156-62.
12. Toumazis I, Bastani M, Han SS, et al. Risk-Based lung cancer screening: A systematic review. Lung Cancer 2020;147:154-86.
13. Sharma V, Ali I, van der Veer S, et al. Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. BMJ Health Care Inform 2021;28:e100253.
14. Schmidt M, Schmidt SA, Sandegaard JL, et al. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol 2015;7:449-90.
15. Frank L. Epidemiology. When an entire country is a cohort. Science 2000;287:2398-9.
16. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13:S31-4.
17. Authority TDHD. Classifications. 2021 [cited 2023 Apr 2]. Available online: https://sundhedsdatastyrelsen.dk/da/english/health_data_and_registers/classifications
18. Organization WH. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). 2019 [cited 2023 Mar 27]. Available online: https://icd.who.int/browse10/2019/en
19. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43:1130-9.
20. Lange P, Tøttenborg SS, Sorknæs AD, et al. Danish Register of chronic obstructive pulmonary disease. Clin Epidemiol 2016;8:673-8.
21. Brusselaers N, Lagergren J. The Charlson Comorbidity Index in Registry-based Research. Methods Inf Med 2017;56:401-6.
22. Armitage JN, van der Meulen JH; Royal College of Surgeons Co-morbidity Consensus Group. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. Br J Surg 2010;97:772-81.
23. Mooney G. The Danish health care system: it ain't broke... so don't fix it. Health Policy 2002;59:161-71.
24. Practitioners TDO of G. Fees for services in General Practise [Internet]. 2021 [cited 2022 Nov 9]. Available online: https://www.laeger.dk/foreninger/plo/overenskomsten-og-aftaler/honorarer/
25. Practitioners TDC of G. Clinical Guideline: Quick-test for air infections in General Practise. 2020 [cited 2023 Apr 17]. Available online: https://vejledninger.dsam.dk/luftvejsinfektioner/?mode=visKapitel&cid=747
26. Cancer IA for R on. Nordcan. 2022. [cited 2023 Apr 1]. Available online: https://nordcan.iarc.fr/en/dataviz/trends?years=2009_2018&populations=208&cancers=160&

sexes=1_2&key=total
27. Authority TDH. The National Patients Registry [Internet]. 2022 [cited 2022 Nov 2]. Available online: https://www.esundhed.dk/Emner/Operationer-og-diagnoser/Landspatientregisteret-Avanceret-udtraek
28. Probst HB, Hussain ZB, Andersen O. Cancer patient pathways in Denmark as a joint effort between bureaucrats, health professionals and politicians--a national Danish project. Health Policy 2012;105:65-70.
29. Borg M, Hilberg O, Andersen MB, et al. Increased use of computed tomography in Denmark: stage shift toward early stage lung cancer through incidental findings. Acta Oncol 2022;61:1256-62.
30. Hyldgaard C, Trolle C, Harders SMW, et al. Increased use of diagnostic CT imaging increases the detection of stage IA lung cancer: pathways and patient characteristics. BMC Cancer 2022;22:464.
31. Statistics Denmark. Population index. 2023 [cited 2022 Nov 23]. Available online: https://www.statistikbanken.dk/statbank5a/default.asp?w=1920
32. Brenner DR, McLaughlin JR, Hung RJ. Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. PLoS One 2011;6:e17479.
33. Koshiol J, Rotunno M, Consonni D, et al. Lower risk of lung cancer after multiple pneumonia diagnoses. Cancer Epidemiol Biomarkers Prev 2010;19:716-21.
34. Sandfeld-Paulsen B, Meldgaard P, Aggerholm-Pedersen N. Comorbidity in Lung Cancer: A Prospective Cohort Study of Self-Reported versus Register-Based Comorbidity. J Thorac Oncol 2018;13:54-62.
35. Tammemagi CM, Neslund-Dudas C, Simoff M, et al. Impact of comorbidity on lung cancer survival. Int J Cancer 2003;103:792-802.
36. Lyratzopoulos G, Neal RD, Barbiere JM, et al. Variation in number of general practitioner consultations before hospital referral for cancer: findings from the 2010 National Cancer Patient Experience Survey in England. Lancet Oncol 2012;13:353-65.
37. McDonald L, Carroll R, Harish A, et al. Suspected cancer symptoms and blood test results in primary care before a diagnosis of lung cancer: a case-control study. Future Oncol 2019;15:3755-62.
38. Shiels MS, Pfeiffer RM, Hildesheim A, et al. Circulating inflammation markers and prospective risk for lung cancer. J Natl Cancer Inst 2013;105:1871-80.
39. Forkasiewicz A, Dorociak M, Stach K, et al. The usefulness of lactate dehydrogenase measurements in current oncological practice. Cell Mol Biol Lett 2020;25:35.
40. Acharya S, Kale J, Rai P, et al. Serum alkaline phosphatase in oral squamous cell carcinoma and its association with clinicopathological characteristics. South Asian J Cancer 2017;6:125-8.
41. Kang HS, Shin AY, Yeo CD, et al. Clinical significance of anemia as a prognostic factor in non-small cell lung cancer carcinoma with activating epidermal growth factor receptor mutations. J Thorac Dis 2020;12:1895-902.
42. Fujii T, Tokuda S, Nakazawa Y, et al. Implications of Low Serum Albumin as a Prognostic Factor of Long-term Outcomes in Patients With Breast Cancer. In Vivo 2020;34:2033-6.
43. Sandfeld-Paulsen B, Aggerholm-Pedersen N, Winther-Larsen A. Hyponatremia in lung cancer: Incidence and prognostic value in a Danish population-based cohort study. Lung Cancer 2021;153:42-8.
44. Gould MK, Huang BZ, Tammemagi MC, et al. Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data. Am J Respir Crit Care Med 2021;204:445-53.
45. Kang J, Chang Y, Ahn J, et al. Neutrophil-to-lymphocyte ratio and risk of lung cancer mortality in a low-risk population: A cohort study. Int J Cancer 2019;145:3267-75.
46. Winther-Larsen A, Aggerholm-Pedersen N, Sandfeld-Paulsen B. Inflammation-scores as prognostic markers of overall survival in lung cancer: a register-based study of 6,210 Danish lung cancer patients. BMC Cancer 2022;22:63.
47. Tian T, Lu J, Zhao W, et al. Associations of systemic inflammation markers with identification of pulmonary nodule and incident lung cancer in Chinese population. Cancer Med 2022;11:2482-91.
48. Ruano-Raviña A, Provencio M, Calvo de Juan V, et al. Lung cancer symptoms at diagnosis: results of a nationwide registry study. ESMO Open 2020;5:e001021.
49. Group DLC. Clinical guideline. 2020 [cited 2022 Nov 19]. Available online: https://www.lungecancer.dk/wp-content/uploads/2020/12/DLCG_visitation_diagn_stadie_AdmGodk141220.pdf
50. Marcus PM, Lenz S, Sammons D, et al. Recruitment methods employed in the National Lung Screening Trial. J Med Screen 2012;19:94-102.
51. van Iersel CA, de Koning HJ, Draisma G, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-

Belgian randomised lung cancer multi-slice CT screening trial (NELSON). Int J Cancer 2007;120:868-74.
52. Statistics Denmark. Access to Data. 2018 [cited 2022 Nov 2]. Available online: https://www.dst.dk/en/TilSalg/Forskningsservice/Dataadgang

Cite this article as: Henriksen MB, Hansen TF, Jensen LH, Brasen CL, Peimankar A, Ebrahimi A, Wiil UK, Hilberg O. A collection of multiregistry data on patients at high risk of lung cancer—a Danish retrospective cohort study of nearly 40,000 patients. Transl Lung Cancer Res 2023;12(12):2392-2411. doi: 10.21037/tlcr-23-495

Supplementary

Table S1 The disease categories, corresponding ICD-10 codes and CCI weights

Disease category	ICD-10 codes	Weights
Myocardial infarction	I21*, I22*, I23*, I252	1
Congestive cardiac failure	I11, I13, I255, I42, I43, I50, I517	1
Peripheral vascular disease	I70–I73, I770, I771, K551, K558, K559, R02, Z958, Z959	1
Cerebrovascular disease	G45, G46, I60–I69	1
Dementia	A810, F00–F03, F051, G30, G31	1
Chronic pulmonary disease	I26, I27, J40–J45, J46*, J47, J60–J67, J684, J701, J703	1
Rheumatological disease	M05, M06, M09, M120, M315, M32–M36	1
Liver disease	B18, I85, I864, I982, K70, K71, K721, K729, K76, R162, Z944	1
Diabetes mellitus	E10–E14	1
Hemiplegia or paraplegia	G114, G81–G83	2
Renal disease	I12, I13, N01, N03, N05, N07, N08, N171*, N172*, N18, N19*, N25, Z49, Z940, Z992	2
Other malignancy	C00–C26, C30–C33, C37–C41, C43, C45–C58, C60–C76, C80–C85, C88, C90–C97	2
Metastatic solid tumour	C77–C79	6
AIDS/HIV	B20–B24	6
Sum-CCI	N/A	0-27

ICD-10, International Classification of Diseases 10th Revision; CCI, Charlson comorbidity index; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus.

https://dx.doi.org/10.21037/tlcr-23-495

Table S2 Cancer location and ICD-10 codes included in other malignancies

Cancer location	ICD-10 code
Colorectal	C17-21, C26
Breast	C50
Prostate	C61
Lymphoma	C82-85
Leukemia	C88, C90-97
Head and neck	C00-C14, C30-33
Melanoma	C43
Bladder	C67-68
Kidney	C64-65
Mesothelioma	C45
Unknown primary tumor	C80
Brain	C70-72
Esophagus and stomach	C15-16
Liver	C22
Gallbladder	C23-24
Pancreas	C25
Mediastinal	C37-39
Bones	C40-41
Kaposi's sarcoma	C46
Nervous	C47
Peritoneum	C48
Connective tissue	C49
Female genitals	C51-52, C57-58
Cervix	C53
Uterus	C54-55
Ovarian	C56
Penis	C60
Testicle	C62
Male genitals	C63
Urethra	C66
Eye	C69
Thyroid	C73
Adrenal	C74-75
Unspecified	C76

ICD-10, International Classification of Diseases 10th Revision.

https://dx.doi.org/10.21037/tlcr-23-495

Table S3 The different types of medication and corresponding ATC-codes included

Included types of medication	ATC-code
Beta-lactamase sensitive penicillins	J01CE
Comb. of penicillins	J01CR
Macrolides	J01FA
Penicillins with extended spectrum	J01CA
Corticosteroids for systemic use	H02AB
Drugs for obstructive airway diseases	R03
Beta-blocking agents	C07A
Calcium channel blockers	C08
ACE inhibitors	C09AA
Non-selective monoamine reuptake inhibitors	N06AA
Selective serotonin reuptake inhibitors	N06AB
Other antidepressants	N06AX
Metformin	A10BA02
Warfarin	B01AA03
Phenprocoumon	B01AA04

ATC, Anatomical Therapeutic Chemical classification.

https://dx.doi.org/10.21037/tlcr-23-495

Table S4 Comparison of variables within the reduced cohort with complete results

Variables	Reduced cohort with complete results: 5,587 individuals		P value
	Non-LC (N=3,733, 66.8%)	LC (N=1,854, 33.2%)	
Demographics			
Age, median [IQR]	70 [58–78]	74 [68–80]	<0.001
Female sex, n (%)	1,605 (43.0)	955 (51.5)	<0.001
Smoking status, n (%)			
Never smokers	1,129 (30.2)	137 (7.4)	<0.001
Former smokers	1,621 (43.4)	1,063 (57.3)	<0.001
Active smokers	983 (26.3)	654 (35.3)	<0.001
Comorbidities, n (%)			
Myocardial infarction	44 (1.2)	14 (0.8)	0.141
Congestive cardiac failure	32 (0.9)	6 (0.3)	0.022
Peripheral vascular disease	105 (2.8)	72 (3.9)	0.031
Cerebrovascular disease	87 (2.3)	48 (2.6)	0.554
Dementia	16 (0.4)	8 (0.4)	0.989
Chronic pulmonary disease	412 (11.0)	173 (9.3)	0.050
Rheumatological disease	71 (1.9)	31 (1.7)	0.546
Liver disease	22 (0.6)	7 (0.4)	0.300
Diabetes mellitus	154 (4.1)	58 (3.1)	0.066
Hemiplegia or paraplegia	3 (0.1)	3 (0.2)	0.381
Renal disease	51 (1.4)	20 (1.1)	0.366
Other malignancies	629 (16.9)	184 (9.9)	<0.001
Metastatic solid tumor	188 (5.0)	42 (2.3)	<0.001
AIDS/HIV infection	0 (0)	1 (0.1)	0.156
Pulmonary tuberculosis	9 (0.2)	0 (0)	0.034
Sarcoidosis	4 (0.1)	2 (0.1)	0.994
Interstitial lung disease	29 (0.8)	5 (0.3)	0.022
Abscess	9 (0.2)	5 (0.3)	0.840
Pleural disease	88 (2.4)	17 (0.9)	<0.001
Pneumonia	291 (7.8)	95 (5.1)	<0.001
CCI =0	2,427 (65.0)	1,337 (72.1)	<0.001
CCI >0	1,306 (35.0)	517 (27.9)	<0.001
Medication, n (%)			
Antibiotics	1,530 (41.0)	734 (39.6)	0.317
COPD	965 (25.9)	595 (32.1)	<0.001
Beta blockers	672 (18.0)	344 (18.6)	0.614
Calcium antagonists	656 (17.6)	394 (21.3)	0.001
ACE inhibitors	425 (11.4)	241 (13.0)	0.080
Glucocorticoids	329 (8.8)	173 (9.3)	0.524
SSRI	265 (7.1)	128 (6.9)	0.789
Metformin	245 (6.6)	134 (7.2)	0.352
TCA	58 (1.6)	41 (2.2)	0.079
Consultations in general practice, n (%)			
0	314 (8.4)	102 (5.5)	<0.001
1–4	1,781 (47.7)	931 (50.2)	
>4	1,638 (43.9)	821 (44.3)	
CRP rapid tests in general practice, n (%)			
0	1,755 (47.0)	844 (45.5)	0.184
1–4	1,827 (48.9)	948 (51.1)	
>4	151 (4.1)	62 (3.3)	
Laboratory analyses, median [IQR]			
B-hemoglobin, mmol/L	8.6 [8.0–9.3]	8.5 [7.8–9.0]	<0.001
B-leucocytes, 10^9/L	7.69 [6.22–9.46]	8.89 [7.40–10.80]	<0.001
B-neutrophils, 10^9/L	4.76 [3.63–6.22]	5.88 [4.63–7.59]	<0.001
B-lymphocytes, 10^9/L	1.78 [1.35–2.34]	1.77 [1.35–2.30]	0.7993
NLR	2.68 [1.85–3.88]	3.20 [2.33–4.80]	<0.001
B-monocytes, 10^9/L	0.65 [0.51–0.84]	0.74 [0.58–0.93]	<0.001
B-basophils, 10^9/L	0.04 [0.03–0.06]	0.05 [0.02–0.07]	0.1664
B-eosinophils, 10^9/L	0.16 [0.09–0.27]	0.14 [0.07–0.24]	<0.001
B-platelets, 10^9/L	273 [224–335]	306 [246–385]	<0.001
P-albumin, g/L	43 [41–45]	42 [39–44]	<0.001
Total calcium, mmol/L	2.36 [2.29–2.43]	2.38 [2.31–2.45]	<0.001
P-CRP, mg/L	3.8 [1.5–11.0]	7.7 [2.7–25.0]	<0.001
P-ALAT, U/L	22 [16–31]	19 [14–27]	<0.001
P-LDH, U/L	199 [176–228]	214 [187–250]	<0.001
P-alkaline phosphatase, U/L	76 [63–92]	83 [68–101]	<0.001
P-bilirubin-total, µmol/L	8 [6–10]	7 [5–9]	<0.001
P-amylase (pancreatic), U/L	25 [19–34]	25 [19–34]	0.678
P-INR	1.0 [0.97–1.1]	1.0 [0.95–1.1]	0.001
P-creatinine, mmol/L	78 [65–91]	72 [60–87]	<0.001
P-sodium, mmol/L	140 [138–141]	139 [136–141]	<0.001
P-potassium, mmol/L	4 [3.8–4.3]	4.1 [3.8–4.3]	0.982
Symptoms, n (%)			
Predispositions	253 (6.8)	167 (9.0)	0.003
Expositions	785 (21.0)	354 (19.1)	0.091
Hemoptysis	694 (18.6)	212 (11.4)	<0.001
Pneumonia	671 (18.0)	303 (16.3)	0.130
Cough	2,012 (53.9)	696 (52.3)	0.249
Dyspnoea	1,365 (36.6)	663 (35.8)	0.556
Fever	268 (7.2)	81 (4.4)	<0.001
Weight loss	822 (22.0)	584 (31.5)	<0.001
Fatigue	684 (18.3)	428 (23.1)	<0.001
Hot flash	402 (10.8)	177 (9.6)	0.158
Hoarseness	174 (4.7)	92 (5.0)	0.619
Back pain	133 (3.6)	129 (7.0)	<0.001
Other pain	340 (9.1)	250 (13.5)	<0.001
Angina	428 (11.5)	256 (13.8)	0.012
Headache	114 (3.1)	65 (3.5)	0.366
Dizziness	161 (4.3)	96 (5.2)	0.146
Edema	196 (5.3)	108 (5.8)	0.372

LC, lung cancer; IQR, interquartile range; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; CCI, Charlson comorbidity index; P-, plasma; B-, blood; ALAT, alanine aminotransferase; CRP, C-reactive protein; INR, international normalized ratio; LDH, lactate dehydrogenase; NLR, neutrophil to lymphocyte ratio; COPD, chronic obstructive pulmonary disease; ACE, angiotensin-converting-enzyme; SSRI, selective serotonin reuptake inhibitors; TCA, tricyclic antidepressants.

https://dx.doi.org/10.21037/tlcr-23-495